LANGUAGE DEVELOPMENT RESEARCH

An Open Science Journal

Volume 4 | Issue 1 | December 2024

ISSN 2771-7976

About the journal

Language Development Research: An Open-Science Journal was established in 2020 to meet the field's need for a peer- reviewed journal that is committed to fully open science: LDR charges no fees for readers or authors, and mandates full sharing of materials, data and analysis code. The intended audience is all researchers and professionals with an interest in language development and related fields: first language acquisition; typical and atypical language development; the development of spoken, signed or written languages; second language learning; bi- and multilingualism; artificial language learning; adult psycholinguistics; computational modeling; communication in nonhuman animals etc. The journal is managed by its editorial board and is not owned or published by any public or private company, registered charity or nonprofit organization.

Child Language Data Exchange System

Language Development Research is the official journal of the **TalkBank system**, comprising the CHILDES, PhonBank, HomeBank, FluencyBank, Multilingualism and Clinical banks, the CLAN software (used by hundreds of researchers worldwide to analyze children's spontaneous speech data), and the Info-CHILDES mailing list, the de-facto mailing list for the field of child language development with over 1,600 subscribers.

Diamond Open Access

Language Development Research is published using the Diamond Open Access model (also known as “Platinum” or “Universal” OA). The journal does not charge users for access (e.g., subscription or download fees) or authors for publication (e.g., article processing charges).

Hosting

The **Carnegie Mellon University Library Publishing Service** (LPS) hosts the journal on a Janeway Publishing Platform with its manuscript management system (MMS) used for author submissions.

License

Language Development Research is published by TalkBank and the Carnegie Mellon University Library Publishing Service. Authors retain the copyright to their published content. This work is distributed under the terms of the **Creative Commons Attribution-Noncommercial 4.0 International license** (https://creativecommons.org/licenses/by-nc/4.0/), which permits any use, reproduction and distribution of the work for noncommercial purposes with no further permissions required provided the original work is attributed as specified under the terms of this Creative Commons license.

Peer Review and Submissions

All submissions are reviewed by a minimum of two peer reviewers, and one of our Action Editors, all well- established senior researchers, chosen to represent a wide range of theoretical and methodological expertise. Action Editors select peer reviewers based on their expertise and experience in publishing papers in the relevant topic area.

Submissions and Publication Cycle

We invite submissions that meet our criteria for rigour, without regard to the perceived novelty or importance of the findings. We publish general and special-topic articles (“Special Collections”) on a rolling basis to ensure rapid, cost-free publication for authors.

Language Development Research is published once a year, in December, with each issue containing the articles produced over the previous 12 months. Individual articles are published online as soon as they are produced. For citation purposes, articles are identified by the year of first publication and digital object identifier (DOI).

Table of Contents

Volume 4, Issue 1, December 2024

Syntactic adaptation and word learning in children and adults

Elizabeth Swanson
University of Maryland, USA

Michael C. Frank
Judith Degen
Stanford University, USA

Abstract: Syntactic adaptation may be a key mechanism underlying children's learning of novel words. Havron et al. (2019) exposed French-speaking children (ages 3 to 4) to a speaker biased toward using either familiar verbs or familiar nouns in a syntactic context which permitted both structures. This prime later influenced participants' interpretations of ambiguous novel words presented in the same syntactic frame. In Experiment 1, we successfully replicated Havron et al. with 77 French-speaking adults, using a web-based eye-tracking paradigm. Experiment 2 adapted the paradigm to English, finding that repeated exposure to a syntactic structure induced 102 English-speaking adults to update their expectations about the meanings of novel words. Experiment 3 found similar evidence of syntactic adaptation in 74 three- to five-year-old English-speaking children. Participants adapted to the specific linguistic structure used, not just the speaker's tendency to mention actions or objects. These findings support the role of rapid adaptation during word learning and demonstrate the feasibility of conducting eye-tracking studies through online platforms.

Keywords: syntactic priming; adaptation; word learning; language acquisition

Corresponding author: Elizabeth Swanson, Department of Linguistics, University of Maryland, College Park, MD 20742. Email: eswan@umd.edu.

ORCID IDs: Elizabeth Swanson: https://orcid.org/0000-0001-8004-4289; Michael C. Frank: https://orcid.org/0000-0002-7551-4378; Judith Degen: https://orcid.org/0000-0003-2513-0234

Citation: Swanson, E., Frank, M.C., & Degen, J. (2024). Syntactic adaptation and word learning in children and adults. *Language Development Research, 4*(1), 1–41. https://doi.org/10.34842/0ef7-5497

Introduction

How do children learn language so quickly? In just a few years, children can learn how to segment a continuous speech stream into words and phrases and map this linguistic content to its meaning. One source that children may draw on when learning unfamiliar words is morphosyntactic information. Specifically, syntactic bootstrapping has been proposed as a process by which children can infer the meanings of unfamiliar words partially based on their morphosyntactic characteristics (Gleitman, 1990). For example, upon hearing a sentence such as *It's daxing*, a child can use the *-ing* affix to infer that *dax* is a verb and therefore likely refers to an action. In this case, the *-ing* affix is a relatively stable and reliable cue to the novel word's part of speech. However, language is highly variable across speakers and situations. To cope with such variability, one mechanism listeners can rely on is linguistic adaptation: the ability to track patterns in the speech of others and update their expectations based on these patterns. Adaptation, including adaptation to a speaker's choice of syntactic structure, is well-studied in adults (Bradlow & Bent, 2008; Chang et al., 2006; Fine et al., 2013; Kleinschmidt & Jaeger, 2015; Kraljic & Samuel, 2007; Ostrand & Ferreira, 2019; Prasad & Linzen, 2021; Ryskin et al., 2019; Schuster & Degen, 2020; Yildirim et al., 2016). Do children also exhibit evidence of syntactic adaptation? And can they use expectations updated during syntactic adaptation to bootstrap word learning?

Havron, de Carvalho, Fiévet, & Christophe (2019) investigated children's capacity to infer novel word meanings by adapting to specific syntactic structures, showing that French-speaking adults and children demonstrated rapid syntactic adaptation after repeated exposure to a particular sentence structure. Furthermore, participants drew on these expectations to guide their learning of unfamiliar words that were presented in the same syntactic context. In this paper, we describe three experiments that replicate the findings of Havron et al. (2019) in a web-based eye-tracking paradigm and extend the findings to English-speaking adults and children. These studies build on prior work examining both syntactic priming and syntactic bootstrapping.

Syntactic Priming in Adults

Syntactic priming in adults is a well-established phenomenon, in which exposure to a particular sentence structure increases the likelihood of participants producing that structure themselves (Bock, 1986; Branigan et al., 2000, 2007; Cleland & Pickering, 2003; Ostrand & Ferreira, 2019; Pickering & Garrod, 2004*)* and demonstrating facilitated comprehension of utterances that contain the structure (Fine et al., 2013; Fine & Jaeger, 2013; Kamide, 2012; Lu et al., 2021; Prasad & Linzen, 2021). On the production side, experimental studies have long shown that participants tend to align their syntactic structures in dialogue (Bock, 1986). Participating actively in a dialogue, rather than listening as a side participant, has been linked to a greater degree of alignment (Branigan et al., 2007). Syntactic alignment effects have also been found with datives and verb particle placement (e.g., *John picked up the book* vs. *John picked the book up*) in a corpus of naturalistic dialogue (Gries, 2005), indicating that syntactic alignment is not merely a product of experimental settings but also a characteristic of natural communication.

In addition, syntactic priming effects have increasingly been investigated in comprehension (Pickering & Ferreira, 2008). One study used a self-paced reading paradigm to examine participants' comprehension of garden path sentences (Fine et al., 2013). After repeated exposures to these sentences, participants adapted to the new syntactic distribution, reducing or even eliminating the processing disadvantage (though cf. Harrington Stack et al., 2018). Syntactic priming can also guide understanding of syntactically ambiguous utterances, with participants interpreting utterances as being consistent with the type of structure they previously heard (Kamide, 2012). Similarly, syntactic adaptation has been proposed as a mechanism underlying satiation effects, where upon repeated exposure listeners are more likely to judge ungrammatical sentences as acceptable (Lu et al., 2021).

Several studies have suggested that syntactic priming involves not just transient activation of representations, but can also have long-term, cumulative effects. An experiment that used a similar picture task as Bock (1986) to elicit sentences containing dative verbs found that syntactic priming still occurred when there was a 20-minute delay between the priming stage and participants' productions (Boyland & Anderson (1998). Even studies in which syntactic priming took place days before the test stage have reported that participants exhibited adaptation to difficult sentence structures, such as ambiguous relative clauses, and came to process them more quickly (Long & Prat, 2008; Wells et al., 2009). Furthermore, even rapid syntactic priming appears to be cumulative, meaning that greater exposure to a particular sentence structure leads to an incrementally larger processing advantage (Fine & Jaeger, 2016; Kaschak, 2007).

While syntactic priming has sometimes been attributed to short-lived activation of representations (Branigan et al., 2000; Pickering & Branigan, 1998; Pickering & Garrod, 2004), the findings of cumulative and long-term priming effects lend support to an explanation of syntactic priming effects as a form of adaptation that is linked to implicit learning about the distributions of sentence structures (Bock & Griffin, 2000; Branigan & Messenger, 2016). Additional evidence for the implicit learning account stems from the finding that the change in listeners' syntactic expectations is influenced by the size of the error signal accompanying a particular syntactic prime (Fine & Jaeger, 2013). Recently, syntactic adaptation has also been modeled as a process of rational belief update, in which the reliability of a cue is taken into account to determine whether listeners should update their expectations (Havron et al., 2020). Differential adaptation depending on a cue's reliability has been found in both adults and four- to five-year-old children (Beretti et al., 2020; Yurovsky et al., 2017). Moreover, some studies have suggested that syntactic priming is speaker-specific (Kamide, 2012; Kroczek & Gunter, 2017; Lu et al., 2021; Yildirim et al., 2016), though others have failed to find such effects (Liu et al., 2017; Ostrand & Ferreira, 2019). Thus, although the exact mechanism remains disputed, syntactic alignment (in production) and syntactic priming (in comprehension) have been clearly demonstrated in adults.

Syntactic Priming in Children

Syntactic priming has the potential to act as a powerful support for children's language acquisition. A number of studies have shown that infants and children are able to engage in statistical learning, meaning that they can extract statistical regularities from an input (Arciuli & Simpson, 2011; Arnon, 2019; Krogh et al., 2013; Saffran et al., 1996; Saffran & Kirkham, 2017; Shufaniya & Arnon, 2018). In the auditory domain, statistical learning appears to develop very early on, from at least the age of 8 months, leading many to suggest that it plays an important role in early language learning (Arciuli & Torkildsen, 2012; Raviv & Arnon, 2018; Romberg & Saffran, 2010). With regard to syntax, in particular, 1-year-old infants have been found to be able to extract grammatical information from statistical regularities in an artificial language after less than two minutes of exposure (Gomez & Gerken, 1999). Such a mechanism could also allow children to rapidly adapt to syntactic patterns in the language input.

Indeed, multiple studies have demonstrated that children are sensitive to syntactic priming, although these effects are sometimes more difficult to detect than with adults depending on the task demands (Shimpi et al., 2007). For instance, children ages three to six and adults showed effects of syntactic alignment with datives, during a task where they were prompted to describe cartoon animations (Peter et al., 2015). Children have also been shown to align with active- and passive-voice sentences, producing more sentences of the type they were previously exposed to (Bencini & Valian, 2008; Messenger et al., 2011).

In addition to alignment studies, children are sensitive to syntactic priming in comprehension. Thothathiri & Snedeker (2008) used an eye-tracking paradigm to measure children's expectations about temporarily ambiguous datives (e.g., direct object: *Show the horse the book* vs. prepositional object: *Show the horn to the dog*). When children had been primed with either DO or PO sentences, they were more likely to interpret a temporarily ambiguous phrase (such as *Show the hor*—) in a manner consistent with the structure used during priming. Like adults, children have also shown cumulative effects of syntactic priming over the course of an experiment (Huttenlocher et al., 2004), including when the priming stimuli used nonsense verbs (Brooks & Tomasello, 1999). Branigan & Messenger (2016) found a difference between priming effects in children and adults: While both groups showed immediate effects of syntactic adaptation, only children demonstrated significant *cumulative* effects in a second session a week later. Cumulative syntactic priming has also been shown over the course of a single session, for the interpretation of ambiguous sentences, with a larger effect in five- to six-year-old children than in adults (Havron et al., 2020). Relatedly, the magnitude of the priming effect has been found to be larger for young children than for older children and adults (Rowland et al., 2012). These results suggest that, at least in some contexts, children may have expectations about sentence structure that are more uncertain or more flexibly updated than adults' expectations. A greater ability to adapt could help children learn more quickly in unfamiliar linguistic contexts. Thus, it is reasonable to propose that syntactic adaptation may play a role in not just children's sentence processing, but also their acquisition of language.

The connection between acquisition and an adaptation account of syntactic priming is motivated by prior work: for instance, Chang et al. (2006) developed a connectionist model of sentence production that used error-based learning to imitate the acquisition of syntax. That is, after encountering a violation of its predictions, the model updated its expectations about upcoming syntactic material. The model was able to account for many syntactic priming effects in adults and children, including the finding that more surprising structures are associated with larger priming effects (Bernolet & Hartsuiker, 2010; Fine & Jaeger, 2013; Jaeger & Snider, 2013). On the other hand, one study did not find evidence of an immediate prime surprisal effect in children, while it did in adults, raising questions about whether children are truly engaging in error-based learning (Fazekas et al., 2020). Both groups did, however, show syntactic priming effects on production, and more surprising input was associated with stronger priming overall.

This work suggests that encountering an unexpected distribution of syntactic structures could lead children to update their expectations and, importantly, recruit those expectations during word learning. For example, in a naturalistic context, a child might hear an adult describing a toy dog using repeated similar syntactic frames, such as *The dog is running, The dog is playing,* etc. Adapting to the use of this syntactic frame would allow the child to more easily learn a novel word presented in the same frame. Such a mechanism has the potential to unify accounts of adaptation in language processing with accounts of language acquisition, which was a key motivation for Havron et al. (2019).

Syntactic Bootstrapping and Word Learning

The syntactic bootstrapping literature provides further motivation for the idea that syntactic information is recruited during word learning. Knowledge of a small number of syntactic cues could prove immensely helpful in constraining children's hypotheses about the meaning of a novel word, such as inferring that *dax* in *It's daxing* is a verb that refers to an action (Brown, 1957; Gleitman, 1990; Waxman, 1999).

Experimental evidence indicates that children are able to draw on syntax during word learning from an early age. Upon hearing *This one is a blicket,* infants as young as 14 months infer that *blicket* refers to an object and not an object property; they make no such inference for *This one is blickish* (Booth & Waxman, 2003). 24-month-olds are sensitive to the syntactic context of novel words and draw on syntactic cues to help them construe images of scenes (Waxman et al., 2009). Using eye-tracking paradigms, studies have reported that 18-month-olds (He & Lidz, 2017) and 23-month-olds (Bernal et al., 2007) can use syntactic cues from phrases such as *It's pooning* vs. *It's a poon* to map novel words to images portraying either actions or objects, respectively. At a broader level, children who are more sensitive to syntactic cues in general have been found to have more accurate interpretations of novel words (Huang & Arnold, 2016).

Much work on syntactic bootstrapping has examined children's ability to use verb arguments to guide their interpretations of verbs (Gleitman et al., 2005). Specifically, a structure-mapping account of verb learning proposes that children have a universal bias to map each noun phrase in a sentence onto a participant role in an event (Fisher, 1994; Fisher et al., 2020; Naigles, 1990). For instance, Yuan & Fisher (2009) played sentences containing novel words that were either transitive (e.g., *She blicked the baby*) or intransitive (e.g., *She blicked*). They found that two-year-olds who heard transitive sentences looked longer at pictures with two people in them rather than one, indicating that they used syntactic cues (i.e., presence of a direct object in transitive sentences) to interpret the novel words. Follow-up work has found similar abilities in 22-month-olds (Messenger et al., 2015) and 15-month-olds (Jin & Fisher, 2014).

Thus, there is ample evidence that children are sensitive to syntactic cues from an early age and use them as a source of information during word learning. Furthermore, computational models have been able to simulate syntactic bootstrapping from limited language input, acquire syntactic categories, and perform well in word-learning tasks (Alishahi & Stevenson, 2008; Brusini et al., 2021; Christodoulopoulos et al., 2016; Christophe et al., 2016). This supports the proposal that syntactic bootstrapping plays an important role in children's word learning. However, syntactic cues are useful especially because they are relatively stable across language—to what extent would children be able to bootstrap novel word meanings based on recently updated expectations, as in syntactic adaptation?

Havron et al. (2019) and the Current Studies

To sum up, both children and adults exhibit syntactic priming in comprehension and production. In addition, syntactic cues appear to play a key role in children's word learning via syntactic bootstrapping. Havron et al. (2019) brought these two lines of work together by investigating whether syntactic adaptation is a driving force in children's acquisition of novel words. Specifically, the study examined whether priming French-speaking children with a particular syntactic structure would influence the meaning they assigned to novel words in an ambiguous context. During training trials, three- and four-year-old children were exposed to repeated trials of a French phrase (*La petite*) that can be followed by either a noun or a verb (e.g., *La petite grenouille [The little frog]* vs. *La petite dort [The little one sleeps]*). On test trials, children heard novel words presented in the same syntactic frame (e.g., *La petite nuve*), and their eye movements were measured to see whether children looked more at an image depicting a novel object or an image depicting a novel action. Children (and an adult comparison group) appeared to update their predictions about which syntactic structure a speaker would use, and they drew on these predictions to infer the meaning of a novel word.

The studies reported here build on the work of Havron et al. (2019) in several ways. First, in Experiment 1, we tested whether these results would directly replicate in a new context: an eye-tracking study conducted entirely online, with adults. Next, we conducted a crosslinguistic replication of the study in English, using a syntactic frame (*The girls/The girl's*) that can similarly be followed by either a noun or a verb (e.g., *The girls sleep* vs. *The girl's book*). We first ran this study online with adults (Experiment 2) and then carried it out with three- to five-year-old children (Experiment 3). These studies examined whether the results of Havron et al. (2019) would replicate in a different language and using novel methods: eye-tracking in a web-based environment. Thus, Experiment 1 provides a validation of the novel method, while Experiments 2 and 3 constitute a cross-linguistic test of the main hypothesis: if syntactic adaptation is a mechanism underlying word learning, then upon encountering an unfamiliar word, English-speaking adults and children should look more at the image (action or object) matching the type of phrase (verb or noun) they heard during training trials.[1]

Experiment 1

Experiment 1 was a direct replication of Havron et al. (2019) that was carried out using web-based eye-tracking. This study served the dual purpose of both replicating the original study and validating web-based eye-tracking as a paradigm suitable for studying the interaction of syntactic bootstrapping and adaptation.

[1] We preregistered all three experiments on the Open Science Framework at: https://osf.io/3j6rw/. All stimuli, data, and analyses for Experiments 1, 2, and 3 can be found at: https://github.com/eswanson166/syntactic-adaptation-and-word-learning.

Method

Participants

We collected data from 77 participants (31 female; 46 male) using Prolific (www.prolific.co), an online crowdsourcing website. All were adults who reported speaking French as their first language.

Procedure

A diagram of the experimental set-up is shown in Figure 1. The stimuli used in the study, as well as the structure of the trials, were identical to those used in Havron et al. (2019) and were downloaded from the authors' repository at https://osf.io/zzd9y/. Every participant was randomly assigned to either the noun condition (37 participants) or the verb condition (40 participants). Participants completed a 9-point calibration, which was adapted from the original study to work with the web-based eye-tracking Javascript library WebGazer (Papoutsaki et al., 2016). The study consisted of two phases: a training phase and a test phase. The total experiment included ten trials and lasted about twelve minutes.

On each training trial, all participants saw two videos. One showed a girl performing a familiar action (such as jumping), while the other showed the same girl holding a familiar object (such as a toy car). The structure of each training trial was identical. First, the participant saw a preview of one video only, followed by a preview of the other video. Then, during the contrast phase, the participant saw both videos together. For these parts of the trial, a female narrator told the child to look at the videos in a child-friendly voice, but she did not comment on what the videos depicted. The last part of the trial was the event phase, during which children saw both videos again, but the narrator described what was in just one of the two videos. If participants were in the noun condition, she said a phrase such as *La petite grenouille* ("The little frog"). If participants were in the verb condition, she said a phrase such as *La petite dort* ("The little one [feminine] is sleeping"). Thus, participants in both conditions heard the same syntactic frame: *La petite [X],* but it was followed by either a noun (meaning "The little X") or a verb (meaning "The little one is Xing"). Participants were exposed to four training trials. The side of the screen where the target video appeared was counterbalanced, and the order of the training trials was randomized.

In between the first two training trials and the last two training trials, participants watched two filler trials. These trials had the same structure as the training trials except that the narrator referred to the type of video that was *not* referred to in the training trials, using a structure that was unambiguous. Therefore, participants in the noun condition heard a description of the action video in a sentence such as *Elle écrit*

("She writes"), since *Elle…* cannot be followed by a noun. Similarly, participants in the verb condition heard a description of the object video in a sentence such as *C'est une poussette* ("It's a baby-stroller"), because *C'est une…* cannot precede a verb. These filler trials were included so that participants would understand that the narrator could refer to either the action video or the object video. It was simply with the structure *La petite…* that the narrator was biased toward using either nouns or verbs. This also reduced the possibility that participants would look toward the action or object video on test trials purely because they were used to looking at that type of video.

After the training trials, all participants watched three test trials, which were identical regardless of condition (though the order was again randomized). Test trials had the same structure as training trials, but the two videos depicted a novel object and a novel action. Also, participants heard the narrator's description once before the event phase started so that looks could be measured from the beginning of the event phase. The narrator used the same *La petite…* context as before, but it was followed by an unfamiliar word that does not actually exist in French, such as *La petite nuve.* Since *La petite…* can be followed by a noun or a verb, participants could in principle interpret *nuve* as a noun or a verb. However, if participants adapt to the structure

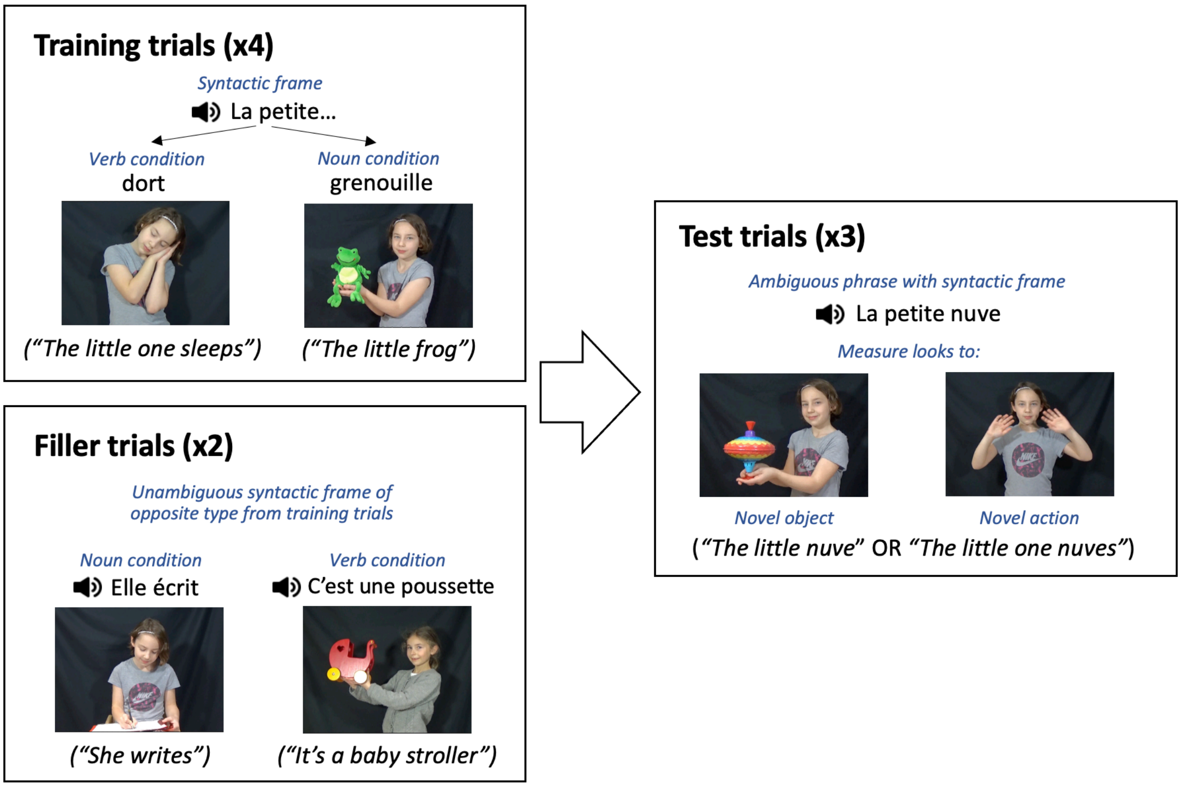


Figure 1. ***Diagram of experimental set-up for Experiment 1.***

preferred by the speaker during training trials, they should behave differently in the different conditions. In particular, they should interpret novel words as nouns in the noun condition, and therefore look more at the object video during test trials; conversely, they should interpret novel words as verbs in the verb condition, and therefore look more at the action video during test trials. In line with previous eye-tracking studies, we considered a greater proportion of looks to a video to be an indicator that participants interpreted the word as matching what was depicted in the video.

As in Havron et al. (2019), there was also one trial at the end of the experiment which used the structure *Le petit [X],* the masculine form of the *La petite [X]* structure, and which showed videos depicting a boy rather than a girl. This was an exploratory trial to examine whether the adaptation effect would generalize to a slightly different structure.

Measures

We measured participants' eye movements using WebGazer, a program that estimates the coordinates of participants' eye movements on the computer screen using a webcam (Papoutsaki et al., 2016). WebGazer is a novel method for conducting eye-tracking studies, and as a direct replication of Havron et al. (2019), Experiment 1 was an ideal way to examine the utility of WebGazer for psycholinguistic research.

All analyses were conducted in R (R Core Team, 2021). WebGazer recorded 81% of total looks as being directed to the screen; the remaining 19% presumably reflected participants looking away or blinking, or WebGazer losing track of their gaze. We followed the common practice of only analyzing looks that were to relevant regions of the display, in this case either the action video or the object video (46% of the total looks in the dataset). In the analyses, we report only the looks to the action video, because when only the regions of interest are examined, any look not to the action video is to the object video.

Results

Proportion of Looks

We calculated each participant's proportion of looks to the action video on each test trial and then averaged these three proportions to obtain each participant's mean proportion of looks to the action video across the three test trials. Since participants heard the full target phrase once before the videos appeared in the event phase, we measured looks from the beginning of the event phase when both videos appeared on screen together. Figure 2a shows the overall mean proportion of looks to the action video in each condition, as well as dots representing individual participants' mean proportions of looks. As hypothesized, participants in the verb condition ($M = 0.585$, $SD = 0.171$) were more likely to look at the action video than participants in the noun condition ($M = 0.395$, $SD = 0.171$).

We conducted a preregistered mixed effects linear regression analysis predicting the arc-sin transformed proportion of looks to the action image during a trial (the same as in the Havron et al. study).[2] The lme4 package was used to conduct the regression analyses (Bates et al., 2015), and the reported p-values were calculated using Satterthwaite's degrees of freedom method via the lmerTest package (Kuznetsova et al., 2017).

In the mixed effects linear regression, we predicted participants' arc-sin transformed mean proportion of looks to the action video as a function of condition, with a random by-participant intercept. Condition was centered to avoid high collinearity with the intercept. We did not include a random intercept for item since there were only three test items. There was a main effect of condition in the direction expected: Participants in the verb condition were significantly more likely to look at the action video than participants in the noun condition ($\beta = 0.218$, $SE = 0.048$, $p < 0.01$).

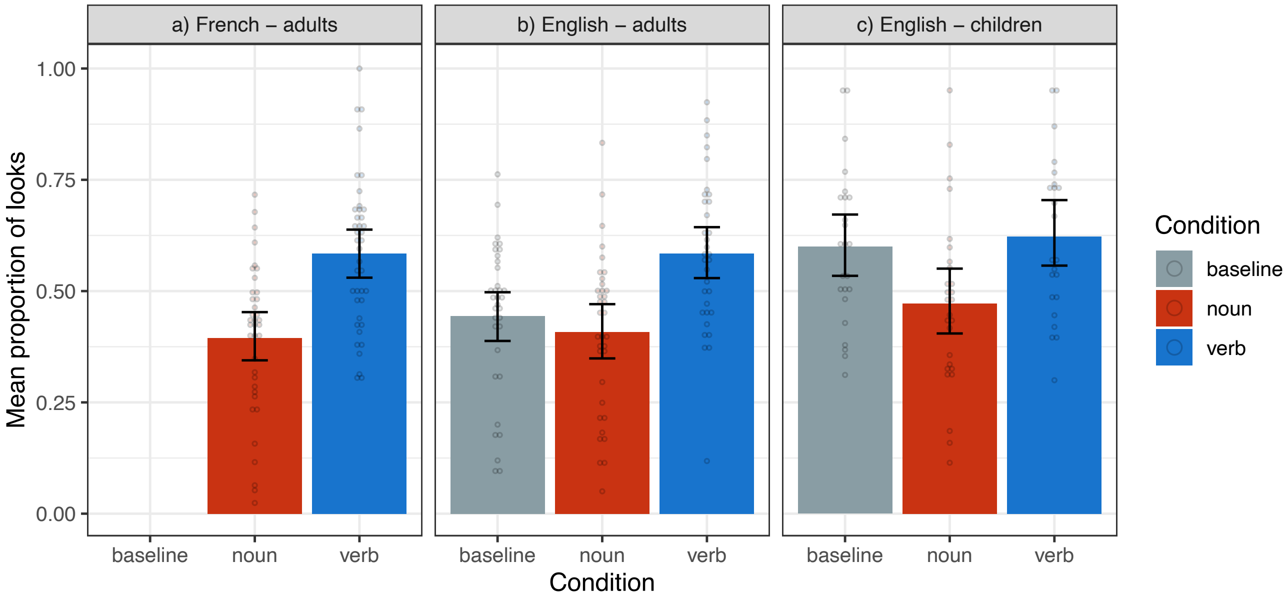


Figure 2. ***Mean overall proportion of looks to the action video or image for a) Experiment 1, b) Experiment 2, and c) Experiment 3. Results are shown for the noun, verb, and (when applicable) baseline conditions during test trials, with bootstrapped confidence intervals. Semi-transparent dots correspond to the mean proportion of looks for individual participants, averaged across the test trials.***

[2] Across all three experiments, we also preregistered a mixed effects logistic regression analysis that directly predicted individual looks to the action image. All results agreed between the two types of models, so the logistic regression analyses are reported in the Supplementary Materials.

Time Course

While the results for proportion of looks demonstrate that adults are indeed using syntactic adaptation to bootstrap novel word meanings, an additional question of interest is how quickly this information can be recruited. Time course data can provide insight into this question. If participants were quickly adjusting their expectations based on the use of the frame *La petite…,* we should see a bias to the action or object video (depending on condition) from the very start of the test trial. Because with the Havron et al. stimuli, participants heard the test trial audio once before the videos appeared on-screen, we do not have information about their eye movements during the first instance of hearing *La petite [novel word].* However, in Figure 3a we present a time course plot which suggests that participants in the verb condition looked significantly more at the action video throughout almost the entire event phase of the test trial, and participants in the noun condition consistently looked more at the object video. In Experiments 2 and 3, we showed participants the images before they heard the first instance of the novel words, in order to examine whether their looking patterns changed over the course of the trial.

Training and Filler Trials

We also conducted exploratory post-hoc analyses of training and filler trials to confirm that participants did in fact look at the video described during training trials. This was important to ensure that (a) the eye-tracker reliably measured looks and (b) participants reacted to the descriptions they heard in expected ways. On filler trials, participants should look at the opposite video of their assigned condition. Doing so would indicate their understanding that the narrator could refer to both types of videos, and that it was just with the structure *La petite…* that she was biased toward one type of video.

As expected, during training trials, participants in the verb condition looked significantly more to the action video than those in the noun condition (β = 0.518, *SE* = 0.044, $p < 0.001$). The pattern was reversed on filler trials (β = -0.433, *SE* = 0.056, $p <$ 0.001). More detailed analysis and visualization of training and filler trials, as well as of the exploratory generalization trial[3], is available in the GitHub repository.

[3] On the exploratory generalization trial, participants in the verb condition looked significantly more at the action video than participants in the noun condition. More detail is provided in the Supplementary Materials.

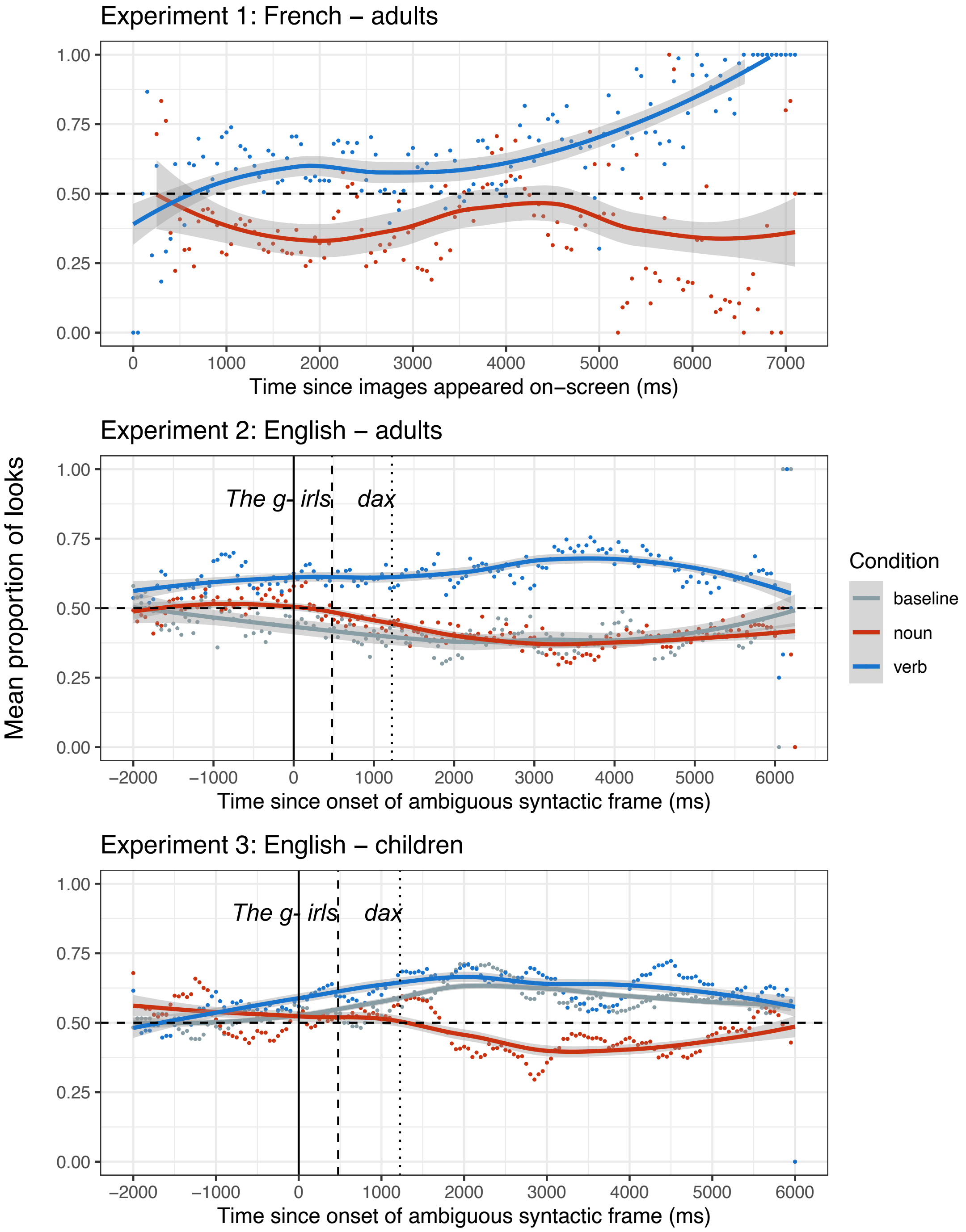


Figure 3. ***Proportion of looks to the action video or image over time on test trials of Exp. 1 (top), Exp. 2 (middle), and Exp. 3 (bottom). Gray areas represent overall confidence intervals. For Experiments 2 and 3, the zero point (indicated by the vertical black line) corresponds to the onset of the ambiguous syntactic frame (The g-); the dashed line represents the mean time point of the end of the syntactic frame, The girls/girl's...; and the dotted line indicates the mean end time point of the first utterance of the novel word, such as The girls/girl's dax. For Experiment 1, participants heard the full target phrase once before the videos appeared on-screen, so we do not mark these time points.***

Discussion

Experiment 1 directly replicated the adult results of Havron et al. (2019), which examined whether syntactic priming influenced word learning. The original study found that participants adapted to a repeated syntactic structure and that they used their updated expectations to interpret an unfamiliar word. Our results were similar: We observed a significant effect of condition such that, compared to participants who heard *La petite (noun)* on training trials, participants who heard *La petite (verb)* looked significantly more at the action video on test trials. Additionally, the time course data suggests that the effect remained consistent throughout the trial. Thus, we found evidence that participants interpreted the ambiguous words on test trials to be consistent with the syntactic structure (noun vs. verb) that had previously been used by the narrator.

The difference we found between conditions appears to be smaller than in the original paper. Havron et al. (2019) reported a mean proportion of looks of 0.653 in the verb condition (compared to our 0.585) and 0.275 in the noun condition (compared to our 0.395); the size of the standard deviations was similar. The smaller effect size is not surprising given that it was a replication (Open Science Collaboration, 2015) and that online eye-tracking is noisier than eye-tracking with in-lab devices (Degen et al., 2021; Madsen et al., 2021; Semmelmann & Weigelt, 2018).

Is it possible that web-based eye-tracking is the wrong tool for investigating our questions of interests? We think not. First, despite the smaller effect size, we replicated the results of Havron et al. (2019). Furthermore, participants were quite clearly looking at the expected videos during both training and filler trials, when it was obvious which video was being described. WebGazer's rate of track loss in our study (19%) was just slightly worse than the upper range (11.1%—17.6%) reported in a study that compared 12 different in-lab eye-trackers with adults (Holmqvist, 2017), and it is on par with the values reported in a comparison of two in-lab eye-trackers (17% and 20%) with three-year-old children (De Kloe et al., 2022). This aligns with previous findings that WebGazer is slightly less accurate than in-lab eye-trackers (with an average offset of 207 pixels vs. 172 pixels for in-lab) and shows higher variance, while still replicating results established in lab-based eye-tracking (Semmelmann & Weigelt, 2018). In our experiment, the relatively low number of data points included in the analysis of looks to action vs. object video (46%) may be due to the conservative way we defined the regions of interest, such that they included just the coordinates of the videos themselves and a small amount of padding (150 pixels) on each side. Because of WebGazer's lower accuracy compared to in-lab eye-trackers, it may be preferable to define wider regions of interest—before beginning analysis—as in Yang & Krajbich (2021), who also replicated lab-based findings using WebGazer. This could help ensure that genuine looks to the region of interest are not excluded due to WebGazer's lower accuracy.

WebGazer was not suited to fine-grained temporal analysis at the time our study was conducted, with previous visual world replication studies finding 300—700 ms delays in the time that effects appeared compared to the original studies (Degen et al., 2021; Semmelmann & Weigelt, 2018; Slim & Hartsuiker, 2022). However, its temporal resolution is substantially improved in newer versions (Vos et al., 2022; Yang & Krajbich, 2021). Overall, it is encouraging that the results of the original paper replicated using the novel method of web-based eye-tracking, and we expect that future versions of the WebGazer software will continue to increase its suitability for behavioral research.

One limitation of the study design in Experiment 1 is that there are four training trials but only two filler trials. While the filler trials indicate that the speaker *can* talk about both actions and objects, it is still the case that the speaker in the verb condition is overall more likely to talk about actions, and the speaker in the noun condition is overall more likely to talk about objects. Thus, the design results in participants being directed to look more frequently at action (verb condition) or object (noun condition) videos during training. We aimed to eliminate this possible confound in Experiment 2.

Experiment 2

Having validated the method via replication of Havron et al. in Experiment 1, we sought to test the main hypothesis—that syntactic adaptation can support word learning—in English. To this end, we created a version of the study using the English syntactic frame *The girls/The girl's*. Like *La petite,* this frame can be followed by either a noun or a verb (e.g., *The girl's book* vs. *The girls sleep*). The cross-linguistic replication allowed us to test whether the adaptation effect observed in Experiment 1 would generalize to a new syntactic frame in a different language. If so, it would provide additional evidence for the role of syntactic adaptation as a general mechanism that can be drawn on during language learning.

A diagram of the trial structure for Experiments 2 and 3 is shown in Figure 4. We made several modifications to the study design that reduced the possible confounds and made it easier to run the study online. First, the trials used object and action images rather than videos, which simplified the task. In addition, we increased the number of test trials from three to four. We also increased the number of filler trials from two to four to match the number of training trials. This ensured that participants in the noun and verb conditions were not biased by looking at more images of the type that matched their condition (action for verb; object for noun) during the training phase. Now, participants were directed to look at equal numbers of action and object images

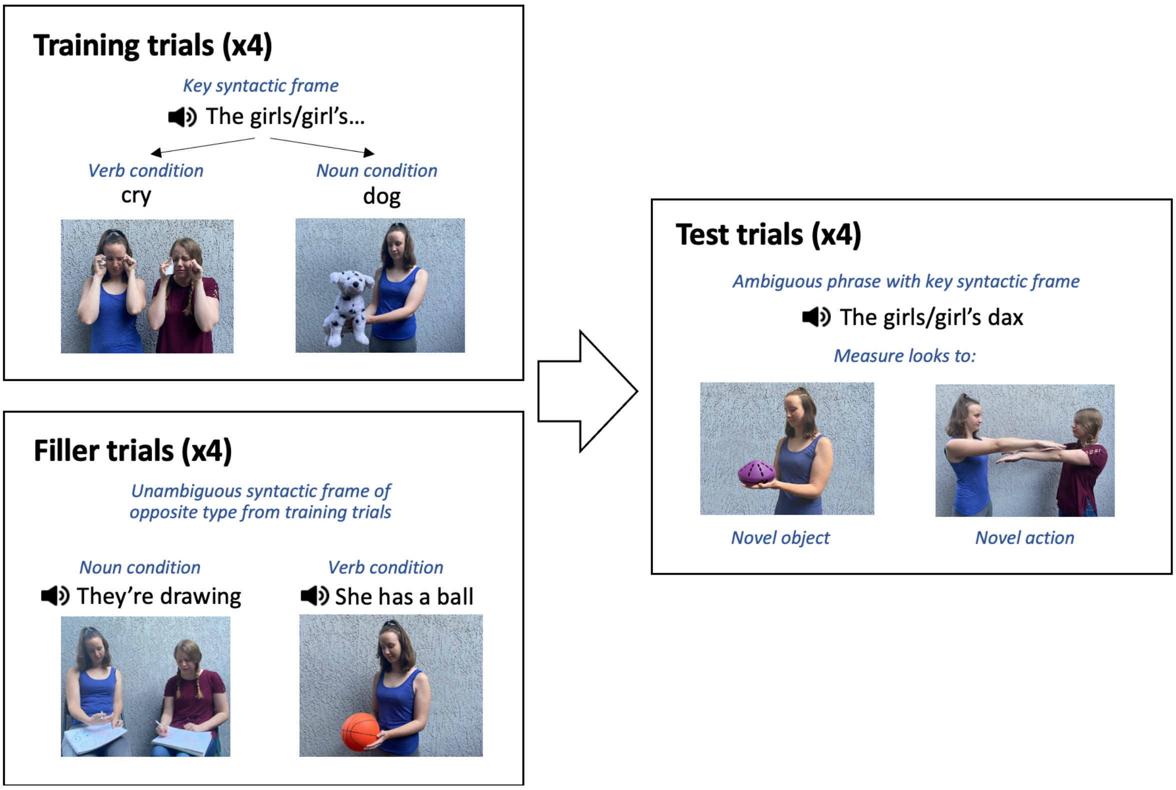


Figure 4. ***Diagram of experimental set-up for Experiments 2 and 3.***

during training trials; the only difference was in the type of linguistic content they heard following the key syntactic frame *The girls/girl's….* In the noun condition, participants heard *The girl's (noun)* on training trials, and in the verb condition, they heard *The girls (verb)* on training trials.

We also added a baseline condition to the study to examine whether participants would demonstrate bias toward looking at a particular image type even if they did not hear the structure *The girls/The girl's* at all before the test phase. In the baseline condition, participants' training trials included only the filler phrases used in both the noun and verb conditions (*They're Xing* in the noun condition and *She has an X* in the verb condition). Like the noun and verb conditions, the baseline condition was balanced so that participants would be directed to look at an equal number of action and object images. The inclusion of a baseline condition was an important step to take to investigate whether the adaptation effect appeared to occur in both the noun and the verb conditions, or whether it was primarily driven by participants in one condition.

We ensured that participants were not biased toward a particular interpretation by factors such as prosody by running an online norming experiment beforehand with 30 adult participants who were native English speakers. In the norming study, we played only the audio clips (such as *The girls/girl's dax*) and asked participants whether they thought the novel word referred to an action or an object. Participants judged that the novel words referred to actions 51.1% of the time, suggesting that the verb and noun interpretations were about equally plausible.

On the final trial, we directly asked participants to click on the image they thought the narrator was talking about. The image selection constituted an explicit measure of participants' comprehension of the phrase containing *The girls/The girl's,* in addition to the implicit evidence provided by eye-tracking. We added the explicit measure only on the final trial to avoid potential interference with participants' eye movements.

Method

Participants

We added an additional baseline condition for Experiment 2 and therefore recruited a larger total of 104 participants (57 female; 41 male; 6 other). Again, we collected data using Prolific and specified that participants had to speak English as their first language. They were randomly assigned to one of the three conditions (35 in the noun condition; 35 in the verb condition; 34 in the baseline condition).

Procedure

Besides the modifications described above, the experiment design was identical to Experiment 1. The number of trials was kept similar to Havron et al. (2019) due to limits in children's ability to maintain attention; the English version of the experiment lasted approximately fifteen minutes. Trial order was randomized, except that we did not allow more than two training or filler trials in a row. Image sides were counterbalanced.

Measures

Experiment 2 was carried out with WebGazer using the same measures as Experiment 1. WebGazer recorded 87% of looks as being directed toward the screen. Again, we analyzed only looks to the action image or the object image (62% of the total looks in the dataset).

Results

Proportion of Looks

Figure 2b shows the mean proportion of looks to the action image in each condition, with dots representing individual participants' mean proportions of looks. We included only looks after the onset of the ambiguous syntactic frame: *The g-...* in *The girls/girl's....* As in Experiment 1, participants in the verb condition (M = 0.596, SD = 0.193) were more likely to look at the action image than participants in the noun condition (M = 0.389, SD = 0.212). These effects were very similar in size to those observed in Experiment 1. The proportion of looks to the action image in the baseline condition (M = 0.435, SD = 0.187) fell in between the noun and verb condition, but the confidence interval for the baseline condition overlapped with the confidence interval for the noun condition (though not with the verb condition).

For Experiment 2, we compared the noun and verb conditions to the baseline condition. As in Experiment 1, we carried out a mixed effects linear regression which predicted the arc-sin transformed mean proportion of looks to the action image as a function of condition, with a random intercept for participant. In this model and all others for Experiments 2 and 3, condition was dummy-coded using the baseline condition as the reference. There was a significant main effect of condition such that participants in the verb condition looked more to the action image compared to participants in the baseline condition (β = 0.161, SE = 0.053, $p < 0.01$). However, there was not a significant difference between looks to the action image in the noun condition compared to the baseline condition (β = -0.05, SE = 0.051, p = 0.322).[4]

Time Course

To better understand at what time participants recruited their updated expectations, we plotted the time course of the mean proportion of looks to the action image, averaged across the four test trials, in Figure 3b. Specifically, we wished to know whether participants might begin looking at the action or object image even before hearing the full phrase *The girls/girl's [novel word]*. For instance, upon hearing *The g-*, participants could have realized that they were likely about to hear a sentence containing *The girls...* and could have drawn on their updated expectations to look at either the action or object image.

[4] Although the comparisons with the baseline condition are our primary statistical analyses, it may be of interest to directly examine the difference between the noun and verb conditions. In Experiment 2, participants in the verb condition looked at the action image significantly more than participants in the noun condition (β = 0.212, SE = 0.052, $p < 0.01$).

The time course plot reveals several interesting descriptive patterns. First, participants in the verb condition appeared more likely to look at the action image for almost the entire duration of the trial, even before hearing the key syntactic frame for the first time (*The girls/girl's [novel word]*). Participants in the baseline condition, on the other hand, were more likely to look at the object image slightly before the naming event occurred and throughout the trial. Finally, participants in the noun condition looked more at the object image than participants in the verb condition, and this effect appeared mostly after hearing the syntactic frame (*The girls/girl's*) for the first time. The pattern of results raises the question of whether participants were making anticipatory looks to the action image in the verb condition, and to the object image in the baseline condition, even before hearing the syntactic frame and the novel word.

The presence of anticipatory looks might raise the concern that the effects are not driven by interpretation of the sentences, but by something else—for instance, a preference for image type despite the equal number of filler and training trials. To address this, we conducted a post-hoc exploratory analysis examining whether there is a detectable change in looks before vs. after the linguistic event of interest: for each participants, on each trial, we calculated the mean difference in proportion of looks to the action image before the end of the audio *The g-* vs. during the rest of the trial. Figure 5a presents the mean difference in proportion of looks to the action image for each condition, with dots representing trial-level differences in proportions of looks across test trials. Then, we conducted an exploratory mixed effects regression analysis which predicted the difference in proportion of looks to the action image as a function of condition, with a random intercept for participant. There was a marginally significant difference between the proportion of looks for participants in the noun condition vs. the baseline condition ($\beta = -0.087$, $SE = 0.047$, $p = 0.068$), but no significant difference for participants in the verb condition vs. the baseline condition ($\beta = 0.06$, $SE = 0.048$, $p = 0.217$).

A likelihood ratio test between this model and a model without the effect of condition revealed an overall significant main effect of condition ($\chi 2(1) = 9.22$, $p < 0.01$), and the confidence intervals for the noun and verb conditions do not overlap. These results suggest that there was a difference in proportion of looks to the action image before vs. after the syntactic frame depending on participants' condition.[5] Therefore, while some of the difference between conditions may have been driven by initial image preferences, the time course provides evidence that participants' looking patterns

[5] In fact, in another post-hoc exploratory analysis where condition was recoded with the noun condition as the reference, participants in the verb condition had a significantly higher difference in proportion of looks to the action image than did participants in the noun condition ($\beta = 0.147$, $SE = 0.048$, $p < 0.01$).

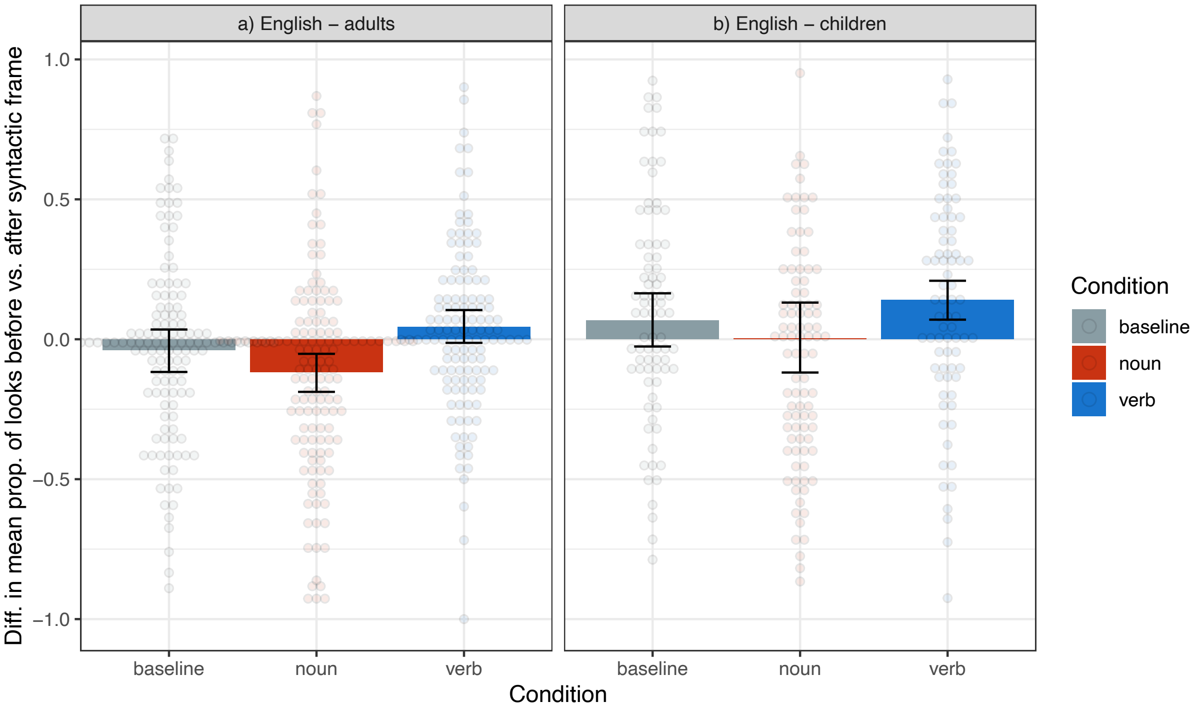


Figure 5. ***Mean overall difference in proportion of looks to the action image for a) Experiment 2 and b) Experiment 3. The difference is calculated by subtracting the proportion of looks before the end of "The g-" from the proportion of looks after the end of "The g-". Results are shown for the noun, verb, and baseline conditions during test trials, with bootstrapped confidence intervals. Semi-transparent dots show the distribution of trial-level data points (these are not by-participant averages).***

changed as the sentence unfolded. As shown in Figure 5a, the change was in the expected direction, with participants in the verb condition looking more at the action image and participants in the noun condition looking more at the object image.

Explicit Selection

The final trial of the experiment was identical to other test trials, but once it was completed, we directly asked participants to select the image they thought the narrator had described. There were large differences by condition, as shown in Figure 6. Participants in the baseline condition were about equally likely to select the action image (54.5%) or the object image (45.5%). In contrast, 85.7% of participants in the noun condition selected the object image, and 70.1% of participants in the verb condition selected the action image. To test these differences, we carried out post-hoc pairwise comparisons of the proportion of participants in each condition who selected the action image, using the Bonferroni adjustment for multiple comparisons. We found that compared to participants in the noun condition, those in the baseline condition (p < 0.01) and the verb condition (p < 0.01) were significantly more likely to select the action image. There was not a significant difference between the baseline and verb conditions (p = 0.81). Despite having selection data for only one trial, the difference between the noun and verb conditions is quite striking: In their explicit judgments about the meaning of a novel word, participants tended to interpret the word in line with the examples they had heard during training trials, which were presented in the same syntactic frame.

Discussion

The results in the verb and noun conditions of Experiment 2 were similar to those obtained in Experiment 1. Participants' mean proportion of looks to the action image was very similar in the verb (0.585 in Experiment 1 compared to 0.596 in Experiment 2) and noun (0.395 in Experiment 1 compared to 0.389 in Experiment 2) conditions. Again, this effect is not as large as the one observed by Havron et al. (2019), but the attenuation of effect size may be due to the noisiness of web-based eye-tracking. The rate of data loss was slightly lower than in Experiment 1, and we again defined regions of interest fairly conservatively; future experiments with WebGazer may wish to adjust this.

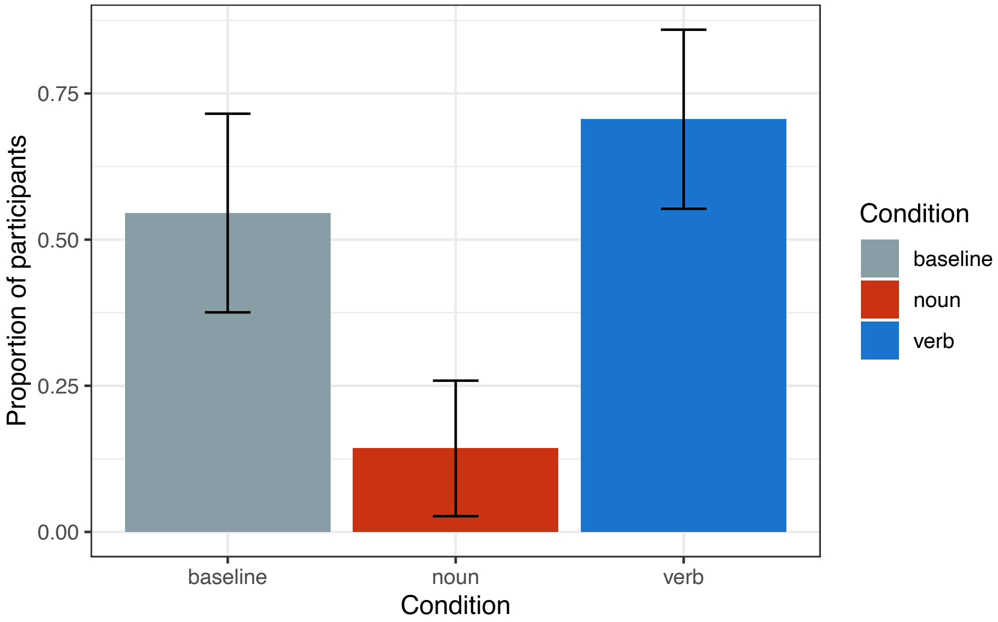


Figure 6. ***Proportion of participants in the baseline, noun, and verb conditions who selected the action image when explicitly asked to click on the image they thought the narrator was talking about.***

Participants' proportion of looks to the action image in the baseline condition fell in between that of the noun and verb conditions. However, based on the 95% confidence interval, which does not include 0.5, baseline participants appeared to show a slight preference for looking at the object image. This could be due to several factors. One possibility is that baseline participants were biased to think that *The girls/girl's X* was more likely to refer to an object image than an action image, either based on sentence prosody or on differences in the frequencies with which they hear the plural *The girls* and the possessive *The girl's…* preceding verbs vs. nouns.

To investigate the hypothesis that baseline participants were influenced by the distributions of the two structures, we conducted a corpus analysis using the Corpus of Contemporary American English (Davies, 2008), which draws from both speech and written text. In this analysis, we found 2,125 instances of *The girl's [noun]* and 1,013 instances of *The girls [verb]*. That is, the plural structure was half as frequent as the possessive structure. While these results align with baseline participants' preference for the object image, which matches the possessive *The girl's [noun]* interpretation of the structure, we have two reasons to doubt that baseline participants were drawing inferences about the meanings of the novel words.

First, the norming study we conducted before running the experiment did not find a preference for the noun or verb interpretation, suggesting that participants were not biased by prosody or by prior expectations about the meanings of the novel words. Second, the results we obtained using explicit selection on the final trial did not indicate that baseline participants were drawing inferences about the meanings of the novel words. Participants in the baseline condition performed essentially at chance when asked which image they thought the speaker was referring to, while a large majority (over 70%) of the participants in the noun and verb conditions selected the object image or the action image, respectively.

Thus, we favor a second possible explanation: participants in the baseline condition may have found the object images to be more salient or interesting. We consider this to be a plausible possibility because two other norming studies[6] found conflicting results regarding the salience of the object and action images. Participants in one norming study thought a speaker would be more likely to talk about the object images overall. However, in the second norming study, where images were matched on salience based on the results from the first study, participants thought a speaker would be more likely to refer to the action images overall. These findings suggest that participants' preferences related to the salience of the images are variable, and it is possible that participants in the baseline condition simply found the object images more

[6] More details about the procedure and analysis for these studies can be found in the norming section of the GitHub repository.

interesting than the action images.

On the whole, the results of Experiment 2 provide evidence that participants in the noun and verb conditions updated their expectations about whether the speaker was likely to follow *The girls/girl's* with a noun or a verb, while participants in the baseline condition maintained uncertainty.

Experiment 3

In Experiment 3, we extended the paradigm from Experiment 2 to ask whether three- to five-year-old English-speaking children would show similar patterns of syntactic adaptation during word learning. If children's behavior is similar to adults, it would support the proposal that adaptation is an important mechanism supporting child language acquisition.

Method

Participants

We collected data through the online Lookit platform (Scott & Schulz, 2017), where children can easily participate in looking-time experiments from home. There were 74 participants (42 female; 32 male). Children were assigned to the same three conditions as in Experiment 2 (27 in the noun condition; 23 in the verb condition; 24 in the baseline condition). We preregistered this smaller sample size compared to Experiments 1 and 2 primarily due to the greater difficulty of recruiting children online compared to adults; the sample size was similar to that of Havron et al. (2019). Children had to be native English speakers to be eligible for the study.

Procedure

Children either completed the study while sitting on their caregiver's lap, with the caregiver closing their eyes, or while seated on their own. The experiment procedure was nearly identical to Experiment 2, except that the instructions at the beginning of the study were made more child-friendly. We also added attention-getters at the beginning of each trial and took a calibration video of the child looking to the left and right sides of the screen, rather than using a 9-point automatic calibration. The trial structure was the same as in Experiment 2, and we maintained the same modifications to the Havron et al. procedure, implementing an equal number of filler and training trials and using image stimuli rather than videos.

Because a caregiver was not always present with the child, we designed the experiment to run by itself on a computer. As a result, we were not able to pause the experiment and ask children to explicitly select which image they thought the speaker was referring to.

Measures

Rather than using web-based eye-tracking, which proved to be noisy and frustrating for participants in pilot testing, we recorded videos of children through Lookit as they completed the study. The first author hand-coded the children's eye movements as being directed towards the left or right side of the screen. Coding was done blindly, without knowledge of the experimental condition a trial appeared in or which image appeared on which side of the screen.

Results

Proportion of Looks

The mean proportion of looks to the action image in each condition is shown in Figure 2c. Children in the verb condition (M = 0.629, SD = 0.17) were more likely to look at the action image than children in the noun condition (M = 0.481, SD = 0.187). The proportion of looks to the action image in the baseline condition (M = 0.597, SD = 0.175) fell in between the noun and verb condition. The confidence interval for the baseline condition overlapped with the confidence intervals for both the noun and verb conditions.

We repeated the analyses from Experiment 2: a mixed effects linear regression analysis predicted the arc-sin transformed mean proportion of looks to the action image as a function of condition, with random by-participant intercepts. There was a significant main effect of condition such that children in the noun condition looked less to the action image compared to children in the baseline condition (β = -0.173, SE = 0.063, $p < 0.01$). There was not a significant difference in looks between the verb condition and the baseline condition (β = 0.037, SE = 0.065, p = 0.572).[7]

Time Course

The time course of children's looks to the action image over time, averaged across the four test trials, is depicted in Figure 3c.

In contrast to the Experiment 2 adults, in Experiment 3, the children in all three

[7] Again, comparing the noun and verb conditions directly, children in the verb condition looked significantly more at the action image than did children in the noun condition ($\beta = 0.210$, $SE = 0.064$, $p < 0.01$).

conditions showed a slight preference for looking at the action image before hearing the key syntactic frame containing the novel word (e.g., *The girls/girl's dax*). This preference may have been due to the presence of two people in the action images, which could be more salient for children, compared to the presence of only one person in the object image. However, the time course indicates that shortly after the beginning of the ambiguous syntactic frame, *The g-*, children's looking patterns began to diverge. Children in the noun condition appeared to look consistently less at the action image than children in the verb condition. Children in the baseline condition fell in between the two, though they still showed a preference for the action image later in the trial.

As in Experiment 2, to determine the point in the trial at which these effects appeared, we performed a post-hoc exploratory analysis in which we calculated the mean difference in each participants' proportion of looks to the action image before the end of the audio *The g-* vs. during the rest of the trial. The results are illustrated in Figure 5b. We then used a mixed effects regression model to predict the difference in proportion of looks to the action image as a function of condition, with a random intercept for participant. The results showed a significant difference between the change in proportion of looks for participants in the noun condition compared to participants in the baseline condition (β = -0.165, *SE* = 0.068, *p* = 0.017). There was no significant difference between participants in the verb condition and those in the baseline condition (β = 0.022, *SE* = 0.07, *p* = 0.757). For children in the noun condition, who appeared to drive the effects in the results, there were changes in their eye movements over the course of the trial. As in Experiment 2, this provides evidence that children's looking preferences were updated as they recognized the familiar syntactic frame.

General Discussion

The three experiments reported here investigated whether syntactic adaptation is a mechanism implicated in word learning, as suggested by Havron et al. (2019). Experiment 1 was a direct replication of Havron et al. (2019) with French-speaking adults. Experiment 2 was a cross-linguistic replication with English-speaking adults and a novel syntactic frame. Experiment 3 was identical to Experiment 2, but with three- to five-year-old English-speaking children. All three experiments provided evidence that participants adapted to the usage of the syntactic frame they encountered. In the English experiments, participants in the noun condition had a stronger expectation that the speaker would use *The girl's [noun],* and participants in the verb condition had a stronger expectation that the speaker would use *The girls [verb].* These updated expectations then guided their interpretation of an ambiguous novel word presented in the same syntactic frame, such as *The girls/girl's dax.* Participants in the verb condition exhibited a preference for looking at the action image over the object image on test trials, and vice versa for participants in the noun condition. This effect was weaker in children than adults, but present in both groups.

Across experiments, the baseline condition also demonstrated variable results: English-speaking adults in the baseline condition appeared to show a preference for the object image, while English-speaking children in the baseline condition appeared to show a preference for the action image. However, participants in the baseline condition always showed a proportion of looks to the action image that fell in between the noun and verb conditions, as we would expect. In addition, the norming studies and the explicit selection task discussed in Experiment 2 provide evidence that adult participants in the baseline condition were not forming interpretations about the meanings of the ambiguous novel words.

Children, on the other hand, may have shown the opposite pattern from adults due to differences in their baseline expectations and preferences. Recall that in an adult corpus, we found about twice as many instances of *The girls [verb]* as *The girls [noun],* which aligned with baseline adults' preference for the object image. We carried out a second corpus analysis using the CHILDES corpus (MacWhinney, 2000) to examine child-directed speech, and found 10 instances of *The girl's [noun]* and 60 instances of *The girls [verb].* That is, the plural structure was six times as frequent as the possessive structure. We are reluctant to draw conclusions from such a small sample, but it may be possible that the plural structure is relatively more frequent compared to the possessive structure in child-directed language than it is in adult language (including written text). Thus, children's baseline preference for the action image may be the result of a baseline expectation for the observed signal to underlyingly have the plural structure. If so, children in the noun condition could be displaying stronger adaptation to the more surprising structure, and vice versa for adults in the verb condition, which would align with the results of Havron et al. (2019) as well as Jaeger & Snider's (2013) finding that the more unexpected primes have bigger priming effects.

However, visual saliency effects could also have influenced both child and adult looking patterns in the baseline condition. For instance, children may have found the action images more salient because they featured two people in them, while adults may have found the novel objects in the object images to be more salient, because they are more knowledgeable about the improbability of encountering such objects in everyday life (children might be more likely to see similarly strange-looking toys). Further examination of children's baseline expectations for structures and their visual saliency preferences is needed to determine whether either of these factors, or both, drives the differences in the baseline condition for adults and children.

In addition to the overall looking time analyses, exploratory time course analyses provided some evidence that participants in the noun and verb conditions adjusted their looking patterns as they listened to the sentence unfold. Both children and adults showed differences in looking patterns, in the expected directions, before vs.

after the onset of the novel word. These effects appeared to be driven by participants in the noun condition for both adults and children (though the comparison with the baseline condition for adults did not reach significance). Nonetheless, since WebGazer is not currently suitable for fine-grained temporal analysis, other methods are likely needed to shed more light upon the question of exactly when in the syntactic frame children and adults begin using their updated expectations to guide their interpretations.

Our results are similar to the key findings of Havron et al. (2019). One contribution of our work was the equal number of filler trials and training trials in Experiments 2 and 3. This modification ensured that participants heard the speaker refer to action and object images with equal frequency; it was only with the specific structure *The girls/girl's...* that participants developed an expectation about whether the speaker would use a noun or a verb. Thus, we can be confident that our results reflect adaptation to the usage of a particular linguistic structure and not to the speaker's general likelihood to talk about actions or objects. The adaptation effect then guided participants' interpretations of an ambiguous novel word that was presented in the same syntactic frame.

Another contribution of these experiments is that they demonstrate the feasibility of conducting eye-tracking studies through web-based platforms. Both WebGazer and Lookit are relatively new tools in the research community and are still undergoing development and expansion. However, both platforms have enormous potential in allowing eye-tracking studies—which have historically not been possible to conduct outside of research labs—to be carried out with larger and more diverse populations (Gosling et al., 2010). The fact that we replicated the findings of Havron et al. (2019) directly and cross-linguistically suggests that conducting studies on these platforms is viable for experiments such as this one, where looking time is computed over a large analysis window. With continuing improvements to the software, WebGazer may become suitable for even finer-grained spatial and temporal analyses (Semmelmann & Weigelt, 2018; Yang & Krajbich, 2021).

Overall, these results support and extend those of Havron et al. (2019). The similar findings across French and English, and between children and adults, lend support to the proposal that syntactic adaptation may be an important mechanism in both language processing and language acquisition. In fact, Havron et al. (2021a) tested whether syntactic adaptation might allow children to update their interpretation of familiar homophones by exposing them to repeated uses of either *La petite [noun]* or *La petite [verb].* When 3- to 4-year-olds then heard an ambiguous sentence such as *La petite ferme* (which could mean "The little farm" or "The little one is closing"), they tended to interpret the homophone as either a noun or a verb depending on which kind of sentences they had heard during training. The combination of these studies illustrates that syntactic adaptation can affect both familiar word processing and novel word learning.

More broadly, the results of the current studies add to the growing literature emphasizing the role of prediction in language acquisition (Babineau et al., 2022). Prior work has called into question whether prediction operates during children's language learning or only in mature processing (Rabagliati et al., 2016). Recent work has found evidence that children can use semantic information to predict upcoming linguistic content from 2 years old (Gambi et al., 2018), and 4- to 5-year-olds were shown to be able to adapt their interpretations of a sentence to rely more on syntactic or semantic information depending on which cue had previously been reliable (Beretti et al., 2020). These studies suggest that children are not only able to make the kinds of predictions that could support language learning, but also adapt the type of information they are drawing on to make those predictions. Other findings have provided support for the claim that linguistic prediction skills may be linked to general vocabulary development in infants and children (Gambi et al., 2021; Mani & Huettig, 2012; Ylinen et al., 2016).

Havron et al. (2019; 2021) and the cross-linguistic extension of their findings reported here contribute to this literature by demonstrating directly that children can update their syntactic predictions and recruit them during novel word learning. As noted by Babineau et al. (2022), however, future work must examine the extent to which prediction plays a role in infants' language acquisition, as some studies have not found such abilities in children 2 or younger (Havron et al., 2021a; 2021b; Gambi et al., 2018)—although this finding could also be due to infants lacking sufficient linguistic experience on which to base their predictions. If prediction is demonstrated to figure significantly in language learning from an early age, it may allow us to provide a more unified account of language acquisition and processing.

Regarding syntactic adaptation specifically and its relationship to word learning, open questions remain about adults' and children's baseline expectations of structure frequency, as well as how these preferences interact with new statistical information about a speaker's usage of syntax. Additional studies that carefully tease apart these factors will contribute to a formal model of expectation update during syntactic adaptation. Furthermore, this research has concentrated on French and English thus far (children's linguistic prediction skills have been studied in German, in Mani & Huettig, 2012, but not how they may adapt those predictions). However, other languages may contain even more frequent examples of ambiguous structures where syntactic adaptation could be useful in children's learning of novel words.

Future work should also further examine the specificity of syntactic adaptation in word-learning contexts. For instance, since we used the same speaker throughout the experiment, we do not know whether the adaptation effect is speaker-specific or whether it could generalize to other speakers and contexts (adult studies have found conflicting results: e.g., Kamide, 2012; Kroczek & Gunter, 2017; Lu et al., 2021; Schuster & Degen, 2019; Yildirim et al., 2016). In addition, future studies could vary the particular lexical content used within the syntactic structure (e.g., *The boys/boy's X*) to determine whether participants generalize their expectations about the underlying syntactic structure to a phrase with differing lexical content. If children are likely to encounter repeated syntactic structures in short bursts within specific contexts, as in the example where a caregiver utters similar phrases such as *The dog is running, The dog is playing,* etc., then we might expect syntactic adaptation to be relatively specific to the speaker and the lexical content. A deeper understanding of the mechanisms of syntactic adaptation in children, including the extent to which it is specific and cumulative, would allow us to examine whether it is in fact a form of error-based learning that could contribute to syntax acquisition (Chang et al., 2006). Moreover, it will be important to study how syntactic adaptation may take place in naturalistic contexts, where children are likely to repeat or respond to novel words that they hear and begin to use them in conversation right away, rather than hearing them repeated multiple times uninterrupted by a single speaker (Clark, 2007). During language acquisition, syntactic adaptation could be one of many tools that children can draw upon—along with speaker cues, prior knowledge, visual context, and more—as they rapidly learn new words.

While the role that syntactic adaptation, and prediction more broadly, plays in children's language learning merits further investigation, these three experiments provide evidence that children and adults can not only flexibly update their expectations about a speaker's syntactic preferences, but also draw on these expectations to guide novel word learning.

References

Alishahi, A., & Stevenson, S. (2008). A Computational Model of Early Argument Structure Acquisition. *Cognitive Science, 32*(5), 789–834. https://doi.org/10.1080/03640210801929287

Arciuli, J., & Simpson, I. C. (2011). Statistical learning in typically developing children: The role of age and speed of stimulus presentation. *Developmental Science, 14*(3), 464–473. https://doi.org/10.1111/j.1467-7687.2009.00937.x

Arciuli, J., & Torkildsen, J. (2012). Advancing Our Understanding of the Link between Statistical Learning and Language Acquisition: The Need for Longitudinal Data. *Frontiers in Psychology, 3*, 324. https://doi.org/10.3389/fpsyg.2012.00324

Arnon, I. (2019). Statistical Learning, Implicit Learning, and First Language Acquisition: A Critical Evaluation of Two Developmental Predictions. *Topics in Cognitive Science,* tops.12428. https://doi.org/10.1111/tops.12428

Babineau, M., Havron, N., Dautriche, I., de Carvalho, A., & Christophe, A. (2022). Learning to predict and predicting to learn: Before and beyond the syntactic bootstrapper. *Language Acquisition, 0*(0), 1–24. https://doi.org/10.1080/10489223.2022.2078211

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1). https://doi.org/10.18637/jss.v067.i01

Bencini, G., & Valian, V. (2008). Abstract sentence representations in 3-year-olds: Evidence from language production and comprehension. *Journal of Memory and Language, 59,* 97–113. https://doi.org/10.1016/j.jml.2007.12.007

Beretti, M., Havron, N., & Christophe, A. (2020). Four- and 5-year-old children adapt to the reliability of conflicting sources of information to learn novel words. *Journal of Experimental Child Psychology, 200,* 104927. https://doi.org/10.1016/j.jecp.2020.104927

Bernal, S., Lidz, J., Millotte, S., & Christophe, A. (2007). Syntax Constrains the Acquisition of Verb Meaning. *Language Learning and Development, 3,* 325–341. https://doi.org/10.1080/15475440701542609

Bernolet, S., & Hartsuiker, R. J. (2010). Does verb bias modulate syntactic priming? *Cognition, 114*(3), 455–461. https://doi.org/10.1016/j.cognition.2009.11.005

Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive Psychology, 18*(3), 355–387. https://doi.org/10.1016/0010-0285(86)90004-6

Bock, K., & Griffin, Z. M. (2000). The persistence of structural priming: Transient activation or implicit learning? *Journal of Experimental Psychology: General, 129*(2), 177–192. https://doi.org/10.1037/0096-3445.129.2.177

Booth, A. E., & Waxman, S. R. (2003). Mapping Words to the World in Infancy: Infants' Expectations for Count Nouns and Adjectives. *Journal of Cognition and Development, 4*(3), 357–381. https://doi.org/10.1207/S15327647JCD0403_06

Boyland, J. T., & R. Anderson, J. (1998). Evidence that Syntactic Priming is Long Lasting. *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society,* 1205. https://doi.org/10.1184/R1/6614720.v1

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition, 106*(2), 707–729. https://doi.org/10.1016/j.cognition.2007.04.005

Branigan, H. P., & Messenger, K. (2016). Consistent and cumulative effects of syntactic experience in children's sentence production: Evidence for error-based implicit learning. *Cognition, 157,* 250–256. https://doi.org/10.1016/j.cognition.2016.09.004

Branigan, H. P., Pickering, M. J., & Cleland, A. A. (2000). Syntactic co-ordination in dialogue. *Cognition, 75*(2), B13–B25. https://doi.org/10.1016/S0010-0277(99)00081-5

Branigan, H. P., Pickering, M. J., McLean, J. F., & Cleland, A. A. (2007). Syntactic alignment and participant role in dialogue. *Cognition, 104*(2), 163–197. https://doi.org/10.1016/j.cognition.2006.05.006

Brooks, P. J., & Tomasello, M. (1999). Young children learn to produce passives with nonce verbs. *Developmental Psychology, 35*(1), 29–44. https://doi.org/10.1037/0012-1649.35.1.29

Brown, R. W. (1957). Linguistic determinism and the part of speech. *The Journal of Abnormal and Social Psychology, 55*(1), 1–5. https://doi.org/10.1037/h0041199

Brusini, P., Seminck, O., Amsili, P., & Christophe, A. (2021). The Acquisition of Noun and Verb Categories by Bootstrapping From a Few Known Words: A Computational Model. *Frontiers in Psychology, 12*. https://www.frontiersin.org/articles/10.3389/fpsyg.2021.661479

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review, 113*(2), 234–272. https://doi.org/10.1037/0033-295X.113.2.234

Christodoulopoulos, C., Roth, D., & Fisher, C. (2016). An incremental model of syntactic bootstrapping. *Proceedings of the 7th Workshop on Cognitive Aspects of Computational Language Learning,* 38–43. https://doi.org/10.18653/v1/W16-1906

Christophe, A., Dautriche, I., de Carvalho, A., & Brusini, P. (2016, December 1). *Bootstrapping the Syntactic Bootstrapper.*

Clark, E. V. (2007). Young children's uptake of new words in conversation. *Language in Society, 36*(2), 157–182. https://doi.org/10.1017/S0047404507070091

Cleland, A. A., & Pickering, M. J. (2003). The use of lexical and syntactic information in language production: Evidence from the priming of noun-phrase structure. *Journal of Memory and Language, 49*(2), 214–230. https://doi.org/10.1016/S0749-596X(03)00060-3

Davies, M. (2008). *The Corpus of Contemporary American English (COCA). Available online at https://www.english-corpora.org/coca/.*

De Kloe, Y. J. R., Hooge, I. T. C., Kemner, C., Niehorster, D. C., Nyström, M., & Hessels, R. S. (2022). Replacing eye trackers in ongoing studies: A comparison of eye-tracking data quality between the Tobii Pro TX300 and the Tobii Pro Spectrum. *Infancy, 27*(1), 25–45. https://doi.org/10.1111/infa.12441

Degen, J., Kursat, L., & Leigh, D. (2021). *Seeing is believing: Testing an explicit linking assumption for visual world eye-tracking in psycholinguistics* [Unpublished]. https://github.com/thegricean/eyetracking_replications/blob/b99241562dc6a1eea603b67e7da5a33fbc25bed0/writing/2021_cogsci/sunbrehenyreplication.pdf

Fazekas, J., Jessop, A., Pine, J., & Rowland, C. (2020). Do children learn from their prediction mistakes? A registered report evaluating error-based theories of language acquisition. *Royal Society Open Science, 7*(11), 180877. https://doi.org/10.1098/rsos.180877

Fine, A. B., & Jaeger, T. F. (2013). Evidence for Implicit Learning in Syntactic Comprehension. *Cognitive Science, 37*(3), 578–591. https://doi.org/10.1111/cogs.12022

Fine, A. B., & Jaeger, T. F. (2016). The role of verb repetition in cumulative structural priming in comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(9), 1362–1376. https://doi.org/10.1037/xlm0000236

Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid Expectation Adaptation during Syntactic Comprehension. *PLOS ONE, 8*(10), e77661. https://doi.org/10.1371/journal.pone.0077661

Fisher, C. (1994). Structure and meaning in the verb lexicon: Input for a syntax-aided verb learning procedure. *Language and Cognitive Processes, 9*(4), 473–517. https://doi.org/10.1080/01690969408402129

Fisher, C., Jin, K., & Scott, R. M. (2020). The developmental origins of syntactic bootstrapping. *Topics in Cognitive Science, 12*(1), 48–77. https://doi.org/10.1111/tops.12447

Gambi, C., Gorrie, F., Pickering, M., & Rabagliati, H. (2018). The development of linguistic prediction: Predictions of sound and meaning in 2- to 5-year-olds. *Journal of Experimental Child Psychology, 173*. https://doi.org/10.1016/j.jecp.2018.04.012

Gambi, C., Jindal, P., Sharpe, S., Pickering, M. J., & Rabagliati, H. (2021). The Relation Between Preschoolers' Vocabulary Development and Their Ability to Predict and Recognize Words. *Child Development, 92*(3), 1048–1066. https://doi.org/10.1111/cdev.13465

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition: A Journal of Developmental Linguistics, 1*(1), 3–55. https://doi.org/10.1207/s15327817la0101_2

Gleitman, L. R., Cassidy, K., Nappa, R., Papafragou, A., & Trueswell, J. C. (2005). Hard Words. *Language Learning and Development, 1*(1), 23–64. https://doi.org/10.1207/s15473341lld0101_4

Gomez, R. L., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition, 70*, 109–135. https://doi.org/10.1016/S0010-0277(99)00003-7

Gosling, S. D., Sandy, C. J., John, O. P., & Potter, J. (2010). Wired but not WEIRD: The promise of the Internet in reaching more diverse samples. *Behavioral and Brain Sciences, 33*(2–3), 94–95. https://doi.org/10.1017/S0140525X10000300

Gries, S. Th. (2005). Syntactic Priming: A Corpus-based Approach. *Journal of Psycholinguistic Research, 34*(4), 365–399. https://doi.org/10.1007/s10936-005-6139-3

Harrington Stack, C. M., James, A. N., & Watson, D. G. (2018). A failure to replicate rapid syntactic adaptation in comprehension. *Memory & Cognition, 46*(6), 864–877. https://doi.org/10.3758/s13421-018-0808-6

Havron, N., Babineau, M., Fiévet, A.-C., de Carvalho, A., & Christophe, A. (2021). Syntactic Prediction Adaptation Accounts for Language Processing and Language Learning. *Language Learning, 71*(4), 1194–1221. https://doi.org/10.1111/lang.12466

Havron, N., Scaff, C., Carbajal, M. J., Linzen, T., Barrault, A., & Christophe, A. (2020). Priming syntactic ambiguity resolution in children and adults. *Language, Cognition and Neuroscience, 35*(10), 1445–1455. https://doi.org/10.1080/23273798.2020.1797130

He, A. X., & Lidz, J. (2017). Verb Learning in 14- and 18-Month-Old English-Learning Infants. *Language Learning and Development, 13*(3), 335–356. https://doi.org/10.1080/15475441.2017.1285238

Holmqvist, K. (2017). *Common predictors of accuracy, precision and data loss in 12 eye-trackers*. https://doi.org/10.13140/RG.2.2.16805.22246

Huang, Y. T., & Arnold, A. R. (2016). Word learning in linguistic context: Processing and memory effects. *Cognition, 156,* 71–87. https://doi.org/10.1016/j.cognition.2016.07.012

Huttenlocher, J., Vasilyeva, M., & Shimpi, P. (2004). Syntactic priming in young children. *Journal of Memory and Language, 50*(2), 182–195. https://doi.org/10.1016/j.jml.2003.09.003

Jaeger, T. F., & Snider, N. E. (2013). Alignment as a consequence of expectation adaptation: Syntactic priming is affected by the prime's prediction error given both prior and recent experience. *Cognition, 127*(1), 57–83. https://doi.org/10.1016/j.cognition.2012.10.013

Jin, K., & Fisher, C. (2014). *Early evidence for syntactic bootstrapping: 15-month-olds use sentence structure in verb learning*. https://www.semanticscholar.org/paper/Early-evidence-for-syntactic-bootstrapping-%3A-use-in-Jin-Fisher/3f668d2daf85ac65e8adf15abd4fff0ef72cbfd0

Kamide, Y. (2012). Learning individual talkers' structural preferences. *Cognition, 124*(1), 66–71. https://doi.org/10.1016/j.cognition.2012.03.001

Kaschak, M. P. (2007). Long-term structural priming affects subsequent patterns of language production. *Memory & Cognition, 35*(5), 925–937. https://doi.org/10.3758/BF03193466

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review, 122*(2), 148–203. https://doi.org/10.1037/a0038695

Kleinschmidt, D., Fine, A., & Jaeger, T. F. (2012, January 1). *A belief-updating model of adaptation and cue combination in syntactic comprehension*. Proceedings of the 34rd Annual Meeting of the Cognitive Science Society (CogSci12).

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language, 56*(1), 1–15. https://doi.org/10.1016/j.jml.2006.07.010

Kroczek, L. O. H., & Gunter, T. C. (2017). Communicative predictions can overrule linguistic priors. *Scientific Reports, 7*(1), Article 1. https://doi.org/10.1038/s41598-017-17907-9

Krogh, L., Vlach, H., & Johnson, S. (2013). Statistical Learning Across Development: Flexible Yet Constrained. *Frontiers in Psychology, 3*. https://www.frontiersin.org/article/10.3389/fpsyg.2012.00598

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software, 82*(13). https://doi.org/10.18637/jss.v082.i13

Liu, L., Burchill, Z., Tanenhaus, M. K., & Jaeger, T. F. (2017). Failure to replicate talker-specific syntactic adaptation. *Proceedings of the 39th Annual Meeting of the Cognitive Science Society,* 2616–2621.

Long, D. L., & Prat, C. S. (2008). Individual differences in syntactic ambiguity resolution: Readers vary in their use of plausibility information. *Memory & Cognition, 36*(2), 375–391. https://doi.org/10.3758/MC.36.2.375

Lu, J., Lassiter, D., & Degen, J. (2021). Syntactic satiation is driven by speaker-specific adaptation. *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society.*

MacWhinney, B. (2000). *The Childes Project: Tools for Analyzing Talk. Third Edition* (3rd ed.). Lawrence Erlbaum Associates. https://doi.org/10.4324/9781315805641

Madsen, J., Júlio, S. U., Gucik, P. J., Steinberg, R., & Parra, L. C. (2021). Synchronized eye movements predict test scores in online video education. *Proceedings of the National Academy of Sciences, 118*(5), e2016980118. https://doi.org/10.1073/pnas.2016980118

Mani, N., & Huettig, F. (2012). Prediction during language processing is a piece of cake—But only for skilled producers. *Journal of Experimental Psychology: Human Perception and Performance, 38*(4), 843–847. https://doi.org/10.1037/a0029284

Messenger, K., Branigan, H., & McLean, J. (2011). Evidence for (shared) abstract structure underlying children’s short and full passives. *Cognition, 121,* 268–274. https://doi.org/10.1016/j.cognition.2011.07.003

Messenger, K., Yuan, S., & Fisher, C. (2015). Learning verb syntax via listening: New evidence from 22-month-olds. *Language Learning and Development, 11*(4), 356–368. https://doi.org/10.1080/15475441.2014.978331

Naigles, L. (1990). Children use syntax to learn verb meanings. *Journal of Child Language, 17*(2), 357–374. https://doi.org/10.1017/S0305000900013817

Open Science Collaboration. (2015). *Estimating the reproducibility of psychological science.* https://www.science.org/doi/10.1126/science.aac4716

Ostrand, R., & Ferreira, V. S. (2019). Repeat after us: Syntactic alignment is not partner-specific. *Journal of Memory and Language, 108,* 104037. https://doi.org/10.1016/j.jml.2019.104037

Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J., & Hays, J. (2016). WebGazer: Scalable Webcam Eye Tracking Using User Interactions. *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI),* 3839–3845.

Peter, M., Chang, F., Pine, J., Blything, R., & Rowland, C. (2015). When and how do children develop knowledge of verb argument structure? Evidence from verb bias effects in a structural priming task. *Journal of Memory and Language, 81,* 1–15. https://doi.org/10.1016/j.jml.2014.12.002

Pickering, M. J., & Ferreira, V. S. (2008). Structural priming: A critical review. *Psychological Bulletin, 134*(3), 427–459. https://doi.org/10.1037/0033-2909.134.3.427

Pickering, M. J., & Garrod, S. (2004). Toward a Mechanistic Psychology of Dialogue. *Behavioral and Brain Sciences, 27*(2), 169–190. https://doi.org/10.1017/S0140525X04000056

Prasad, G., & Linzen, T. (2021). Rapid syntactic adaptation in self-paced reading: Detectable, but only with many participants. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 47*(7), 1156–1172. https://doi.org/10.1037/xlm0001046

R Core Team. (2021). *R: A language and environment for statistical computing.* [Computer software]. https://www.R-project.org/

Rabagliati, H., Gambi, C., & Pickering, M. J. (2016). Learning to predict or predicting to learn? *Language, Cognition and Neuroscience, 31*(1), 94–105. https://doi.org/10.1080/23273798.2015.1077979

Raviv, L., & Arnon, I. (2018). The developmental trajectory of children's auditory and visual statistical learning abilities: Modality-based differences in the effect of age. *Developmental Science, 21*(4), e12593. https://doi.org/10.1111/desc.12593

Romberg, A. R., & Saffran, J. (2010). Statistical learning and language acquisition. *WIREs Cognitive Science, 1*(6), 906–914. https://doi.org/10.1002/wcs.78

Rowland, C. F., Chang, F., Ambridge, B., Pine, J. M., & Lieven, E. V. M. (2012). The development of abstract syntax: Evidence from structural priming and the lexical boost. *Cognition, 125*(1), 49–63. https://doi.org/10.1016/j.cognition.2012.06.008

Ryskin, R., Kurumada, C., & Brown-Schmidt, S. (2019). Information Integration in Modulation of Pragmatic Inferences During Online Language Comprehension. *Cognitive Science, 43*(8), e12769. https://doi.org/10.1111/cogs.12769

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

Saffran, J., & Kirkham, N. (2017). Infant Statistical Learning. *Annual Review of Psychology, 69*, 1–23. https://doi.org/10.1146/annurev-psych-122216-011805

Schuster, S., & Degen, J. (2020). I know what you're probably going to say: Listener adaptation to variable use of uncertainty expressions. *Cognition, 203*, 104285. https://doi.org/10.1016/j.cognition.2020.104285

Schuster, S., & Degen, J. (2019). Speaker-specific adaptation to variable use of uncertainty expressions. *Proceedings of the 41st Annual Meeting of the Cognitive Science Society.*

Scott, K., & Schulz, L. (2017). Lookit (Part 1): A New Online Platform for Developmental Research. *Open Mind.* https://doi.org/10.1162/OPMI_a_00002

Semmelmann, K., & Weigelt, S. (2018). Online webcam-based eye tracking in cognitive science: A first look. *Behavior Research Methods, 50*(2), 451–465. https://doi.org/10.3758/s13428-017-0913-7

Shimpi, P. M., Gámez, P. B., Huttenlocher, J., & Vasilyeva, M. (2007). Syntactic priming in 3- and 4-year-old children: Evidence for abstract representations of transitive and dative forms. *Developmental Psychology, 43*(6), 1334–1346. https://doi.org/10.1037/0012-1649.43.6.1334

Shufaniya, A., & Arnon, I. (2018). Statistical Learning Is Not Age-Invariant During Childhood: Performance Improves With Age Across Modality. *Cognitive Science, 42*(8), 3100–3115. https://doi.org/10.1111/cogs.12692

Slim, M. S., & Hartsuiker, R. J. (2022). Moving visual world experiments online? A web-based replication of Dijkgraaf, Hartsuiker, and Duyck (2017) using PCIbex and WebGazer.js. *Behavior Research Methods*. https://doi.org/10.3758/s13428-022-01989-z

Thothathiri, M., & Snedeker, J. (2008). Syntactic priming during language comprehension in three- and four-year-old children. *Journal of Memory and Language, 58*(2), 188–213. https://doi.org/10.1016/j.jml.2007.06.012

Vos, M., Minor, S., & Ramchand, G. C. (2022). Comparing infrared and webcam eye tracking in the Visual World Paradigm. *Glossa Psycholinguistics, 1*(1). https://doi.org/10.5070/G6011131

Waxman, S. R. (1999). The dubbing ceremony revisited: Object naming and categorization in infancy and early childhood. In *Folkbiology* (pp. 233–284). The MIT Press. https://doi.org/10.1002/wcs.150

Waxman, S. R., Lidz, J. L., Braun, I. E., & Lavin, T. (2009). Twenty-four-month-old infants' interpretations of novel verbs and nouns in dynamic scenes. *Cognitive Psychology, 59*(1), 67–95. https://doi.org/10.1016/j.cogpsych.2009.02.001

Wells, J. B., Christiansen, M. H., Race, D. S., Acheson, D. J., & MacDonald, M. C. (2009). Experience and sentence processing: Statistical learning and relative clause comprehension. *Cognitive Psychology, 58*(2), 250–271. https://doi.org/10.1016/j.cogpsych.2008.08.002

Yang, X., & Krajbich, I. (2021). Webcam-based online eye-tracking for behavioral research. *Judgment and Decision Making, 16*(6), 1485–1505. https://doi.org/10.1017/S1930297500008512

Yildirim, I., Degen, J., Tanenhaus, M. K., & Jaeger, T. F. (2016). Talker-specificity and adaptation in quantifier interpretation. *Journal of Memory and Language, 87*, 128–143. https://doi.org/10.1016/j.jml.2015.08.003

Ylinen, S., Bosseler, A., Junttila, K., & Huotilainen, M. (2016). Predictive coding accelerates word recognition and learning in the early stages of language development. *Developmental Science, 20*. https://doi.org/10.1111/desc.12472

Yuan, S., & Fisher, C. (2009). “Really? She blicked the baby?”: two-year-olds learn combinatorial facts about verbs by listening. *Psychological Science, 20*(5), 619–626. https://doi.org/10.1111/j.1467-9280.2009.02341.x

Yurovsky, D., Case, S., & Frank, M. C. (2017). Preschoolers Flexibly Adapt to Linguistic Input in a Noisy Channel. *Psychological Science, 28*(1), 132–140. https://doi.org/10.1177/0956797616668557

Data, code and materials availability statement

All stimuli, data, and analyses for all three experiments can be found at: https://github.com/eswanson166/syntactic-adaptation-and-word-learning.

Ethics statement

Data collection for these studies was approved by the Stanford Institutional Review Board (IRB), protocol 19960.

Authorship and contributorship statement

ES proposed the project, carried out data collection and analysis, and wrote the first draft of the manuscript. MCF and JD helped design the studies, provided advice on data collection and analysis, and reviewed and edited the manuscript. All authors approved the final version of the manuscript and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Acknowledgements

We thank Naomi Havron for her support and helpful discussion, as well as an anonymous reviewer for their comments. We would also like to thank everyone at the Language and Cognition Lab and the ALPS lab. Finally, we are grateful to the participants and families who took part in this research through Prolific and Lookit.

Supplementary materials

Mixed Effects Logistic Regression Analysis

The main analysis reported in the paper is the mixed effects linear regression predicting the arc-sin transformed proportion of looks to the action image during a trial. Across all three experiments, we also preregistered a mixed effects logistic regression analysis directly predicting individual looks to the action image. We chose to carry out both analyses because each has advantages and drawbacks, with proportion of looks collapsing information about individual looks, while models of raw looks may not fully account for correlations between neighboring looks (though we did include previous look as a predictor). However, converging evidence from these two models would provide promising support for the hypothesis. Indeed, the results of the models agreed across all three experiments, so to save space, we only reported the linear regression on arc-sin transformed proportion of looks in the final paper. The results of the logistic regression on individual looks are summarized below.

Experiment 1

In Experiment 1, the mixed effects logistic regression predicted the log odds of looking to the action video as a function of condition and previous look (to the action video or not). It included a random intercept for participant and a random slope that accounted for participant differences in the effect of previous look. There was a significant main effect of condition ($\beta = 0.855$, SE = 0.175, $p < 0.001$), such that participants in the verb condition were more likely to look at the action video. There was also a significant main effect of previous look ($\beta = 4.761$, SE = 0.225, $p < 0.001$) such that if a participant looked at the action video on their previous look, they were more likely to look at the action video on the following look as well.

Experiment 2

Similarly, in Experiment 2, the mixed effects logistic regression analysis directly predicted the log odds of looking to the action image as a function of condition and previous look. It included random by-participant intercepts and a random by-participant slopes for previous look. Participants in the noun condition were marginally less likely to look at the action image compared to those in the baseline condition ($\beta = -0.395$, SE = 0.23, $p < 0.09$), while participants in the verb condition were marginally more likely to look at the action image ($\beta = 0.441$, SE = 0.231, $p < 0.06$). There was also a significant main effect of previous look ($\beta = 4.96$, SE = 0.245, $p < 0.001$) such that participants were more likely to look at the action image if their previous look was to the action image.

Experiment 3

In Experiment 3, the mixed effects logistic regression model again directly predicted the log odds of looking to the action image as a function of condition and previous look. As before, we included random by-participant intercepts and random by-participant slopes for previous look. This model also revealed a significant effect of condition, such that children in the noun condition were less likely than children in the baseline condition to look at the action image (β = -0.263, SE = 0.101, p < 0.01). There was not a significant effect for children in the verb condition compared to children in the baseline condition (β = 0.138, SE = 0.106, p = 0.194). The effect of previous look was significant, such that if a child's look on the previous sample was towards the action image, their current look was also more likely to be directed towards the action image (β = 7.27, SE = 0.087, p < 0.001).

Exploratory Generalization Trial

In Experiment 1, following Havron et al., we included an exploratory generalization trial. On this trial, participants heard an ambiguous structure using the masculine *Le petit…* frame rather than the feminine *La petite…* that had appeared during training trials. Results indicated that participants in the verb condition looked significantly more to the action video than participants in the noun condition (β = 0.195, *SE* = 0.085, *p* = 0.024). Although we should be cautious given that it is based on a single trial, this finding suggests that syntactic adaptation may generalize to slightly different structures. Though this question was outside the scope of Experiments 2 and 3, it merits further investigation to determine the extent to which syntactic adaptation is structure-specific.

License

Children’s developing conversational and reading inference skills: a call for a collaborative approach

Elspeth Wilson
University of Cambridge, UK

Kate Cain
Lancaster University, UK

Catherine Davies
University of Leeds, UK

Jenny Gibson
University of Cambridge, UK

Holly Joseph
University of Reading, UK

Ludovica Serratrice
University of Reading, UK
UiT The Arctic University of Norway, Norway

Margreet Vogelzang
University of Cambridge, UK

Abstract: In this perspectives article, we call for a collaborative approach to research on children’s development of conversational inferences and of reading inferences. Despite the clear commonalities in their focus, the two rich research traditions have remained almost entirely separate, primarily within the fields of Developmental Psychology and Experimental Pragmatics, on the one hand, and Cognitive, Developmental and Educational Psychology on the other. We briefly survey research on conversational and reading inferences, and show how both similarities and differences in theoretical approach, methodologies and findings raise significant questions, including: What effect does both context (conversation or reading) and modality (oral, visual, written) have on the need for children to make inferences, and for the opportunities for them to learn to do so? And how do linguistic and background knowledge, socio-cognitive and environmental factors support different inferences across contexts and modalities? We propose that a collaborative agenda is timely and crucial for interdisciplinary work. Researchers need to develop theoretical models of how different types of inference cluster together and are supported or affected by the context, modality, and other linguistic, socio-cognitive and environmental factors. They must also develop methodologies which enable reliable and valid measures of inferencing ability that can capture quantitative and qualitative changes across development. Ultimately this will contribute to better understanding children’s pragmatic development, as well as teaching and intervention practices in communication and reading comprehension.

Keywords: pragmatic development; pragmatic inference; reading comprehension; reading inference.

Corresponding author(s): Elspeth Wilson, Faculty of Education, University of Cambridge, CB2 8PQ, UK. Email: elspeth@elspethwilson.uk.

ORCID IDs:
Elspeth Wilson: 0000-0001-6114-1294
Kate Cain: 0000-0003-2780-188X
Catherine Davies: 0000-0001-9347-7905
Jenny Gibson: 0000-0002-6172-6265
Holly Joseph: 0000-0003-4325-4628
Ludovica Serratrice: 0000-0001-5141-6186
Margreet Vogelzang: 0000-0003-2811-5419

Citation: Wilson E., Cain, K., Davies, C., Gibson, J., Joseph, H., Serratrice, L., & Vogelzang, M. (2024). Children’s developing conversational and reading inference skills: a call for a collaborative approach. *Language Development Research*, *4*(1), 42–79. https://doi.org/10.34842/89as-7d31

Introduction

Accumulating evidence highlights the importance of pragmatic inferences for conversation and reading comprehension skills (Bohn & Frank, 2019; Matthews, 2014; Oakhill, 2020; O'Brien et al., 2015). For instance, children have to learn what a speaker means when they say, 'My sister's a hedgehog', or how two sentences relate in a text such as: 'There was a loud crash in the kitchen; "Where is the dustpan and brush?", asked Ben'. Pragmatic language skills, broadly defined, are crucial not only for successful communication and comprehension (Cain & Oakhill, 1999; Nation, 2005; Norbury & Bishop, 2002), but also for building peer relationships and socio-emotional and behavioural development across childhood and adolescence (Conti-Ramsden et al., 2019; Coplan & Weeks, 2009; Helland et al., 2014; Mok et al., 2014; St Clair et al., 2011). In addition, reading comprehension in particular enables access to learning materials and contributes to educational and employment outcomes (OECD, 2019; World Literacy Foundation, 2018).

Developmental research on conversational inferences and reading inferences has remained almost entirely separate, despite their common focus. Pragmatic inferences in conversation have largely been studied within the fields of Developmental Psychology and Experimental Pragmatics (a branch of Linguistics), while reading inferences have primarily been investigated within the domains of Educational, Developmental and Cognitive Psychology. Our working assumption, though, is that just as a child might encounter a new word in conversation, and then extend their understanding from a book, or vice versa, so too when they learn to understand ironic utterances, resolve anaphoric reference or derive any kind of inference in one context (conversing or reading), they will likely be able to call on and develop these skills in the other context, notwithstanding some interesting differences which we will discuss. Modality – whether the language is oral (or visual in the case of sign languages) or written – is a dimension that actually cuts across these contexts: while conversational inferences broadly align with oral language, and reading inferences with written language, there is no neat mapping, in research programmes or in the real-world. For instance, studies on reading inferences might involve *listening* to texts, just as children listen to books in the context of shared book reading; studies on conversational inferences may present short utterances without discourse context, possibly in written modality.

In this perspectives article, we therefore aim to highlight commonalities and differences in research findings about children's development of inference skills across contexts in conversation and reading; to show how these commonalities and differences raise some fundamental questions about the development of inferencing; and to set out a collaborative agenda for future research. As authors we have taken the first step in this collaboration, combining our expertise as theoretical and experimental linguists, developmental and cognitive psychologists, speech and language therapists and educators, who have researched either conversational or reading inferences, or both. We intentionally take a broad view in our survey of the state of the art, bringing in research on a range of inferences across a range of ages, based on a range of theoretical frameworks. And we intentionally raise more questions than answers – to show how integrating current research on inferencing reveals directions to address outstanding issues in future research.

Deepening our empirical understanding and honing our theoretical models of inferencing development will ultimately contribute to more effective teaching and intervention practices. In England, "making inferences on the basis of what is being said and done" is set out in the National Curriculum as a requirement for teaching reading comprehension from age 5 (Key Stage 1), as is, more generally, "listen[ing] and respond[ing] appropriately to adults and their peers" (Department for Education, 2013). Similarly, for instance, making inferences from texts is also part of the US Common Core, while formulating ideas about the author's intention is an aim for readers in primary school in Saxony, Germany (National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010; Staatsministerium für Kultus Freistaat Sachsen, 2019). Collaboration and alignment between academic researchers in the fields mentioned above, and crucially, with educational and clinical practitioners, is essential for optimal support of children's language and literacy development (Davies et al., 2022, 2023). In addition, inferencing may be the target of intervention with particular clinical populations, such as children with Developmental Language Disorder or autistic children (Adams et al., 2009; Bishop et al., 2016; Dawes et al., 2019; Matthews et al., 2018).

First, we summarise the state of the art in the study of conversational and reading inferences and their development. Then, we reflect on the similarities and differences between findings on inference development across research traditions – similarities in some conceptions of pragmatic inference and an increased focus on factors associated with developing inference ability, and differences in motivations for research, inference types, methods, and theory. This review highlights some fundamental questions which still need to be addressed to develop both theory and practice: how modality affects the learning of inferences; which inferences children learn over development; and how cognitive, psychological and environmental factors influence their development. Finally, we identify some promising ways forward by laying out a collaborative research agenda.

Conversational inferences

Research on children's development of pragmatic inferences has investigated a range of inference types, including quantity implicatures (for reviews see Papafragou & Skordos, 2016; E. Wilson & Katsos, 2020), relevance implicatures (E. Wilson & Katsos, 2022), metaphor (Pouscoulous & Tomasello, 2020), metonymy (Köder & Falkum, 2020), irony (Filippova, 2014; Köder & Falkum, 2021; Zajączkowska et al., 2020), genericity (Lazaridou-Chatzigoga et al., 2019), reference and anaphora resolution (Davies & Kreysa, 2018; Rabagliati & Robertson, 2017; Serratrice & Allen, 2015), and indirect speech acts (Schulze et al., 2013) – see Table 1 for examples of these kinds of inference. As well as receiving attention within Developmental Psychology, much of this research has been situated within the field of Experimental Pragmatics, which aims to test and develop pragmatic theory. It is particularly inspired by the work of Grice (1975), but also neo-Gricean and post-Gricean approaches, including Relevance Theory (Sperber & D. Wilson, 1995; D. Wilson & Sperber, 2012) and, more recently, probabilistic pragmatic approaches (e.g. Goodman & Frank, 2016). Grice's key insight was that conversation is a co-operative act, with speakers and listeners sharing expectations that a speaker will be informative, relevant and true, and will observe conventions of language use; this enables listeners to infer speakers' *intended* meaning in a particular context. Take an example of a simple quantity implicature: 'What did you pack in your bag?' 'I packed a book'. Here, the questioner can reason that the addressee means a book *and nothing else*, assuming that the addressee is being informative (giving the most possible information that is relevant) and is knowledgeable about the situation. Then imagine, instead, the following scenario: 'Anyone who packed a book can take a bookmark' 'I packed a book'. In this case, the inference that the speaker packed a book and nothing else is unlikely to be derived, as that is no longer relevant to the discourse. Again, if the speaker instead says, 'I'm not sure, but I think he packed a book', then an exhaustive inference, a book and nothing else, is less likely to be made. These examples show that, crucially, the inferential process is dependent on the context – including the discourse context – and the listener's knowledge of the speaker – including the speaker's knowledge or certainty about the situation being described.

Developmental research has investigated when and how children become able to derive a range of different inferences; a complementary line of research has examined children's production of implicated meaning, either in naturalistic contexts, such as corpus data, or in experimental ones (e.g. Davies & Katsos, 2010; Eiteljoerge et al., 2018; Serratrice & Allen, 2015). Thanks to increasingly child-appropriate methodologies (Veenstra & Katsos, 2018), one consistent finding has been that, for most inference types, children actually begin to be able to derive the speaker's intended meaning at a much younger age than initially thought: at 3 or 4 years, for example, for some quantity or relevance implicatures (Horowitz et al., 2018; Stiller et al., 2015; E. Wilson & Katsos, 2022) and simple perceptual metaphors (Pouscoulous & Tomasello, 2020). Irony, however, is consistently later in developing, from 6 years and throughout childhood (Filippova, 2014; Köder & Falkum, 2021; Zajączkowska et al., 2020), which could be due to the way irony inferences draw on more complex social-cognitive skills (Mazzarella & Pouscoulous, 2021). The aim of this line of research is to understand children's communication in conversation, and so the measures which are used predominantly present stimuli in an oral modality (while written is frequently used with adults). However, the experimental context is often far from a naturalistic conversation, which has important consequences for the interpretation of data and our understanding of the development of inferencing – we return to this important point below. There is also an extensive and valuable literature on pragmatic development in clinical contexts, such as children with Developmental Language Disorder or autistic children. Such studies may use bespoke experimental measures or a variety of standardised global measures of pragmatic ability, such as the Test of Pragmatic Language (Phelps-Terasaki & Phelps-Gunn, 1992), the relevant Comprehensive Assessment of Spoken Language subscales (Carrow-Woolfolk, 1999), the Language Use Inventory (O'Neill, 1996) and the Children's Communicative Checklist (Bishop, 2003). However, these typically include a battery with a broad range of different inferences, as well as skills which are considered pragmatic only in a very broad sense, such as turn-taking (for a review see Matthews et al., 2018). Understanding pragmatic development in a variety of languages and learning experiences is highly important, although in this article we focus predominantly, for the sake of space, on research with typically developing children. Likewise, for brevity, and because this is where the bulk of current research lies, we focus on inference comprehension, but a similar contribution on the production of inferences in conversation and in writing would be welcome.

Recent studies have also begun to investigate factors which support inference development: that is, the skills and knowledge children need to make an inference. This includes structural language (lexical and syntactic knowledge and processing), sociocognitive skills (such as mentalising or Theory of Mind), and Executive Function (EF), including inhibition and working memory. These skills may themselves have complex direct and mediating associations. For instance, children only begin to reliably derive scalar quantity implicatures with 'some' around the age of 5 (see Table 1 for an example), though there is cross-linguistic variability (Katsos et al., 2016). Studies can then examine whether this is because younger children lack the necessary semantic knowledge (Horowitz et al., 2018), have not yet formed a lexical scale such as <some, all> (Barner et al., 2011), or have difficulty tracking what the relevant alternatives are in the discourse context (Skordos & Papafragou, 2016). More generally, developing implicature inferencing is associated with vocabulary and grammatical knowledge, and it could be that better vocabulary aids inferencing, that inferencing skills aid vocabulary acquisition, or, most likely, both (Foppolo et al., 2020; E. Wilson & Katsos, 2022).

To take another example, Gricean theory is often taken to imply a key role for Theory of Mind in pragmatic inferences, as the listener has to reason about the speaker's mental states in assuming that the speaker is knowledgeable and truthful. Studies have examined whether Theory of Mind correlates with children's pragmatic abilities, such as irony (Zajączkowska et al., 2020) and pronoun resolution (Kuijper et al., 2021), and whether children are able to take into account another's perspective, which may be different from their own, in implicature derivation (Kampa & Papafragou, 2020; E. Wilson et al., 2022). So far the evidence is mixed – findings both support and do not support the role of mentalising in quantity implicatures, depending, for instance, on the precise inference required and the measure used (Barner et al., 2018; Hochstein et al., 2016; Kampa & Papafragou, 2020; E. Wilson et al., 2022). More generally in Experimental Pragmatics, the evidence on children's development is also scattered, with research having focused on some inference types, and having used some methodological paradigms more than others, providing a fragmented and incomplete picture of competence in conversational inference to date. Critical calls have challenged the field to expand the phenomena studied and consider more carefully the effects of context on pragmatic strategies and children's inferencing abilities (Andrés-Roqueta & Katsos, 2017; Falkum, 2022). Likewise, there is an increasing awareness of the need to include a diversity of languages and learning experiences in this research, including bi-multilingual children, as linguistic experience could be an important factor itself in pragmatic development (Antoniou et al., 2020; Antoniou & Katsos, 2017; Fortier et al., under review; Katsos et al., 2016; Zhao et al., 2021).

Table 1. ***Examples of inferences typically studied as conversational inferences***

Inference	Example
Implicature (ad hoc quantity)	'Did you meet his parents?' 'I met his mum.' +> not his dad.
Implicature (scalar quantity)	I packed some of the books. +> I pack some but not all of the books.
Implicature (relevance)	'How was the theatre trip?' 'There was a train strike.' +> I couldn't go.
Irony	'I am sorry to announce that the 09:10 train to Cambridge has been cancelled.' 'Superb!' +> Disastrous!
Metaphor	The tree was wearing a white hat. +> The tree was covered in snow.
Metonymy	The nursery emailed some information. +> A member of staff at the nursery emailed some information.
Presupposition	I went to Paris again. +> I'd been to Paris before.
Indirect speech act	Can you give me your shoes? +> Give me your shoes.

Reading inferences

Making inferences is acknowledged as a crucial part of learning to read, both in research (Castles et al., 2018; Kendeou et al., 2016) and in teaching practice (e.g. Such, 2021). Indeed, inferencing ability is found to be a key predictor of reading comprehension (Bowyer-Crane & Snowling, 2005; Oakhill & Cain, 2012). For example, take the text: 'Finally the family arrived. They flung open the car doors, heard the gulls, and felt the salt spray on their faces'. To explain where the action took place, a good comprehender might make the global coherence inference that the family had arrived at the seaside, although that is not explicitly stated in the text. In contrast, a poor comprehender might struggle to draw on information across the text and fill in information from background knowledge to make this inference. To take another example, reading 'Jake gave the book to Tom. He thought he'd like it', the reader can infer that the two instances of the pronoun 'he' refer to Jake and Tom respectively, and that the two sentences are related *causally* (*because* Jake thought Tom would like the book) – examples of local cohesion inferences. Again, a poor comprehender could struggle to make these inferences and therefore to access the full meaning of the text.

Table 2. ***Examples of inferences typically studied as reading inferences***

Inference	Example
Global coherence	The delicate glass vase fell to the floor. Sue went to fetch a brush. +> The vase broke and Sue intended to clear up the mess.
Local cohesion – anaphor resolution	Jake gave the book to Tom because he thought he'd like it. +> Jake gave the book to Tom because he (Jake) thought he (Tom) would like it.
Local cohesion – causative	Jake gave the book to Tom. He thought he'd like it. +> Jake gave the book to Tom *because* he (Jake) thought he (Tom) would like it.

Research on reading inferences shares with research on conversational inferences some of its origins in early psycholinguistics (Graesser et al., 1994). The motivation is to understand how children comprehend *texts,* and as such it is concerned not just with children's reading but also *listening* to texts, which is particularly important for younger children who are still learning to decode (Language and Reading Research Consortium & Muijselaar, 2018). Methodologically, this means texts may be presented within studies in written or oral modality, testing reading or listening comprehension – again, a point we return to below. One important driver in this line of research is the need to identify components of reading comprehension which, separately or together, may present challenges to children, resulting in poor comprehension (Oakhill, 2020), and this ultimately contributes to the development of interventions to boost reading skills (Elbro & Buch-Iversen, 2013; Elleman, 2017; Kispal, 2008; Whatmuff, n.d.).

A widely-adopted theoretical framework for reading comprehension is The Simple View of Reading (Hoover & Gough, 1990), which defines reading comprehension as the product of decoding (reading individual words) and listening comprehension (for reviews of other frameworks, see Cain & Barnes, 2017; McNamara & Magliano, 2009). That is not to say that reading comprehension is simple; rather, the framework simply encapsulates its two main components, with listening comprehension being the result of complex linguistic and cognitive processes that contribute to building a mental model of a text. The complementary Rope Model (Scarborough, 2009) breaks down listening comprehension into multiple strands: knowledge of and access to vocabulary, background knowledge, understanding of sentence structure, inferencing, and knowledge of texts and their structures; to this have been added other factors like comprehension monitoring (Oakhill et al., 2015). The outcome of successful reading (and listening) comprehension is a situation model: a mental representation of the state of affairs described by a text, which goes beyond the literal meaning, and includes meanings integrated across sentences and inferences constructed from the text and background knowledge (Kintsch, 1998).

Inferences for text comprehension are classified or modelled in a variety of ways. First, a distinction is sometimes made between necessary and elaborative inferences. Necessary inferences, as in the examples above, are required to build a coherent mental model, whereas elaborative inferences enrich the mental model but are usually not regarded as essential for comprehension (Barnes et al., 1996; Cain et al., 2001). For example, a reader might infer that the family in the example above had arrived at a sandy beach with blue sea, or were happy to be there after a long journey – but these inferences are not necessary for a coherent mental model. Concentrating on necessary inference, we have already illustrated the distinction between global coherence and local cohesion – see Table 2 for examples of these types of inference. Studies on the development of reading inferences may use just one type of inference, local or global (Oakhill, 1982; Oakhill & Yuill, 1986; Yuill et al., 1989), or both (e.g. Barnes et al., 1996; Cain, Oakhill, & Bryant, 2004; Cain & Oakhill, 1999; Davies et al., 2019; Joseph et al., 2021), but they typically use a range of inferences for each type (e.g., anaphor resolution and causal relations for local cohesion). In addition, they may present texts aurally (listening comprehension (Currie & Cain, 2015)) or visually (reading comprehension (Barnes et al., 1996)). Alternative taxonomies focus on different functional distinctions of inferences as well as their sources of information, such as connecting inferences or backward elaborations (van den Broek et al., 1993). According to still other approaches, categorising different inference types is less important than characterising a general inference skill, which "depends on the core, fundamental processes of activation and integration of information and generalises across contexts" (Kendeou, 2015, p. 160).

One key finding from large-scale, longitudinal studies is that inferencing skills in reading (or listening to texts) improve across childhood, from 4 to 15 years (e.g. Barnes et al., 1996; Language and Reading Research Consortium & Muijselaar, 2018). In general, very young children can make inferences about causal, spatial and temporal relations, in real-world situations and then in linguistic communication. However, in linguistic communication and especially reading, there are a number of complex interacting factors which constrain the number and the type of inferences a child can actually make during comprehension (Cain & Barnes, 2017; Kendeou, 2015). Availability and accessibility of background knowledge, working memory, inhibition and cognitive load (e.g. from conflicting sequencing of information or from decoding) all change over development, meaning that on the whole children make more inferences with age (Barnes et al., 1996; Currie & Cain, 2015). Further, there is a clear relationship between inferencing and vocabulary within age groups, and a reciprocal relationship over development. This is particularly the case for vocabulary depth – how much a reader knows about words – rather than just vocabulary breadth – how many words they know. Vocabulary depth predicts later inferencing, and inferencing predicts later vocabulary depth (Cain & Oakhill, 2014; Language and Reading Research Consortium et al., 2019). These studies often require participants to answer comprehension questions about the text, to assess which inferences children have made, i.e. their explicit knowledge of inferred meaning is assessed. In older children, online eye-tracking while reading can be used to address questions about the time course of inferencing. For instance, 8- to 13-year-olds prioritise efficiency when reading: they initially only make the most necessary inferences, and then go back if they meet inconsistent information and need to revise their interpretation (Joseph et al., 2021).

A significant body of work has examined factors associated with reading comprehension in general (which typically includes inferencing). One focus has been on EF, and especially working memory, based on the assumption that this is required for keeping information in mind and integrating it across sentences to contribute to a mental model of the text (Follmer, 2018; Language and Reading Research Consortium et al., 2019; Nouwens et al., 2021). Theory of Mind, although given less attention so far in reading studies, has also been argued to be important (Dore et al., 2018). It has been found to be related to listening comprehension more generally (Kim, 2020; Kim & Phillips, 2014), and to predict reading comprehension longitudinally (Atkinson et al., 2017). Finally background or world knowledge is crucial for inferencing, as particular coherence inferences result from integrating information provided explicitly by the text with background knowledge (Smith et al., 2021).

Taking stock: similarities and differences between conversational and reading inferences

Our review highlights striking similarities in the findings to date across research on children's conversational and reading inferences. However, there are also differences in approach, methods, and findings which lead us to some fundamental questions about children's development of inferences – and which invite a collaborative research agenda to address them.

In both conversation and reading, learning to make inferences is crucial for understanding meaning, as well as for learning about language and about the world (Bohn et al., 2021; Horowitz & Frank, 2016). In both contexts, arguably, inferences are made for coherence and relevance, either to arrive at the intended meaning of the speaker or the writer. Further, a variety of factors have been identified which are at the very least correlated with inferencing skills, and which may well be contributors to their development. For both conversational and reading inferences, studies have shown associations with vocabulary; background or world knowledge; Theory of Mind; and EF, including working memory and inhibition. We take up this point of convergence below, but also note that the availability and strength of evidence varies across contexts, inference types and age groups, and so there are still gaps in our empirical understanding and theoretical models of the development of inferencing.

There are also differences between the dominant research traditions on children's conversational and reading inferences, which we summarise in Table 3. First, research on conversational inferences has typically sought to identify qualitative changes in development: when children become able to derive certain inferences, and which theoretically-motivated prerequisite factors might prevent or allow inferences. On the other hand, research on reading inferences has often focussed on quantitative change, observing a gradual improvement of children's inferential skills over time, perhaps in the number of inferences made or the number of cues required for an inference (Currie & Cain, 2023; Van den Broek et al., 2015). The difference in age group studied is important here: studies on conversational inferences typically sample 3- to 7-year-olds, depending on the type of inference studied, whereas studies of reading, by their very nature, typically begin around 5 years at the start of reading instruction, right through to the teenage years, although tasks which involve listening to written texts may be used with younger children.

Second, different phenomena have been the focus of research across studies on conversational and reading inferences. In general, for conversational inferences the focus has been on implicatures, alongside metaphor, metonymy, irony, anaphora and more; whilst for reading the focus has been on global coherence and local cohesion inferences, alongside some other figurative language use such as idioms. Thus, the study of conversational inferences has focused on classic pragmatic phenomena, which are clearly communicative (linked to speaker intention) and linguistic (based on speaker utterances); approaches to reading inferences have examined either pragmatic–syntactic or pragmatic–lexical phenomena (e.g., lexical disambiguation and anaphora resolution), or potentially general inferences, like causality or character intent, most analogous to relevance implicatures.

Third, these two areas of research – on conversational inferences and reading inferences – are set in different theoretical frameworks: in a Gricean approach to pragmatics, the listener's goal is to arrive at the speaker's intended meaning; in typical models of reading comprehension, the reader's goal is to construct a coherent mental model, which does involve the author's intended meaning (Kintsch, 1998). Gricean approaches tend to model inferencing at the computational level of explanation (answering 'what' and 'why' questions) in terms of logical or rational steps in reasoning; where the nature of reading inferences is specified, it tends towards a psycholinguistic notion of spreading activation – inferencing skill depends on activation of information from the text or background knowledge and integration of this with new information (Kendeou, 2015). These approaches are though by no means mutually exclusive, of course.

Fourth, the different research traditions have both employed a whole range of experimental designs and paradigms. There is, though, a tendency for research on developing conversational inferences to involve small-scale, tightly controlled bespoke tests on a single conversational inference type, using implicit measures like picture-selection. Research on the development of reading inferences, meanwhile, has additionally involved large-scale studies, and has included both standardised and experimenter-designed tests, requiring expressive responses from participants, such as answering questions explicitly (alongside eye-tracking for fluent readers).

Below we examine how these differences raise a number of important questions about the type of inferences children are learning to derive, and the factors playing a role in children's development of inferencing. First, though, we address the issue of modality and what effects it might have on that development.

Table 3. ***Summary of key differences between research programmes on conversational and reading inferences***

Feature of research	In research on conversational inferences	In research on reading inferences
Typical types of research question	Are the predictions of pragmatic theory fulfilled in pragmatic development? When do children acquire the ability to derive a particular type of pragmatic inference? What are the socio-cognitive and linguistic factors which facilitate or hinder children's inferencing ability?	How does inferencing relate to reading comprehension? How does general inferencing ability develop with age? What are the socio-cognitive and linguistic factors associated with inferencing ability?
Prominent theoretical frameworks	Gricean and neo-Gricean pragmatic theory (e.g., Degen & Tanenhaus, 2014; Grice, 1975; Levinson, 2000) Relevance Theory (Sperber & D. Wilson, 1995; D. Wilson & Sperber, 2012) Probabilistic Pragmatics, including Rational Speech Act theory (Frank & Goodman, 2012; Franke & Jäger, 2016) Speech Act theories (Austin, 1962; Searle, 1969)	The Simple View of Reading, and the Rope Model (Hoover & Gough, 1990; Scarborough, 2009) Construction-Integration model of text comprehension (Kintsch, 1998) Connectionist models of text comprehension (Graesser et al., 1994) For other accounts see McNamara & Magliano (2009)
Age group typically studied	3–7 years, and older for later developing inferences like irony	From 5 years, with some listening comprehension studies at younger ages
Typical methodologies	Truth Value Judgement or Felicity Judgement Sentence-to-picture matching (with reaction time) Visual world paradigm eye-tracking Action-based tasks	Question-and-answer comprehension tasks (explicit responses) Eye-tracking while reading (for older children)
Typical research designs	Cross-sectional, with participants grouped by age or age taken as a continuous variable Focussed on a single inference type with experimental manipulation Uses a bespoke measure Often small-scale, conducted in psychology and linguistic labs	Cross-sectional or longitudinal May include a range of inferences in an inference or reading comprehension task May use a bespoke or validated or standardised measure Can be large-scale, conducted in schools

What is the effect of modality on inferencing?

Children start learning how to derive communicative inferences in conversation as they develop oral language skills; from a young age, they also make inferences from wordless picture books, and when they are read to during shared book reading (e.g. Paris & Paris, 2003; Silva & Cain, 2015); and they then bring these skills to the task of learning to read. How might modality affect how children develop their inferencing abilities? Of the many differences between text and spoken language, there are some which seem important both for inferencing itself and for the opportunity to learn inferencing.

In a conversation the interlocutor is typically co-present, whereas when reading a text, the author is not. Spoken utterances therefore include cues such as prosody and gesture, and are supported by facial expressions and immediate context. As mentioned previously, in Gricean models of conversational inference, the listener's reasoning about the speaker's epistemic state plays a fundamental role, but it has also been suggested that the co-presence of the speaker can be an important cue for this mentalising (Katsos & Andrés-Roqueta, 2021). A text, on the other hand, gives its own kind of context, including descriptions of characters or the writer's epistemic state, together with genre and background knowledge, and there may well be pictures in children's books. Theory of Mind has also been suggested as an important factor in reading, but primarily to follow characters' mental perspectives and emotions in narrative texts (Dore et al., 2018).

Furthermore, the opportunities to learn inferential processes and cues to meaning may differ across modalities. In conversation, there is the opportunity to repair miscommunication through questioning – something we know to be important for referential production, at least (Matthews et al., 2012); and miscommunication itself may be revealed by speaker feedback. When reading, there is the possibility of going back over a text, for example if it becomes clear that something earlier was misunderstood, or a necessary inference was not made (Joseph et al., 2021). For children, this revision may be prompted by questioning, which may be particularly effective when it immediately follows the inference-triggering text, rather than comes at the end of the text (Butterfuss et al., 2022; Freed & Cain, 2017).

If we consider existing research, we can see that the distinction between oral and written language does not map onto studies targeting conversational and reading inferences, respectively. First, as part of the suite of rigorous methods used in Experimental Pragmatics, carefully controlled stimuli often involve utterances being presented to the listener somewhat 'out-of-the-blue', with little information about the speaker, and, in the case of adults or older children, often as text. Indeed, such studies typically pay little attention to whether an utterance is read or heard. That is: much of the body of research on conversational inferences to date does not include much *conversation*, and does not focus on the affordances of the oral modality. The aim of this approach is to break down the 'building blocks' of communication, and be able to focus on a particular type of inference, reducing the effects of confounding factors, but this may involve removing supporting cues like prosody, gesture and facial expression, as well as discourse context (Noveck, 2018). Second, studies on reading inferences include children's inferences when listening to texts read aloud and when viewing wordless picture books. This reflects the ways children encounter texts not just when they themselves are reading, but, more often in early childhood, when they are being read to, in a shared book context. It is also motivated by the need to mitigate for the influence of developing word reading skills: in the early stages of reading acquisition, a focus on decoding written words on the page takes cognitive effort which can obscure children's comprehension skills, including inferencing. We summarise these relationships between modalities and typical research paradigms in Table 4.

Table 4. ***Summary of the ways in which the features of modality intersect with current studies on conversational and reading inferences***

	Studies on conversational inferences	Studies on reading inferences
Modality	May be presented in oral or written modality	May be presented in oral or written modality (listening or reading comprehension)
Discourse context	Sentences may be presented out of the blue, or in simple question-and-answer pairs, with little or no discourse context	Texts may consist of a few short paragraphs
Social context	Interlocutor may or may not be co-present; stimuli may be presented on a computer screen and/or as spoken by an avatar or fictional character; participant may be an observer rather than interlocutor.	Texts may or may not be presented in a shared book reading context.
Nature of experimental stimuli	Sentences are often highly controlled in vocabulary and grammatical structure; there may or may not be naturalistic prosody.	Sentences may be controlled and may not reflect lexical and syntactic patterns typically found in children's books.

When we then think of the implications of bringing together findings on reading and conversational inferences, we first need to be careful to take them in their experimental context. Second, we can see that in those controlled experimental contexts there may actually be fewer differences between stimuli targeting ‘conversational inferences’ and those targeting reading inferences, than we would see between naturalistic conversation or reading. Naturalistic conversation and reading may differ substantially, for example, in lexical or syntactic complexity (Dawson et al., 2021). Furthermore, the affordances of oral and written modalities give rise to some interesting questions: how do children learn how to look out for and give appropriate weight to different cues that need to be taken into consideration when deriving inferences, across different modalities? And how do spoken and written language provide differing opportunities to do this? To give an example: deriving a late-developing inference like irony, associated with mentalising skills, is likely to be aided by the cues of a co-present speaker, which in itself provides a strong signal of the need for mentalising, along with features like prosody. On the other hand, another inference type like anaphora resolution may be less affected by modality, although the temporal affordances of reading – being able to go back over text – might be beneficial. In general, the role of modality needs to be addressed with a collaborative approach to the study of inferencing development.

Which inferences are children learning?

We observed that research on conversational and reading inferences has tended to have different foci in terms of types of inferences. Why might this be? A first possible explanation is that different theoretical frameworks or simply historical precedent could have played a role: there is no reason a priori to think quantity implicatures, for example, could not be studied in a text, or coherence inferences studied in conversation, as indeed has been the case under some approaches, such as in Literary Pragmatics (e.g. Chapman & Clark, 2019). Second, given the tendency for studies on conversational inferences to start with young children, aged 3 years and upwards, and for those on reading inferences to examine older children, from 5 years, another explanation is that these studies focus on those inference types particularly developing in those periods. However, we have already seen that this is not (solely) the case: some typical conversational inferences such as irony are relatively late developing, from around 6 years (Filippova, 2014); typical reading inferences such as coherence inferences are surely required in conversation too, before learning to read, and indeed phenomena like anaphor and reference resolution have been studied across modalities and contexts (e.g. Arnold et al., 2007; Pyykkönen et al., 2010; Serratrice, 2007; Song & Fisher, 2007). That said, there could be a third explanation, based on either qualitative or quantitative differences in the kinds of inferences which children develop in conversation and in reading, due to the different nature of the input. Just as there are differences in vocabulary, syntax or structure between language typically used in conversation and in reading and writing (Castles et al., 2018), there could be differences in the inferences required to understand the speaker's or the author's intended meaning: certain inference types could only be encountered in one context or, more likely, encountered more frequently in one context than another. This is an important empirical question that requires further investigation – and is indeed crucial to understanding the pragmatic challenges in learning to communicate and learning to read. In order to achieve an accurate and complete understanding of children's pragmatic development, we need to look at all kinds of inference in all contexts – conversation, reading and listening to texts.

This leads us to the next set of questions: what is the interaction between pragmatic skills and reading development? In other words, which inferencing skills do children bring to learning to read, and which do they develop *for* reading? And then what is the effect of learning to read on pragmatic development more generally? In other realms of linguistic development, reading, and being read to, are key contributors: for example, a reciprocal relationship between vocabulary and reading comprehension has been observed (e.g. Oakhill & Cain, 2012), and that relationship is, in part, mediated by inferential skills (Cain, Oakhill, & Lemmon, 2004; Elleman, 2017; Language and Reading Research Consortium et al., 2019). We expect transfer of linguistic skills across modalities and across contexts, and it would be surprising if this was not also the case in the realm of pragmatics, notwithstanding the possible effects of modality on learning to make inferences which we have already discussed. To date, however, developmental studies within Experimental Pragmatics have paid very little attention to whether children are readers or not (though Katsos et al, 2016, did find an effect of being in school on quantifier understanding). One study, however, which actually compared typical cases of conversational and reading inferences directly in 7–13-year-olds, found a surprisingly low correlation between a textual local inference task and an implicature task, about the same as with vocabulary and grammar skills, with analysis suggesting that task-specific skills play an important role (A. C. Wilson & Bishop, 2022). These questions clearly need further research, by adopting developmental approaches to this kind of comparative data, taking into account modality, context, and inference type. The answers to these questions are particularly important for the first years of formal education: inferencing skills are known to be developing significantly in both conversation and reading; children are exposed to texts both as readers and as listeners; and they are given a new linguistic experience in the classroom.

What are the explanatory factors in children's inferencing development?

Research on the development of conversational and reading inferences has identified a variety of knowledge, skills, processes and experiences which are involved in deriving inferences: conceptual and structural knowledge (background and world knowledge, vocabulary and grammar); social cognition; environmental factors (linguistic and multilingual experience, and socioeconomic status); and EF. These can be related to inferencing abilities directly or in a mediated way, but to date the amount of research and strength of evidence across different factors and inference types is very variable. For example, vocabulary knowledge is required to understand the semantic content of an utterance or piece of text, which is needed for deriving any one inference, but vocabulary knowledge also provides more opportunities in general to access at least some meaning in a discourse or text, and thereby practise pragmatic skills (LARRC et al., 2019; Oakhill & Cain, 2012; E. Wilson & Katsos, 2021). To take another example, social cognition (particularly Theory of Mind) has been widely implicated in the development of conversational inferences, but there is growing evidence that its role may depend on the inference and discourse context at hand (Katsos & Andrés-Roqueta, 2021). For reading inferences, social cognition has been particularly linked to inferences about characters' perspectives and emotions (Dore et al., 2018). The effect of socioeconomic factors on pragmatic skills has received relatively little attention; a whole number of factors associated with socioeconomic experience could impact inferencing, including access to books, libraries, and material resources more broadly, diversity of linguistic input, structural language skills, and cognitive skills including mentalising (e.g. Cutting & Dunn, 1999; Hughes et al., 1999). Cross-linguistic work has also started to reveal the effect of language for conversational inferences, as languages grammaticalize or lexicalise different information (e.g. Katsos et al., 2016 for quantifiers), while studies with multilingual children have so far yielded mixed evidence on the effects of learning more than one language on inferencing development (Antoniou et al., 2020; Antoniou & Katsos, 2017; Dupuy et al., 2019). A systematic review of empirical research of these factors across conversational and reading inferences is needed to identify consistencies, inconsistencies, and gaps in knowledge to inform the development of testable theoretical models of inference development.

To illustrate in a little more detail the task of building a model with factors which contribute to the development of inference-making, take the example of Executive Function. EF is itself a complex construct, most commonly conceived as including working memory, inhibition and cognitive flexibility. Crucially, it is developing over the preschool and early school years – both in its components and their integration (De Cat, 2015; Diamond, 2006). Within Experimental Pragmatics, children's challenges with inferences have sometimes been attributed to 'processing difficulties', sometimes EF in particular (e.g. Huang & Snedeker, 2009; Pouscoulous et al., 2007; Siegal et al., 2010). Recently, there has been an increased understanding of which particular cognitive skills may be required, in theory, for particular types of inference. For example, quantity implicatures require generating and accessing alternatives: when a speaker says, 'I ate some of the biscuits', the listener has to generate the alternative, 'all', as both a lexically plausible and contextually relevant alternative, and then negate it, to arrive at the meaning, *I ate some but not all of the biscuits.* This has led to the hypothesis that inhibition might play an important role in negating the literal meaning of the utterance. Developmental studies, however, have so far offered mixed findings: they have not observed an association between inhibition and inferencing skills, when testing whether performance on an implicature task is predicted by performance in an inhibition task. For example, Antoniou, Veenstra, Kissine and Katsos (2020) found no evidence for an effect of inhibition, but did find that performance on a battery of pragmatic inferences was predicted by a combined working memory measure in 10–12-year-olds (see too Horowitz et al., 2018; Nordmeyer et al., 2016; see also Zajączkowska & Abbot-Smith, 2020 for cognitive flexibility and irony).

When it comes to reading inferences, the focus has largely been on working memory, given the need for the reader to hold in mind information from across sentences and then integrate that with the mental model of the text and with newly activated background knowledge (Oakhill et al., 2015). For example, in LARRC, Currie and Mujselaar's (2019) large-scale longitudinal study with children aged 4 to 9 years, children heard short texts, including sentences such as: 'Even though Tim's thumb was bruised and sore, he was smiling. He put the hammer that had caused the pain away in his toolbox'. They then had to answer questions like, 'Why did Tim have a sore thumb?', which require integration of information from the two sentences, prior text and background knowledge. After variance associated with vocabulary was taken into account, they found little influence of working memory on inferencing at each grade; this reflects a trend in results across studies that suggests working memory alone is not a unique predictor of inferencing (see too for a meta-analytic review Peng et al., 2018). This sits against a backdrop, though, of a large body of work which has found evidence for the role of all Executive Functions, including working memory, in reading comprehension in general (see Follmer, 2018 for a meta-analytic review).

Research on explanatory factors in children's inferencing development, including contradictory findings and incomplete evidence, opens up a number of important questions. First, are the key predictors the same for conversational and reading inferences? Or are there differences in the required cognitive, linguistic and social resources which are due to either the context (conversation or text) or the modality itself? Socio-cognitive capacities, linguistic experience and knowledge, and learning strategies and processes are also developing significantly in early childhood, so we would expect cascading development across these domains to affect inferencing skills (Bohn & Frank, 2019; Oakes & Rakison, 2019). We might expect the relative contribution of related skills to change over time too. Second, how do different inference types vary in the knowledge and skills they require? For instance, inferences might call on more or less challenging vocabulary and grammatical knowledge, depending on the features of the utterance or text; or require higher or lower levels of inhibition and working memory, depending on the strength of relevant information to be inhibited, or length of discourse or text implicated in an inference; or engage more or less with social cognition, depending on whether the speaker or writer's perspective has to be actively taken into account to resolve the intended meaning. An informative approach to understand apparently contradictory findings could be to consider the different inference types at stake in different experimental contexts, and the differences in cognitive load and processes potentially involved. This needs to be done at quite a fine-grained level: for example, reading inference studies testing 'local cohesion inferences' can often include both anaphora resolution and bridging inferences, which draw on different levels of lexical and grammatical knowledge, and potentially other areas of knowledge and cognitive functions to differing extents too.

Third, how can these complex interactions be modelled and motivated theoretically? When considering associations between two complex constructs, like inferencing skills and EF, there are multiple mutually-inclusive possible linking hypotheses (Matthews et al., 2018): even for a factor like vocabulary knowledge, there are potentially different roles of vocabulary breadth and depth, and immediate or long-term ways that these contribute to inferencing. Finally, to what extent do the experimental measures used contribute both to hypothesised predictors and to their observed effects? Longer texts used to test reading inferences, often with multiple questions at the end of a text, are likely to reveal more individual differences in working memory, for instance, than the single-sentence stimuli which are designed to trigger inferences within Experimental Pragmatics studies. Furthermore, it can be hard to disentangle confounds in the measures used: for instance, verbal working memory tasks may rely on verbal skills, but vocabulary and verbal intelligence are themselves predictive of reading inferencing ability (Cain, Oakhill, & Bryant, 2004; Kidd et al., 2018). In sum, there are many complex interactions still to map out, a challenge which lends itself to a collaborative approach.

A collaborative approach to children's inferences

We have suggested that bringing together research, and researchers, on conversational and reading inferences brings to light a number of core questions for children's development of inferencing, which we have collated in Table 5. These include: What effect does both context (conversation or reading) and modality (oral, visual, written) have on the need for children to make inferences, and for the opportunities for them to learn to do so? And how do linguistic and background knowledge, socio-cognitive skills and environmental factors support different inferences across contexts and modalities? We suggest that a collaborative approach is the best way of addressing these outstanding questions.

First, a collaborative approach to the study of children's development of communicative inferences means that linguists, cognitive psychologists, developmental psychologists and educational psychologists across a number of research approaches have to work together, possibly in adversarial collaborations where differing theoretical frameworks are tested empirically. This paper itself was born out of a workshop hosted at the University of Cambridge which brought together researchers from different research areas with a common interest in children's inferencing. Working together throws light on differing assumptions – for instance, about what a communicative inference *is*, or how to test children's inferencing skills – as well as common or contradictory findings, which give rise to the kinds of questions we have outlined. For example, several papers discussing reading comprehension argue and provide evidence for the idea that inferencing in reading has earlier precursors in oral language or general inferencing skills (Cain & Barnes, 2017; Kendeou, 2015; Van den Broek et al., 2015), while linguists approaching developmental pragmatics questions would assume this was the case – but they in turn often do not pay attention to whether children are readers, and which modality an utterance is presented in.

Table 5. ***Summary of questions raised for future collaborative research on conversational and reading inference development***

What is the effect of modality on inferencing?	How do children learn how to look out for and give appropriate weight to different cues that need to be taken into consideration when deriving inferences, across different modalities? How do spoken and written language provide differing opportunities to do this?
Which inferences are children learning?	What is the interaction between pragmatic skills and reading development? Which inferencing skills do children bring to learning to read, and which do they develop for reading? What is the effect of learning to read on pragmatic development more generally?
What are the explanatory factors in children's inferencing development?	Are the key predictors the same for conversational and reading inferences? Or are there differences in the required cognitive, linguistic and social resources which are due to either the context (conversation or text) or the modality itself? How do different inference types vary in the knowledge and skills they require? How can these complex interactions be modelled and motivated theoretically? To what extent do the experimental measures used contribute both to hypothesised predictors and to their observed effects?

Second, it means considering in more detail how existing theories which were developed to account for developing conversational and reading inference skills overlap and interact. We echo the call of Matthews, Biney and Abbot Smith (2018), writing in light of their review of individual differences in pragmatic skills, "to integrate the results of modelling individual differences data and complementary experimental work ... into psycholinguistic models of language processing" (2018:202). We need to be clear about the levels of analysis that current models are operating at (Franke & Jäger, 2016; Geurts & Rubio-Fernández, 2015), and aim ultimately for a mechanistic model which can connect inferencing to related areas of cognition. The aim would be to establish how different types of inference cluster together, both in terms of their developmental trajectory in children's comprehension, and also in how they are supported or affected by the context (conversation or reading), modality, and other linguistic, socio-cognitive and environmental factors. This may show that inferences, which were previously separately categorised or differently labelled, may pattern together, or that grouped inferences actually behave differently. Such a model, on the one hand, is informed by and informs theories of meaning which motivate inference making; and on the other hand, can then be related via linking hypotheses to underlying processes, including EF. The key is that the ongoing, interactive development of oral language skills and literacy are taken into consideration.

Third, it means combining methods: by comparing children's performance on tightly controlled experimental methods and more naturalistic measures to explore and begin to explain task factors from different experimental contexts; by combining insights from these different designs to improve both experimental and naturalistic measures; and by implementing them longitudinally, as has been a particular tradition in studies on reading inferences. A particular challenge is the availability of standardised measures which are psychometrically valid and reliable, but which also measure particular pragmatic inferences rather than a mix of pragmatic and communication skills (see Matthews, Biney and Abbot Smith, 2018, for a review). Likewise, task reliabilities for cognitive measures can also be surprisingly poor, both in terms of test–retest reliability and order of presentation in an experimental session (Schuch et al., 2022). A further problem for the studies of individual differences is the widespread correlation, or positive manifold, across different cognitive measures – sometimes attributed to a g factor or to interacting developmental processes (Van Der Maas et al., 2006). One study in pragmatics that has moved in a promising direction, A. C. Wilson and Bishop (2022), found evidence for a family of pragmatic skills, with only modest correlation between them, and differing levels of association with vocabulary and grammatical skills with a test battery for older children aged 7–13 years; a particular strength of this study was its testing of the reliability of the measures in an adequately powered sample. Similarly, Bohn et al. (2023) tested six tasks for pragmatic inferencing in 3-5-year-olds (including quantity implicature and informativeness inferences) for retest reliability, formalised the shared features of these inferences theoretically, and then tested their association with other cognitive skills including EF in an individual differences study; they found evidence for a systematic relationship between the pragmatic and EF tasks. The next steps are to extend these kinds of approaches to more inference types and across age groups, particularly from preliteracy through the primary school years as children learn to read.

This is a substantial challenge, but it is one that has important consequences. Poor inferencing skills have been identified as one of the causes of poor communication and poor reading comprehension outcomes (e.g. Botting & Adams, 2005; Cain & Oakhill, 1999; Oakhill & Cain, 2012). Ultimately, we need to identify exactly which of the family of pragmatic skills and processes are most at risk across contexts, and then develop and test interventions which can boost those inferential skills. It is even an open question as to whether targeted inferencing interventions are most effective given limited educational time and resources (Davies et al., 2019; Elleman, 2017; Kendeou et al., 2020; Whatmuff, n.d.), or whether a focus on vocabulary, grammar, background knowledge or high level communication skills have enough positive influence on pragmatics (West et al., 2021). Butterfuss, Kendeou, McMaster, Orcutt & Bukut (2022) developed and tested a reading inferencing intervention with audiovisual and non-reading contexts in preliterate pre-schoolers, and while they did find a boosting effect of questioning, scaffolding and feedback, this was greater for children who already had higher language skills and EF – a Matthew effect, where children who already have more advanced skills develop even more than those who do not. In either case, understanding the similarities and differences in inferencing in both conversation and reading contexts, and across modalities, is likely to be crucial. A collaborative approach to researching children's development of inferences across oral language and reading has the potential to provide a more accurate and fuller picture of children's developing pragmatic skills, and a deeper understanding of how they can be improved.

In sum, in this perspectives article we have called attention to the two distinct bodies of research on inferencing development – targeting conversation and reading. In general, they share some basic assumptions about what inferencing is for, and an increasing focus on the factors which are associated with developing inference skills. However, there are also some interesting and potentially critical differences, in the phenomena studied, in methodologies, and in motivation, which may account for apparently contradictory findings and provide insight into future avenues of research that will provide more comprehensive accounts of linguistic and cognitive development. Not least this includes how learning to make inferences in conversation relates to learning to make inferences when listening to or reading texts, and vice versa. We have argued that combining theoretical and empirical expertise on inferencing in conversation and reading is crucial for gaining a full understanding of children's pragmatic development.

References

Adams, C., Clarke, E., & Haynes, R. (2009). Inference and sentence comprehension in children with specific or pragmatic language impairments. *International Journal* *of Language & Communication Disorders, 44*(3), 301–318. https://doi.org/10.1080/13682820802051788

Andrés-Roqueta, C., & Katsos, N. (2017). The contribution of grammar, vocabulary and theory of mind in pragmatic language competence in children with Autistic Spectrum Disorders. *Frontiers in Psychology, 8,* 996. https://doi.org/10.3389/fpsyg.2017.00996

Antoniou, K., & Katsos, N. (2017). The effect of childhood multilingualism and bilectalism on implicature understanding. *Applied Psycholinguistics.,* 1–47. https://doi.org/10.1017/S014271641600045X

Antoniou, K., Veenstra, A., Kissine, M., & Katsos, N. (2020). How does childhood bilingualism and bi-dialectalism affect the interpretation and processing of pragmatic meanings? *Bilingualism: Language and Cognition, 23*(1), 186–203. https://doi.org/10.1017/S1366728918001189

Arnold, J. E., Brown-Schmidt, S., & Trueswell, J. (2007). Children's use of gender and order-of-mention during pronoun comprehension. *Language and Cognitive Processes, 22*(4), 527–565. https://doi.org/10.1080/01690960600845950

Atkinson, L., Slade, L., Powell, D., & Levy, J. P. (2017). Theory of mind in emerging reading comprehension: A longitudinal study of early indirect and direct effects. *Journal of Experimental Child Psychology, 164,* 225–238. https://doi.org/10.1016/j.jecp.2017.04.007

Austin, J. L. (1962). *How to Do Things with Words.* Oxford University Press.

Barner, D., Brooks, N., & Bale, A. (2011). Accessing the unsaid: The role of scalar alternatives in children's pragmatic inference. *Cognition, 118*(1), 87–96. https://doi.org/10.1016/j.cognition.2010.10.010

Barner, D., Hochstein, L. K., Rubenson, M. P., & Bale, A. (2018). Four-year-old children compute scalar implicatures in absence of epistemic reasoning. *Semantics in Language Acquisition, 24,* 325–349.

Barnes, M. A., Dennis, M., & Haefele-Kalvaitis, J. (1996). The Effects of Knowledge Availability and Knowledge Accessibility on Coherence and Elaborative Inferencing in Children from Six to Fifteen Years of Age. *Journal of Experimental Child Psychology, 61*(3), 216–241. https://doi.org/10.1006/jecp.1996.0015

Bishop, D. V. M. (2003). *Children's Communication Checklist (vol 2).* Harcourt Assessment.

Bishop, D. V. M., Snowling, M. J., Thompson, P. A., Greenhalgh, T., & Consortium, C. (2016). CATALISE: A Multinational and Multidisciplinary Delphi Consensus Study. Identifying Language Impairments in Children. *PLOS ONE, 11*(7), e0158753. https://doi.org/10.1371/journal.pone.0158753

Bohn, M., & Frank, M. C. (2019). The Pervasive Role of Pragmatics in Early Language. *Annual Review of Developmental Psychology, 1*(1), 223–249. https://doi.org/10.1146/annurev-devpsych-121318-085037

Bohn, M., Tessler, M. H., Kordt, C., Hausmann, T., & Frank, M. C. (2023). An individual differences perspective on pragmatic abilities in the preschool years. *Developmental Science, n/a*(n/a), e13401. https://doi.org/10.1111/desc.13401

Bohn, M., Tessler, M. H., Merrick, M., & Frank, M. C. (2021). How young children integrate information sources to infer the meaning of words. *Nature Human Behaviour, 5*(8), Article 8. https://doi.org/10.1038/s41562-021-01145-1

Botting, N., & Adams, C. (2005). Semantic and inferencing abilities in children with communication disorders. *International Journal of Language & Communication Disorders, 40*(1), 49–66. https://doi.org/10.1080/13682820410001723390

Bowyer-Crane, C., & Snowling, M. J. (2005). Assessing children’s inference generation: What do tests of reading comprehension measure? *British Journal of Educational Psychology, 75*(2), 189–201. https://doi.org/10.1348/000709904X22674

Breheny, R., Ferguson, H. J., & Katsos, N. (2013). Taking the epistemic step: Toward a model of on-line access to conversational implicatures. *Cognition, 126*(3), 423–440. https://doi.org/10.1016/j.cognition.2012.11.012

Butterfuss, R., Kendeou, P., McMaster, K. L., Orcutt, E., & Bulut, O. (2022). Question Timing, Language Comprehension, and Executive Function in Inferencing. *Scientific Studies of Reading, 26*(1), 61–78. https://doi.org/10.1080/10888438.2021.1901903

Cain, K., & Barnes, M. (2017). Reading comprehension: What develops and when? In K. Cain, D. Compton, & R. Parrila (Eds.), *Theories of Reading Development* (pp. 258–283). John Benjamins Publishing Company.

Cain, K., & Oakhill, J. (1999). Inference making ability and its relation to comprehension failure in young children. *Reading and Writing, 11*(5), 489–503.

Cain, K., & Oakhill, J. (2014). Reading comprehension and vocabulary: Is vocabulary more important for some aspects of comprehension? *L’Année Psychologique, 114*(4), 647–662. https://doi.org/10.3917/anpsy.144.0647

Cain, K., Oakhill, J., & Bryant, P. (2004). Children's Reading Comprehension Ability: Concurrent Prediction by Working Memory, Verbal Ability, and Component Skills. *Journal of Educational Psychology, 96*(1), 31–42. https://doi.org/10.1037/0022-0663.96.1.31

Cain, K., Oakhill, J., & Lemmon, K. (2004). Individual Differences in the Inference of Word Meanings From Context: The Influence of Reading Comprehension, Vocabulary Knowledge, and Memory Capacity. *Journal of Educational Psychology, 96*, 671–681. https://doi.org/10.1037/0022-0663.96.4.671

Cain, K., Oakhill, J. V., Barnes, M. A., & Bryant, P. E. (2001). Comprehension skill, inference-making ability, and their relation to knowledge. *Memory & Cognition, 29*(6), 850–859. https://doi.org/10.3758/BF03196414

Carrow-Woolfolk, E. (1999). *CASL: Comprehensive Assessment of Spoken Language*. American Guidance Services Circle Pines, MN.

Castles, A., Rastle, K., & Nation, K. (2018). Ending the Reading Wars: Reading Acquisition From Novice to Expert. *Psychological Science in the Public Interest, 19*(1), 5–51. https://doi.org/10.1177/1529100618772271

Chapman, S., & Clark, B. (2019). *Pragmatics and Literature*. John Benjamins Publishing Company.

Conti-Ramsden, G., Mok, P., Durkin, K., Pickles, A., Toseeb, U., & Botting, N. (2019). Do emotional difficulties and peer problems occur together from childhood to adolescence? The case of children with a history of developmental language disorder (DLD). *European Child & Adolescent Psychiatry, 28*(7), 993–1004. https://doi.org/10.1007/s00787-018-1261-6

Coplan, R. J., & Weeks, M. (2009). Shy and soft-spoken: Shyness, pragmatic language, and socio-emotional adjustment in early childhood. *Infant and Child Development, 18*(3), 238–254. https://doi.org/10.1002/icd.622

Currie, N. K., & Cain, K. (2015). Children's inference generation: The role of vocabulary and working memory. *Journal of Experimental Child Psychology, 137*, 57–75. https://doi.org/10.1016/j.jecp.2015.03.005

Currie, N. K., & Cain, K. (2023). Developmental differences in children's generation of knowledge-based inferences. *Discourse Processes, 60*(6), 440–456. https://doi.org/10.1080/0163853X.2023.2225980

Cutting, A. L., & Dunn, J. (1999). Theory of Mind, Emotion Understanding, Language, and Family Background: Individual Differences and Interrelations. *Child Development, 70*(4), 853–865. https://doi.org/10.1111/1467-8624.00061

Davies, C., Ebbels, S., Nicoll, H., Syrett, K., White, S., & Zuniga-Montanez, C. (2023). Supporting Adjective Learning by Children with Developmental Language Disorder: Enhancing Metalinguistic Approaches. *International Journal of Language & Communication Disorders, 58*(2), 629–650.

Davies, C., & Katsos, N. (2010). Over-informative children: Production/comprehension asymmetry or tolerance to pragmatic violations? *Lingua, 120*(8), 1956–1972.

Davies, C., & Kreysa, H. (2018). Look before you speak: Children's integration of visual information into informative referring expressions. *Journal of Child Language, 45*(5), 1116–1143. https://doi.org/10.1017/S0305000918000120

Davies, C., McGillion, M., Rowland, C., & Matthews, D. (2019). Can inferencing be trained in preschoolers using shared book-reading? A randomised controlled trial of parents' inference-eliciting questions on oral inferencing ability. *Journal of Child Language,* 1–25. https://doi.org/10.1017/S0305000919000801

Davies, C., Syrett, K., Taylor, L., Wilkes, S., & Zuniga-Montanez, C. (2022). Supporting adjective learning in 5-7 year olds across the curriculum: Insights from psychological research. *Language and Linguistics Compass, 16*(11), e12476.

Dawes, E., Leitão, S., Claessen, M., & Kane, R. (2019). A randomized controlled trial of an oral inferential comprehension intervention for young children with developmental language disorder. *Child Language Teaching and Therapy, 35*(1), 39–54. https://doi.org/10.1177/0265659018815736

Dawson, N., Hsiao, Y., Tan, A. W. M., Banerji, N., & Nation, K. (2021). Features of lexical richness in children's books: Comparisons with child-directed speech. *Language Development Research, 1*(1), Article 1. https://doi.org/10.34842/5we1-yk94

De Cat, C. (2015). The cognitive underpinnings of referential abilities. In L. Serratrice & S. E. M. Allen (Eds.), *The Acquisition of Reference* (Vol. 15, pp. 263–283). John Benjamins Publishing Company.

Degen, J., & Tanenhaus, M. (2014). Processing Scalar Implicature: A Constraint-Based Approach. *Cognitive Science, 39*(4), 667–710. https://doi.org/10.1111/cogs.12171

Department for Education. (2013). *English programmes of study: Key stages 1 and 2 National curriculum in England.* Department for Education.

https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/335186/PRIMARY_national_curriculum_-_English_220714.pdf

Diamond, A. (2006). The early development of executive functions. In E. Bialystok & F. I. M. Craik (Eds.), *Lifespan cognition: Mechanisms of change.* (pp. 70–95). Oxford University Press.

Dore, R. A., Amendum, S. J., Golinkoff, R. M., & Hirsh-Pasek, K. (2018). Theory of Mind: A Hidden Factor in Reading Comprehension? *Educational Psychology Review, 30*(3), 1067–1089. https://doi.org/10.1007/s10648-018-9443-9

Dupuy, L., Stateva, P., Andreetta, S., Cheylus, A., Deprez, V., van der Henst, J.-B., Jayez, J., Stepanov, A., & Reboul, A. (2019). Pragmatic abilities in bilinguals: The case of scalar implicatures. *Linguistic Approaches to Bilingualism, 9*(2), 314–340. https://doi.org/10.1075/lab.17017.dup

Eiteljoerge, S. F., V., Pouscoulous, N., & Lieven, E. V. M. (2018). Some Pieces Are Missing: Implicature Production in Children. *Frontiers in Psychology, 9*. https://doi.org/10.3389/fpsyg.2018.01928

Elbro, C., & Buch-Iversen, I. (2013). Activation of Background Knowledge for Inference Making: Effects on Reading Comprehension. *Scientific Studies of Reading, 17*(6), 435–452. https://doi.org/10.1080/10888438.2013.774005

Elleman, A. M. (2017). Examining the impact of inference instruction on the literal and inferential comprehension of skilled and less skilled readers: A meta-analytic review. *Journal of Educational Psychology, 109*(6), 761–781. https://doi.org/10.1037/edu0000180

Falkum, I. L. (2022). The development of non-literal uses of language: Sense conventions and pragmatic competence. *Journal of Pragmatics, 188*, 97–107. https://doi.org/10.1016/j.pragma.2021.12.002

Filippova, E. (2014). Irony production and understanding. In Matthews, Danielle (Ed.), *Pragmatic Development in First Language Acquisition* (pp. 261–278). John Benjamins.

Follmer, D. J. (2018). Executive Function and Reading Comprehension: A Meta-Analytic Review. *Educational Psychologist, 53*(1), 42–60. https://doi.org/10.1080/00461520.2017.1309295

Foppolo, F., Mazzaggio, G., Panzeri, F., & Surian, L. (2020). Scalar and ad-hoc pragmatic inferences in children: Guess which one is easier. *Journal of Child Language*, 1–23. https://doi.org/10.1017/S030500092000032X

Fortier, M., Kellier, D., Flecha, M. F., & Frank, M. C. (under review). *Ad-hoc pragmatic implicatures among Shipibo-Konibo children in the Peruvian Amazon* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/x7ad9

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science, 336*(6084), 998–998.

Franke, M., & Jäger, G. (2016). Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics. *Zeitschrift Für Sprachwissenschaft, 35*(1), 3–44. https://doi.org/10.1515/zfs-2016-0002

Freed, J., & Cain, K. (2017). Assessing school-aged children's inference-making: The effect of story test format in listening comprehension. *International Journal of Language & Communication Disorders, 52*(1), 95–105. https://doi.org/10.1111/1460-6984.12260

Geurts, B., & Rubio-Fernández, P. (2015). Pragmatics and Processing. *Ratio, 28*(4), 446–469. https://doi.org/10.1111/rati.12113

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences, 20*(11), 818–829. https://doi.org/10.1016/j.tics.2016.08.005

Graesser, A., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text representation. *Psychological Review, 101*(3), 397–95. https://doi.org/10.1037/0033-295X.101.3.371

Grice, H. P. (1975). Logic and conversation. In R. Stainton (Ed.), *Perspectives in the Philosophy of Language* (pp. 41–58). Broadview Press.

Helland, W. A., Lundervold, A. J., Heimann, M., & Posserud, M.-B. (2014). Stable associations between behavioral problems and language impairments across childhood – The importance of pragmatic language problems. *Research in Developmental Disabilities, 35*(5), 943–951. https://doi.org/10.1016/j.ridd.2014.02.016

Hochstein, L., Bale, A., Fox, D., & Barner, D. (2016). Ignorance and Inference: Do Problems with Gricean Epistemic Reasoning Explain Children's Difficulty with Scalar Implicature? *Journal of Semantics, 33*(1), 107–135. https://doi.org/10.1093/jos/ffu015

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing, 2*(2), 127–160. https://doi.org/10.1007/BF00401799

Horowitz, A. C., & Frank, M. C. (2016). Children's pragmatic inferences as a route for learning about the world. *Child Development, 87*(3), 807–819. https://doi.org/10.1111/cdev.12527

Horowitz, A. C., Schneider, R. M., & Frank, M. C. (2018). The Trouble With Quantifiers: Exploring Children's Deficits in Scalar Implicature. *Child Development, 89*(6), E572–E593. https://doi.org/10.1111/cdev.13014

Huang, Y. T., & Snedeker, J. (2009). Online interpretation of scalar quantifiers: Insight into the semantics–pragmatics interface. *Cognitive Psychology, 58*(3), 376–415. https://doi.org/10.1016/j.cogpsych.2008.09.001

Hughes, C., Deater-Deckard, K., & Cutting, A. L. (1999). 'Speak roughly to your little boy'? Sex Differences in the Relations Between Parenting and Preschoolers' Understanding of Mind. *Social Development, 8*(2), 143–160. https://doi.org/10.1111/1467-9507.00088

Joseph, H., Wonnacott, E., & Nation, K. (2021). Online inference making and comprehension monitoring in children during reading: Evidence from eye movements. *Quarterly Journal of Experimental Psychology, 74*(7), 1202–1224. https://doi.org/10.1177/1747021821999007

Kampa, A., & Papafragou, A. (2020). Four-year-olds incorporate speaker knowledge into pragmatic inferences. *Developmental Science, 23*(3), e12920. https://doi.org/10.1111/desc.12920

Katsos, N., & Andrés-Roqueta, C. (2021). Where next for pragmatics and mind reading? A situation-based view (Response to Kissine). *Language, 97*(3), e184–e197. https://doi.org/10.1353/lan.2021.0036

Katsos, N., Cummins, C., Ezeizabarrena, M.-J., Gavarró, A., Kraljević, J. K., Hrzica, G., Grohmann, K. K., Skordi, A., López, K. J. de, Sundahl, L., Hout, A. van, Hollebrandse, B., Overweg, J., Faber, M., Koert, M. van, Smith, N., Vija, M., Zupping, S., Kunnari, S., ... Noveck, I. (2016). Cross-linguistic patterns in the acquisition of quantifiers. *Proceedings of the National Academy of Sciences, 113*(33), 9244–9249. https://doi.org/10.1073/pnas.1601341113

Kendeou, P. (2015). A general inference skill. In A. E. Cook, E. J. O'Brien, & J. Lorch Robert F. (Eds.), *Inferences during Reading* (pp. 160–181). Cambridge University Press.

Kendeou, P., McMaster, K. L., Butterfuss, R., Kim, J., Bresina, B., & Wagner, K. (2020). The Inferential Language Comprehension (iLC) Framework: Supporting

Children's Comprehension of Visual Narratives. *Topics in Cognitive Science, 12*(1), 256–273. https://doi.org/10.1111/tops.12457

Kendeou, P., McMaster, K. L., & Christ, T. J. (2016). Reading Comprehension: Core Components and Processes. *Policy Insights from the Behavioral and Brain Sciences, 3*(1), 62–69. https://doi.org/10.1177/2372732215624707

Kidd, E., Donnelly, S., & Christiansen, M. H. (2018). Individual Differences in Language Acquisition and Processing. *Trends in Cognitive Sciences, 22*(2), 154–169. https://doi.org/10.1016/j.tics.2017.11.006

Kim, Y.-S. (2020). Theory of mind mediates the relations of language and domain-general cognitions to discourse comprehension. *Journal of Experimental Child Psychology, 194,* 104813. https://doi.org/10.1016/j.jecp.2020.104813

Kim, Y.-S., & Phillips, B. (2014). Cognitive Correlates of Listening Comprehension. *Reading Research Quarterly, 49*(3), 269–281. https://doi.org/10.1002/rrq.74

Kintsch, W. (1998). *Comprehension: A Paradigm for Cognition.* Cambridge University Press.

Kispal, A. (2008). Effective Teaching of Inference Skills for Reading. Literature Review. Research Report DCSF-RR031. *National Foundation for Educational Research.* https://eric.ed.gov/?id=ED501868

Köder, F., & Falkum, I. L. (2020). Children's metonymy comprehension: Evidence from eye-tracking and picture selection. *Journal of Pragmatics, 156,* 191–205. https://doi.org/10.1016/j.pragma.2019.07.007

Köder, F., & Falkum, I. L. (2021). Irony and Perspective-Taking in Children: The Roles of Norm Violations and Tone of Voice. *Frontiers in Psychology, 12.* https://doi.org/10.3389/fpsyg.2021.624604

Kuijper, S. J. M., Hartman, C. A., & Hendriks, P. (2021). Children's Pronoun Interpretation Problems Are Related to Theory of Mind and Inhibition, But Not Working Memory. *Frontiers in Psychology, 12.* https://doi.org/10.3389/fpsyg.2021.610401

Language and Reading Research Consortium, Currie, N. K., & Muijselaar, M. M. L. (2019). Inference making in young children: The concurrent and longitudinal contributions of verbal working memory and vocabulary. *Journal of Educational Psychology, 111*(8), 1416–1431. https://doi.org/10.1037/edu0000342

Language and Reading Research Consortium, & Muijselaar, M. M. L. (2018). The Dimensionality of Inference Making: Are Local and Global Inferences Distinguishable?

Scientific Studies of Reading, 22(2), 117–136. https://doi.org/10.1080/10888438.2017.1371179

Lazaridou-Chatzigoga, D., Katsos, N., & Stockall, L. (2019). Generalizing About Striking Properties: Do Glippets Love to Play With Fire? *Frontiers in Psychology, 10,* 1971. https://doi.org/10.3389/fpsyg.2019.01971

Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature.* MIT Press.

Matthews, D. (2014). *Pragmatic Development in First Language Acquisition* (Vol. 10). John Benjamins Publishing Company.

Matthews, D., Biney, H., & Abbot-Smith, K. (2018). Individual differences in children's pragmatic ability: A review of associations with formal language, social cognition, and executive functions. *Language Learning and Development, 14*(3), 186–223. https://doi.org/10.1080/15475441.2018.1455584

Matthews, D., Butcher, J., Lieven, E., & Tomasello, M. (2012). Two-and Four-Year-Olds Learn to Adapt Referring Expressions to Context: Effects of Distracters and Feedback on Referential Communication. *Topics in Cognitive Science, 4*(2), 184–210. https://doi.org/10.1111/j.1756-8765.2012.01181.x

Mazzarella, D., & Pouscoulous, N. (2021). Pragmatics and epistemic vigilance: A developmental perspective. *Mind & Language, 36*(3), 355–376. https://doi.org/10.1111/mila.12287

McNamara, D. S., & Magliano, J. (2009). Toward a Comprehensive Model of Comprehension. *Psychology of Learning and Motivation, 51,* 297–384. https://doi.org/10.1016/S0079-7421(09)51009-2

Mok, P. L. H., Pickles, A., Durkin, K., & Conti-Ramsden, G. (2014). Longitudinal trajectories of peer relations in children with specific language impairment. *Journal of Child Psychology and Psychiatry, 55*(5), 516–527. https://doi.org/10.1111/jcpp.12190

Nation, K. (2005). Children's Reading Comprehension Difficulties. In *The Science of Reading: A Handbook* (pp. 248–265). Blackwell Publishing. https://doi.org/10.1002/9780470757642.ch14

National Governors Association Center for Best Practices, Council of Chief State School Officers. (2010). *Common Core State Standards for English Language Arts.* http://corestandards.org/

Norbury, C. F., & Bishop, D. V. M. (2002). Inferential processing and story recall in children with communication problems: A comparison of specific language impairment, pragmatic language impairment and high-functioning autism. *International Journal of Language & Communication Disorders, 37*(3), 227–251. https://doi.org/10.1080/13682820210136269

Nordmeyer, A. E., Yoon, E. J., & Frank, M. C. (2016). Distinguishing processing difficulties in inhibition, implicature, and negation. *Proceedings of the 37th Annual Conference of the Cognitive Science Society,* 2789–2794.

Nouwens, S., Groen, M. A., Kleemans, T., & Verhoeven, L. (2021). How executive functions contribute to reading comprehension. *British Journal of Educational Psychology, 91*(1), e12355. https://doi.org/10.1111/bjep.12355

Noveck, I. (2018). *Experimental Pragmatics: The Making of a Cognitive Science*. Cambridge University Press.

Oakes, L. M., & Rakison, D. H. (2019). *Developmental Cascades: Building the Infant Mind.* Oxford University Press.

Oakhill, J. (1982). Constructive processes in skilled and less skilled comprehenders' memory for sentences. *British Journal of Psychology, 73*(1), 13–20. https://doi.org/10.1111/j.2044-8295.1982.tb01785.x

Oakhill, J. (2020). Four Decades of Research into Children's Reading Comprehension: A Personal Review. *Discourse Processes, 57*(5–6), 402–419. https://doi.org/10.1080/0163853X.2020.1740875

Oakhill, J., & Cain, K. (2012). The Precursors of Reading Ability in Young Readers: Evidence From a Four-Year Longitudinal Study. *Scientific Studies of Reading, 16*(2), 91–121. https://doi.org/10.1080/10888438.2010.529219

Oakhill, J., Cain, K., & Elbro, C. (2015). *Understanding and teaching reading comprehension: A handbook.* Routledge.

Oakhill, J., & Yuill, N. (1986). Pronoun Resolution in Skilled and Less-Skilled Comprehenders: Effects of Memory Load and Inferential Complexity. *Language and Speech, 29*(1), 25–37. https://doi.org/10.1177/002383098602900104

O'Brien, E. J., Cook, A. E., & Lorch, J., Robert F. (Eds.). (2015). *Inferences during Reading*. Cambridge University Press. https://doi.org/10.1017/CBO9781107279186

OECD. (2019). *PISA 2018 Insights and Interpretations*. OECD Publishing.

O'Neill, D. K. (1996). Two-year-old children's sensitivity to a parent's knowledge state when making requests. *Child Development*, 659–677. https://doi.org/10.1111/j.1467-8624.1996.tb01758.x

Papafragou, A., & Skordos, D. (2016). Scalar Implicature. In J. Lidz, W. Snyder, & J. Pater (Eds.), *Oxford Handbook of Developmental Linguistics* (pp. 611–632). Oxford University Press.

Paris, A. H., & Paris, S. G. (2003). Assessing narrative comprehension in young children. *Reading Research Quarterly, 38*(1), 36–76. https://doi.org/10.1598/RRQ.38.1.3

Peng, P., Barnes, M., Wang, C., Wang, W., Li, S., Swanson, H. L., Dardick, W., & Tao, S. (2018). A meta-analysis on the relation between reading and working memory. *Psychological Bulletin, 144,* 48–76. https://doi.org/10.1037/bul0000124

Phelps-Terasaki, D., & Phelps-Gunn, T. (1992). *Test of pragmatic language: Examiner's manual.* Pro-Ed.

Pouscoulous, N., Noveck, I. A., Politzer, G., & Bastide, A. (2007). A developmental investigation of processing costs in implicature production. *Language Acquisition, 14*(4), 347–375. https://doi.org/10.1080/10489220701600457

Pouscoulous, N., & Tomasello, M. (2020). Early birds: Metaphor understanding in 3-year-olds. *Journal of Pragmatics, 156,* 160–167. https://doi.org/10.1016/j.pragma.2019.05.021

Pyykkönen, P., Matthews, D., & Järvikivi, J. (2010). Three-year-olds are sensitive to semantic prominence during online language comprehension: A visual world study of pronoun resolution. *Language and Cognitive Processes, 25*(1), 115–129. https://doi.org/10.1080/01690960902944014

Rabagliati, H., & Robertson, A. (2017). How do children learn to avoid referential ambiguity? Insights from eye-tracking. *Journal of Memory and Language, 94,* 15–27. https://doi.org/10.1016/j.jml.2016.09.007

Sauerland, U. (2004). Scalar implicatures in complex sentences. *Linguistics and Philosophy, 27*(3), 367–391. https://doi.org/10.1023/B:LING.0000023378.71748.db

Scarborough, H. (2009). Connecting early language and literacy to later reading (dis)abilities: Evidence, theory, and practice. In F. Fletcher-Campbell, J. Soler, & G. Reid (Eds.), *Approaching Difficulties in Literacy Development: Assessment, Pedagogy and Programmes* (Vol. 10, pp. 23–38). SAGE.

Schuch, S., Philipp, A. M., Maulitz, L., & Koch, I. (2022). On the reliability of behavioral measures of cognitive control: Retest reliability of task-inhibition effect, task-preparation effect, Stroop-like interference, and conflict adaptation effect. *Psychological Research, 86*(7), 2158–2184. https://doi.org/10.1007/s00426-021-01627-x

Schulze, C., Grassmann, S., & Tomasello, M. (2013). 3-Year-Old Children Make Relevance Inferences in Indirect Verbal Communication. *Child Development, 84*(6), 2079–2093. https://doi.org/10.1111/cdev.12093

Searle, J. R. (1969). *Speech Acts: An Essay in the Philosophy of Language.* Cambridge University Press.

Serratrice, L. (2007). Cross-linguistic influence in the interpretation of anaphoric and cataphoric pronouns in English–Italian bilingual children. *Bilingualism: Language and Cognition, 10*(3), 225–238. https://doi.org/10.1017/S1366728907003045

Serratrice, L., & Allen, S. E. M. (2015). *The Acquisition of Reference.* John Benjamins Publishing Company.

Siegal, M., Surian, L., Matsuo, A., Geraci, A., Iozzi, L., Okumura, Y., & Itakura, S. (2010). Bilingualism accentuates children's conversational understanding. *PLoS One, 5*(2), e9004.

Silva, M., & Cain, K. (2015). The relations between lower and higher level comprehension skills and their role in prediction of early reading comprehension. *Journal of Educational Psychology, 107*(2), 321–331. https://doi.org/10.1037/a0037769

Skordos, D., & Papafragou, A. (2016). Children's derivation of scalar implicatures: Alternatives and relevance. *Cognition, 153*(6–18). https://doi.org/10.1016/j.cognition.2016.04.006

Smith, R., Snow, P., Serry, T., & Hammond, L. (2021). The Role of Background Knowledge in Reading Comprehension: A Critical Review. *Reading Psychology, 42*(3), 214–240. https://doi.org/10.1080/02702711.2021.1888348

Song, H., & Fisher, C. (2007). Discourse prominence effects on 2.5-year-old children's interpretation of pronouns. *Lingua, 117*(11), 1959–1987. https://doi.org/10.1016/j.lingua.2006.11.011

Sperber, D., & Wilson, D. (1995). *Relevance: Communication and Cognition.* Wiley.

St Clair, M. C., Pickles, A., Durkin, K., & Conti-Ramsden, G. (2011). A longitudinal study of behavioral, emotional and social difficulties in individuals with a history of specific language impairment (SLI). *Journal of Communication Disorders, 44*(2), 186–199. https://doi.org/10.1016/j.jcomdis.2010.09.004

Staatsministerium für Kultus Freistaat Sachsen. (2019). *Lehrplan Grundschule Deutsch*. Sächsisches Staatsministerium für Kultus. www.bildung.sachsen.de/apps/lehrplandb/

Stiller, A. J., Goodman, N. D., & Frank, M. C. (2015). Ad-hoc Implicature in Preschool Children. *Language Learning and Development, 11*(2), 176–190. https://doi.org/10.1080/15475441.2014.927328

Such, C. (2021). *The Art and Science of Teaching Primary Reading*. SAGE.

Van den Broek, P., Beker, K., & Oudega, M. (2015). Inference generation in text comprehension: Automatic and strategic processes in the construction of a mental representation. In E. J. O'Brien, A. E. Cook, & J. Lorch Robert F. (Eds.), *Inferences during Reading* (pp. 94–121). Cambridge University Press. https://doi.org/10.1017/CBO9781107279186

van den Broek, P., Fletcher, C. R., & Risden, K. (1993). Investigations of inferential processes in reading: A theoretical and methodological integration. *Discourse Processes, 16*(1–2), 169–180. https://doi.org/10.1080/01638539309544835

Van Der Maas, H. L. J., Dolan, C. V., Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. J. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review, 113*(4), 842–861. https://doi.org/10.1037/0033-295X.113.4.842

Veenstra, A., & Katsos, N. (2018). Assessing the comprehension of pragmatic language: Sentence judgment tasks. In A. H. Jucker, K. P. Schneider, & W. Biblitz (Eds.), *Methods in Pragmatics* (pp. 257–279). de Gruyter Mouton.

West, G., Snowling, M. J., Lervåg, A., Buchanan-Worster, E., Duta, M., Hall, A., McLachlan, H., & Hulme, C. (2021). Early language screening and intervention can be delivered successfully at scale: Evidence from a cluster randomized controlled trial. *Journal of Child Psychology and Psychiatry, 62*(12), 1425–1434. https://doi.org/10.1111/jcpp.13415

Whatmuff, T. (n.d.). *Inference Training*. https://www.northamptonshire.gov.uk/councilservices/children-families-education/schools-and-education/information-for-school-staff/Documents/NCRP/Part%203i%29%20inference%20intervention%20PPT%20handout.pdf

Wilson, A. C., & Bishop, D. V. M. (2022). A novel online assessment of pragmatic and core language skills: An attempt to tease apart language domains in children. *Journal of Child Language, 49*(1), 38–59. https://doi.org/10.1017/S0305000920000690

Wilson, D., & Sperber, D. (2012). *Meaning and relevance*. Cambridge University Press.

Wilson, E., & Katsos, N. (2020). Acquiring implicatures. In K. Schneider & E. Ifantidou (Eds.), *Developmental and Clinical Pragmatics* (pp. 119–148). de Gruyter Mouton.

Wilson, E., & Katsos, N. (2022). Pragmatic, linguistic and cognitive factors in young children's development of quantity, relevance and word learning inferences. *Journal of Child Language, 49*(6), 1065–1092. https://doi.org/doi:10.1017/S0305000921000453

Wilson, E., Lawrence, R., & Katsos, N. (2022). The role of perspective-taking in children's quantity implicatures. *Language Learning and Development, 19*(2), 167–187. https://doi.org/10.1080/15475441.2022.2050236

World Literacy Foundation. (2018). *The Economic and Social Cost of Illiteracy: A white paper by the world literacy foundation*. www.worldliteracyfoundation.org/wp-content/uploads/2021/07/TheEconomicSocialCostofIlliteracy-2.pdf

Yuill, N., Oakhill, J., & Parkin, A. (1989). Working memory, comprehension ability and the resolution of text anomaly. *British Journal of Psychology, 80*(3), 351–361. https://doi.org/10.1111/j.2044-8295.1989.tb02325.x

Zajączkowska, M., & Abbot-Smith, K. (2020). "Sure I'll help—I've just been sitting around doing nothing at school all day": Cognitive flexibility and child irony interpretation. *Journal of Experimental Child Psychology, 199*, 1170–1188. https://doi.org/10.1016/j.jecp.2020.104942

Zajączkowska, M., Abbot-Smith, K., & Kim, C. S. (2020). Using shared knowledge to determine ironic intent; a conversational response paradigm. *Journal of Child Language, 47*, 1170–1188. https://doi.org/10.1017/S0305000920000045

Zhao, S., Ren, J., Frank, M. C., & Zhou, P. (2021). The Development of Quantity Implicatures in Mandarin-Speaking Children. *Language Learning and Development, 17*(4), 343–365. https://doi.org/10.1080/15475441.2021.1886935

Authorship and Contributorship Statement

All authors conceived the project, and wrote and revised the manuscript. All authors approved the final version of the manuscript and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. Authors excepting EW appear in alphabetical order.

Acknowledgements

Elspeth Wilson was funded by the Economic and Social Research Council (grant ES/S010203/1); the support of the Economic and Social Research Council (ESRC) is gratefully acknowledged. She thanks Katharina Risse for discussion on literacy curricula in Germany. We are very grateful for the insights and engagement of participants of the workshop, 'Inferencing in conversation and in reading: insights from linguistics, psychology and education' (Cambridge, 17 January 2020), discussion at which inspired the conception and content of this paper.

License

Investigating how vocabulary relates to different dimensions of family socioeconomic circumstance across developmental and historical time.

Emma Thornton[1]
Danielle Matthews[2]
Praveetha Patalay[3]
Colin Bannard[4]

Manchester Institute of Education, University of Manchester, United Kingdom.[1]
Department of Psychology, University of Sheffield, United Kingdom.[2]
Social Research Institute and Department of Population Science and Experimental Medicine, University College London, United Kingdom.[3]
Department of Linguistics and English Language, University of Manchester, United Kingdom.[4]

Abstract: Social inequalities in child vocabulary persist, despite decades of efforts to understand and reduce them. Different dimensions of socioeconomic circumstances (SEC), such as parent education, income, occupational status, wealth, and relative neighbourhood deprivation, are likely to represent different mechanisms of effects on child vocabulary. We investigate which aspects of SEC relate to vocabulary, and whether relations are stable over developmental and historical time. Data from two large, national datasets were analysed: the 1970 British Cohort Study (born 1970; N= 14,851) and the Millennium Cohort Study (born 2000-01; N=17,070). Substantial individual differences in vocabulary (ages 3–14) were explained by multiple indicators each making a unique contribution, most notably parent education (partial R^2:6.4%-8.5%), income (partial R^2: 4.3%-6.4%), and occupation (partial R^2: 5.3-8.1). Inequalities were generally stable over developmental and historical time. However, findings suggest a need to focus on widening inequalities at the start and end of compulsory schooling.

Keywords: Language, cognitive ability, child, adolescent, cross-cohort, generation, ontogeny, social inequality, social class.

Corresponding author(s): Emma Thornton, Manchester Institute of Education, University of Manchester, Oxford Road, Manchester, M13 9PL, UK. Email: Emma.Thornton@manchester.ac.uk

ORCID ID(s): https://orcid.org/0000-0003-4623-9538 (ET); https://orcid.org/0000-0003-3562-9549 (DM); https://orcid.org/0000-0002-5341-3461 (PP); https://orcid.org/0000-0001-5579-5830 (CB).

Citation: Thornton, E., Matthews, D., Patalay, P., & Bannard, C. (2024). Investigating how vocabulary relates to different dimensions of family socioeconomic circumstance across developmental and historical time. *Language Development Research*, *4*(1), 80–174. https://doi.org/10.34842/mhqh-9g10

Introduction

Children need good language skills in order to be able to access education and, in turn, the labour market (Law, Charlton, & Asmussen, 2017; Oxford University Press, 2018). For decades, studies have observed social inequalities in vocabulary size (Hart & Risley, 1995; Pace, Luo, Hirsh-Pasek, & Golinkoff, 2017) and policy makers have sought educational interventions to reduce these disparities (Bercow, 2018). Yet randomised controlled trials suggest that such interventions have mixed success (Law, Charlton, Dockrell, et al., 2017). To assist in better directing future research and better targeting interventions, we address three fundamental questions using large, nationally representative, longitudinal UK datasets. First, are all indicators of socioeconomic circumstance (SEC) equal in predicting vocabulary outcomes? Second, does the relation between SEC and language development stay constant over developmental time? And third, is the relation between SEC and language development changing over historical time as our economy becomes increasingly knowledge-based and hourglass-shaped?

While caregiver education, occupational status, income, wealth, and neighbourhood disadvantage statistics are all often used as interchangeable indicators of SEC, each dimension reflects access to different resources that may affect language development (Duncan & Magnuson, 2012). Some have argued that caregiver education is the most relevant SEC indicator for language development as it is most directly related to the *quality* of the language learning environment and/or language related genetic factors (Hirsh-Pasek et al., 2015; Hoff, 2013; Hoff, Laursen, & Bridges, 2012). However, no empirical work has explicitly tested this claim in nationally representative samples and there are plausible pathways by which other indicators of SEC may also exert effects on vocabulary. First, income may affect language development through the availability of learning resources in the household (Duncan, Magnuson, & Votruba-Drzal, 2017; Washbrook & Waldfogel, 2011). Second, the family stress model posits that economic difficulty can influence parenting through its harmful effect on emotions, behaviours and relationships (Conger & Donnellan, 2007). This in turn can affect language development via the interactions parents have with their children (Perkins, Finegood, & Swain, 2013). Therefore, family wealth could be a protective mechanism, acting as a safeguard against any negative effects of sudden income losses, such as unexpected unemployment (Grinstein-Weiss, Williams Shanks, & Beverly, 2014; Killewald, Pfeffer, & Schachner, 2017). Third, occupational status reflects one's social position in the labour market, as well as power and status (Sullivan, Ketende, & Joshi, 2013). It is thought that people's social networks generally consist of people who are similar to them in terms of occupational status, known as occupational homophily. (Griffiths, Lambert, & Tranmer, 2011; McPherson, Smith-Lovin, & Cook, 2001). This may be indirectly related to language development, as children will adopt language used by their parents when talking to them and when talking to individuals in their social network (Sullivan, 2007). Finally, developmental theory emphasises how the immediate caregiving environment is nested within broader societal and cultural spheres (Bronfenbrenner, 1979; Rowe & Weisleder, 2020). As a proxy for this wider environment, neighbourhood-level statistics (such as the UK Indices of Multiple Deprivation) may additionally predict language development (Bennetts et al., 2022; Neuman, Kaefer, & Pinkham, 2018). Directly comparing the predictive value of different SEC indicators can help us understand why vocabulary inequalities exist and which mechanisms to further explore and target if aiming to support development. Our first goal was thus to test whether five key indicators of SEC (caregiver education, income, wealth, occupational status and neighbourhood deprivation) each predict unique variance in child vocabulary and how much relative variance they predict.

Compelling arguments have been made in favour of early intervention to prevent social disadvantage affecting language before children reach formal education (e.g., Doyle, Harmon, Heckman, & Tremblay, 2009), yet there is also evidence that the SEC gap in vocabulary is pronounced among adolescents (Spencer, Clegg, & Stackhouse, 2012; Sullivan & Brown, 2015). In fact, we do not know if or when the word gap shrinks or widens as children grow up. Nor do we know whether the predictive value of different SEC indicators remains stable over developmental time. For example, while caregiver education may be important during the early years, it has been proposed that family wealth may be a more important predictor of outcomes in adolescence and early adulthood. This might be because wealth facilitates access to high quality secondary education or other forms of academic support (Pfeffer, 2018). It is thus possible that the relative effect of different dimensions of SEC changes throughout development. Our second goal was therefore to test whether social disparities in language development have narrowed or widened over developmental time, from early childhood to mid-adolescence, for a contemporary generation born at the start of the 21st Century.

Large societal changes in the UK have seen an increase in the proportion of parents who have attended university, and a reconfiguration of the economy such that fewer people are in middle-ranked jobs, with more in lower grade employment on the one hand and in the higher managerial and professional occupations on the other (often characterised as a move to an hourglass economy; Bolton, 2012; Holmes & Mayhew, 2012). Many more jobs are now also knowledge-based, making language and cognitive skills of great importance for the UK economy (Beddington et al., 2008; Deloitte, 2016), and putting pressure on parents to support their children's cognitive development to open doors to the labour market. Income inequality increased in the UK in the 1980s and 1990s, and at the start of the millennium, income polarisation appeared to increase (those with the highest average incomes appeared to experience the largest increases, whilst those with lower average incomes experienced declines in their income; Dorling et al., 2007). These broad shifts in society have the potential to change the association between different measures of SEC and language development. Our third goal was thus to test whether the relations between different SEC indicators and language development have become more or less pronounced over historical time, comparing children born at the turn of this century with those born in 1970.

In a series of pre-registered analyses, we met the first two goals by analysing data from the Millennium Cohort Study (17,070 children born between 2000-02; MCS2001). We then compared these contemporary trends with those in a cohort born 30 years prior using data from the 1970 British Cohort Study (15,817children born in 1970; BCS1970, and 16,020 children in the MCS2001). Both studies contain measures of vocabulary at multiple ages and we use these as indicators of general language ability. Since different measures of formal language tend to load on to the same factor (Fricke et al., 2017), vocabulary is likely to be a good proxy for broader language ability. Nonetheless, an exclusive focus on vocabulary has implications for the conclusions we can draw, and we return to this in the discussion section.

Method

Data

We used data from two large nationally representative UK birth cohort studies: the Millennium Cohort Study (MCS2001 cohort) and the 1970 British Cohort Study (BCS1970 cohort). Addressing research questions 1-3 involved analyses of the MCS2001 cohort data only, due to the availability of multiple SEC indicators in this cohort, allowing us to examine the unique contribution of different SEC indicators to inequalities in language ability in a contemporary cohort. In addressing research question 4 we used data from the MCS2001 and BCS1970 cohorts in a cross-cohort comparison. The use of these two datasets for a cross-cohort comparison allowed us to examine inequalities in language ability in two generations born 30 years apart, during a period which has seen changes to occupational and educational structures in the UK.

MCS2001. The Millennium Cohort Study is a longitudinal birth cohort study of 19,518 young people, from 19,244 families, born across England, Scotland, Wales and Northern Ireland between 2000-02 (Connelly & Platt, 2014). To date there have been seven sweeps of data collection conducted when cohort members were aged 9 months and ages 3, 5, 7, 11, 14, and 17. More information on the MCS2001 cohort can be found here: https://cls.ucl.ac.uk/cls-studies/millennium-cohort-study/.

BCS1970. The 1970 British Cohort Study is a longitudinal birth cohort study of 16,571 children who were born during one week in 1970 in England, Scotland and Wales (Elliott & Shepherd, 2006). It has 4 childhood sweeps (data collected at birth and 5, 10 and 16 years). More information on the BCS1970 cohort can be found here: https://cls.ucl.ac.uk/cls-studies/1970-british-cohort-study/

Sample Selection. We selected all cohort members with a response on at least one of the language tasks at the time points considered – ages 3, 5, 11 or 14 (RQ 1-3, MCS2001 cohort only) and age 5, 10 or 16 (BCS1970) and ages 5, 11 or 14 MCS2001) for the cross-cohort comparison. Where cohort members were twins, triplets, or there were multiple cohort members from the same family, one of these members was selected at random.

Measures

Vocabulary Measures (MCS2001 Cohort Only)

The MCS2001 cohort members completed a battery of cognitive tests throughout childhood and into early adolescence. Full details about the completed vocabulary tests can be found in Appendix A.

At ages 3, 5 and 11, subscales of the British Ability Scale II (BAS II) were completed (Elliott, Smith, & McCulloch, 1996). The British Ability Scales consist of a series of tests measuring cognitive ability and educational attainment, between ages 2 years 6 months to 7 years 11 months. Progression through these tests depends on performance, and poor performance may result in a different, easier set of items being administered. Cohort members were born over a 1.5 year period (September 2000-January 2002) and assessed over a range of months, so age at the time of testing may differ between cohort members. Therefore, we used t-scores (as published in the data), which are adjusted for item difficulty and age. These were converted to z scores for analyses.

Ages 3 & 5. Cohort members completed the Naming Vocabulary BAS II subscale, as a measure of expressive vocabulary. Cohort members were shown a series of images and were asked to name each item in the image (Moulton et al., 2020).

Age 11. Cohort members completed the Verbal Similarities BAS II subscale. This is a measure of verbal reasoning and verbal knowledge. Sets of three words were read out to the cohort member, usually by the interviewer, and cohort members had to say how the words were related to each other (Moulton, 2020).

Age 14. Word Activity task. This test was a subset of items from the Applied Psychology Unit (APU) Vocabulary Test (Closs, 1986). Cohort members were given a list of 20 target words, each presented alongside 5 other words. Cohort members had to choose the word which meant the same, or nearly the same as the target word, from the 5 options (Moulton, 2020). Total scores out of 20 were converted into z scores for analyses.

Vocabulary Measures (Cross-Cohort Comparison)

For the cross-cohort comparison, we considered vocabulary at three time points in each cohort: age 5 (both cohorts; defined as early language ability), ages 10/11 (BCS1970 and MCS2001 cohorts respectively, referred to as late childhood language ability) and ages 16/14 (BCS1970 and MCS2001 cohorts respectively, referred to as adolescent language ability). There is no age 3 data for the BCS1970 cohort, hence the earliest language measure considered in the cohort comparisons is age 5.

Early Language Ability. For the BCS1970 cohort, receptive vocabulary was measured at age 5 using the English Picture Vocabulary Test (EPVT), a UK version of the Peabody Picture Vocabulary Test (Brimer & Dunn, 1962; Dunn, Dunn, Bulheller, & Häcker, 1965). Cohort members were shown 56 sets of four diverse images and heard a specific word associated with each set of four images. They were asked to select one picture that matched the presented word and were awarded one point for every correct response. For the MCS2001 cohort, expressive vocabulary was measured using the naming vocabulary sub-test of the BAS II (Elliott et al., 1996). We adjusted for age in months at the time of the test in both cohorts. All scores and ages were converted to z scores for analyses.

Late Childhood Language Ability. When the BCS1970 cohort members were aged 10, they completed the BAS word similarities subscale (Elliott, Murray, & Pearson, 1979). The test was made up of 21 items, each of which consisted of three words. The teacher read these sets of items out loud and cohort members had to a) name another word that was consistent with the three words in the item and b) state how the words were related. In order to receive a point, cohort members had to correctly answer both parts of the question (Moulton, 2020). Details on the scoring of this vocabulary measure and the SPSS syntax used can be found in appendix 3 of "Childhood Cognition in the 1970 British Cohort Study" (Parsons, 2014). When MCS2001 cohort members were aged 11, they completed the BAS II verbal similarities subscale (detailed above). As already mentioned, test scores for the MCS2001 cohort were adjusted for item difficulty. In both cohorts, we controlled for age at the time of the test and converted all scores to z scores.

Adolescent Language Ability. When aged 16, BCS1970 cohort members completed the APU Vocabulary Test (Closs, 1986). This consisted of 75 items: an item consisted of a target word, presented with a multiple-choice list, from which cohort members had to select a word that meant the same as the target word (Moulton, 2020). These items got progressively harder throughout the test. Details on the scoring of this vocabulary test can be found in appendix 3 (Parsons, 2014). When MCS2001 cohort members were aged 14, they completed the Word Activity Task (detailed above). Words used in the Word Activity Task were a subset of the words used in the BCS1970cohort Vocabulary Test, which cohort members completed aged 16 (Moulton, 2020). Scores were adjusted for age and converted to z scores for analyses.

Measures Of Socioeconomic Position (MCS2001 Cohort Only)

Five indicators of family SEC were used: parent education, family income, wealth, occupational status, and relative neighbourhood deprivation. Operationalisation of these variables is as follows:

Parent Education. As a measure of parent's education when cohort members were aged 3, highest parent NVQ (National Vocational Qualification) level was used (both academic and vocational qualifications derived into NVQ levels 1-5, with level 5 equating to higher qualifications). It is worth noting that the NVQ levels derived in MCS2001 data differ from those defined by the UK Government (https://www.gov.uk/what-different-qualification-levels-mean/list-of-qualification-levels). In the MCS2001 data, these are:

NVQ level 0: none of these/other qualifications
NVQ level 1: GCSE grades D-G, NVQ/ SVQ/ GSVQ level 1
NVQ level 2: GCSE grades A-C, trade apprenticeships, NVQ/ SVQ/ GSVQ level 2
NVQ level 3: A/ AS/ S levels, NVQ/ SVQ/ GSVQ level 3
NVQ level 4: first degree, diplomas in higher education, professional qualifications at degree level
NVQ level 5: higher degree

To contextualise for readers not familiar with the UK system, GCSEs (or the Scottish equivalent) are subject-specific qualifications. The majority of children will take 9 GCSEs in the academic year they turn 16. A-levels are also subject specific and most people continuing in school on an academic route will specialise to take three subjects at the age of 18. A range of non-vocational qualifications are available at both stages, yielding the mapping noted above. We compared how well *maternal education* and *highest household education* (i.e., the educational qualification of the most qualified parent in the household) predicted vocabulary at each age (see Appendix B) and, based on findings that highest household education consistently accounted for the most variance in vocabulary at each age, we use a measure of highest parent education in our analyses.

Family Income. Here we used UK OECD weighted income quintiles at child age 3 (an indication of household income 1=lowest, 5=highest, accounting for family size). If data was missing, OECD weighted income quintiles at child age 9 months were used instead.

Wealth. Here we used a measure of total net wealth, taken from the age 11 sweep of the MCS2001 cohort — when cohort members were aged 11, parents reported on their savings and assets, total debts owed, the value of their house and the amount of outstanding mortgage owed on their home for the first time. This measure was derived from 4 variables: amount outstanding on all mortgages, house value, amount of investments and assets, and amount of debts owed. Outstanding mortgages were subtracted from the house value, to give a measure of housing wealth. In cases where families were not homeowners, they were given a housing wealth value of 0. Debts owed were taken from the amount of investments and assets, to give a measure of financial wealth. In cases where families reported having no savings or debts, they were given a financial wealth value of 0. Housing wealth and financial wealth were then summed to give an overall measure of total net wealth. Our measure of wealth was heavily positively skewed, in line with the distribution of wealth in the general population, which is heavily influenced by extreme values of the top 1% (Killewald, 2017). Total net wealth was therefore split into quintiles for our analyses.

Occupational Status. Here we used the highest household occupational status (National Statistics Socioeconomic Classification (NS-SEC) 3 categories: higher managerial; intermediate; and routine, with a fourth category for those who were unemployed) at child age 3 years. If data were missing, occupational status at child age 9 months was used instead.

Relative Neighbourhood Deprivation. Indices of multiple deprivation (IMD) are the government official measure of relative deprivation (Mclennan et al., 2019). Based on an individual's postcode (at the level of the street), these are used to rank small areas or neighbourhoods in England, Scotland, Wales, and Northern Ireland from the least deprived to the most deprived area. The IMD is a broad conceptualisation of deprivation, including a wide variety of living circumstances, rather than just a lack of income for adequate financial resources, which often defines people living in poverty. However, people can be considered deprived if they do not have access to any type of resource, not just income (Mclennan, 2019). Therefore, we used IMD deciles at child age 3 (with 1= most deprived and 10=least deprived) as a measure of relative neighbourhood deprivation.

Measures Of Socioeconomic Position (Cross-Cohort Comparison)

The SEC indicators used in RQ1-RQ3 include the full set of five SEC indicators (parent education, income, wealth, occupational status, and neighbourhood deprivation), enabling us to consider the multi-faceted nature of SEC. However, they are not all directly comparable to the data available in the BCS1970 cohort. Therefore, for RQ4, we used a subset of SEC indicators to ensure comparability, to the best of our ability, across the two cohorts. Harmonisation of these measures can be found in Table 1; data harmonisation is the process of making data from different sources (such as different cohorts) more similar to improve comparability between cohorts (O'Neill, Kaye, & Hardy, 2020).

Parent Education. The highest academic qualification achieved by a parent in the household when the cohort member was aged 5. Where this information is missing, information from previous sweeps was used.

Occupational Status. Highest household occupational status at child age 5. For the BCS1970 cohort, this was ascertained with the Registrar General's classification. For the MCS2001 cohort, the NS-SEC classification system was used. Where this information is missing, information from previous sweeps was used.
Family Income. UK OECD weighted income quintiles at child age 10 (BCS1970) and 11 (MCS2001) were used as an indication of household income 1=lowest, 5=highest, accounting for family size). The BCS1970 first measured family income when cohort members were aged 10, hence we take this information from the age 10 (BCS1970) and age 11 (MCS2001) sweeps for the cross-cohort comparison.

Potential Confounders

We adjusted for gender (male= 0, female=1), ethnicity and whether English was spoken as an additional language (EAL) in the home (1= only English, 2=English and another language, 3=Only another language). Harmonisation of these measures for RQ4 can be found in Table 1.

Data Analysis

All analyses were pre-registered on the Open Science Framework website (https://osf.io/482zw/).

Missing Data Strategy.

Missing data in all analyses was accounted for with multiple imputation using chained equations with the *mice* package in R (van Buuren & Groothuis-Oudshoorn, 2011).

Analysis of MCS2001 Cohort Only. Each dataset was imputed 25 times, as this was greater than the percentage of missing data (10.6%)(White, Royston, & Wood, 2011). There was no missing data for gender or neighbourhood deprivation, and the percentage of missing data was less than 1% for ethnicity and EAL status. 14.71% of vocabulary scores at age 3 were missing, 12.41% of age 5 vocabulary scores were missing, 23.92% of age 11 vocabulary scores were missing, and 36.88% of age 14 vocabulary scores were missing. Full proportions of missing data can be found in Appendix C We conducted a series of sensitivity checks whereby we repeated the analyses on a dataset which had complete cases for vocabulary at ages 3, 5, 11 and 14 Missing data among the components of our wealth variable were also high (30.73% (outstanding mortgage); 27.57% (house valuation); 39.85% (total savings); and 28.99 (total debts owed). We therefore conducted sensitivity analyses where we considered all cohort members with a response to at least one wealth component variable and at least two wealth variables. Overall, these sensitivity checks revealed a similar pattern of results to the main analyses; results are available upon request. Combined sampling and attrition weights were applied to the data to account for the stratified clustered design of MCS2001 cohort data and the oversampling of subgroups, as well as for missing data due to attrition.

Cross-Cohort Comparison. Each dataset was again imputed 25 times, as this was greater than the percentage of missing data in each cohort (6.7% MCS2001 cohort, 21.3% BCS1970 cohort (White, Royston, & Wood, 2011). For the MCS2001 cohort, 6.67% of age 5 vocabulary scores were missing, 18.93% of age 11 vocabulary scores were missing, and 32.74% of age 14 vocabulary scores were missing. For the BCS1970 cohort, 20.12% of age 5 vocabulary scores were missing, 6.89% of age 10 vocabulary scores were missing, and 63.92% of age 16 vocabulary scores were missing (as a result of the teachers strike in 1986). Full proportions of missing data in both cohorts can be found in Appendix C. Again, combined sampling and attrition weights available in MCS2001 data were applied to data from this cohort. The BCS1970 cohort does not have the same sample design as the MCS2001cohort and thus sample weights are not necessary. However, attrition weights to account for non-response between birth and age 5 were created and included in analyses for BCS1970 cohort data (Appendix D for details).

Analyses

Analytic Sample. To address the first two research questions in a contemporary cohort, we analysed the data of 17,070 children in the MCS2001 (all cohort members with a response on at least one of the language tasks at ages 3, 5, 11 or 14). 49.05% of cohort members were female, 85.97% were of White ethnicity and 88.49% did not speak English as an additional language. Demographic differences between the children included in the analytic samples for Research Questions 1-3 and the full MCS cohort are negligible (see Table S2, Appendix E).

For the cross-generation comparison, we analysed the data of 14,851children in the BCS1970, and 16,020 children in the MCS2001 with harmonised measures (cohort members with a response on at least one vocabulary task administered in early childhood, late childhood and/or adolescence; see Table 1 for details of harmonisation). 49.45% of BCS1970 cohort members were female, 93.52% were of White ethnicity and 94.97% did not speak English as an additional language. In the cross-cohort comparison, 48.67% of MCS cohort members were female, 86.03% were of White ethnicity and 88.64% did not speak English as an additional language. Demographic differences between the children included in the analytic samples for Research Question 4 and the full MCS2001 and BCS1970 cohorts were also negligible (see Table S3, Appendix E).

Descriptive Statistics. Descriptive statistics were calculated across the 25 imputed datasets. Analytical samples were compared to the full cohort samples to see if there were any differences in characteristics of those included in the analyses. Mean language scores for each SEC group are reported (see Table 2).

Inequalities in vocabulary at ages 3, 5, 11 and 14: what is the variation captured by each indicator of SEC individually? Language scores at ages 3, 5, 11 and 14 were considered as separate outcome variables. For each age, separate models with each SEC predictor in turn (parent education, income, wealth, occupational status, and neighbourhood deprivation, each in a separate model) were built to assess the unadjusted relationship between each predictor and language at each time point. Potential confounding variables were then added to each of the models.

A drop-one analysis was used to assess the unique contribution of each predictor; a model with all 5 SEC predictors was compared to models with each predictor removed in turn. This was done for each age (3, 5, 11 and 14). Improvements in fit were assessed using model comparisons for imputed data, using the method of Meng and Rubin (Meng & Rubin, 1992). If the five-predictor model was a better fit to the data than the four-predictor model following the removal of an SEC indicator, then the SEC variable that was dropped can be said to account for significant unique variance in language ability at that age. Partial R^2 values for each SEC indicator are reported, indicating the proportion of variance explained by each SEC predictor, above that of the potential confounding variables.

How does a composite measure of overall socioeconomic position perform relative to individual measures and combinations of measures? A latent composite factor of

SEC was created using confirmatory factor analysis (see Appendix F for details). This composite factor was then included as the predictor variable in four separate regression models (each one considering vocabulary at each age), adjusting for the potential confounding variables. Relative AIC values were used to compare the marginal predictive value of each SEC predictor. These were calculated for each imputed dataset for each single-predictor model, the composite model and a model with all indicators included simultaneously (Schomaker & Heumann, 2014), and means and confidence intervals of these values across the imputed datasets are reported. This allowed us to consider whether the composite measure provides an equivalent or better fit to the data, compared to all predictors included simultaneously, and in relation to each individual predictor.

How does the relationship between SEC measures and vocabulary change over developmental time? (Vocabulary at ages 3, 5, 11 and 14). Here we addressed whether or not one's position in the language distribution changes at each age, and how much of this is a function of SEC. The models from RQ1 were used to answer this question. Due to the different measures of language ability available at each age, we were unable to model longitudinal changes in language development. However, because the outcome variable of language ability at each age is standardised to the same scale, the coefficients are directly comparable. We also compared the standardised coefficients from the models in RQ2, which consider our composite factor of SEC, allowing us to establish the best predictor across developmental time.

How has the relationship between SEC measures and vocabulary changed with historical time? (Comparison of two nationally representative cohorts, born 30 years apart). We had 3 separate outcome variables in each cohort (early childhood language ability, late childhood language ability, and adolescent language ability). We built three regression models per outcome, one with occupational status as the predictor variable, one with parent education as the predictor variable, and finally, one with family income as the predictor variable. Because our measures of language ability were standardised within each cohort, we were able to directly compare coefficients between cohorts and establish the rate of inequality in language ability at each age in the two cohorts.

Table 1. ***Cross-cohort harmonisation of variables***

Measure	BCS1970	MCS2001	Harmonised
Age 5 language ability	EPVT. Continuous measure.	Naming vocabulary. Continuous measure.	Total vocabulary score: continuous cohort specific standardised *z* score
Late childhood language ability	Age 10. BAS word similarities	Age 11. BAS II verbal similarities	Total vocabulary score: continuous cohort specific standardised *z* score
Adolescent language ability	Age 16. Vocabulary Test	Age 14. Word activity task,	Total vocabulary score: continuous cohort specific standardised *z* score. Note that a harmonised version of the BCS1970 Vocabulary Test with the same words included in the MCS2001 Word activity task was also created, however this correlated 0.93 with the full BCS1970 measure, so we did not conduct this sensitivity analysis.
Occupational status at birth	Age 5. Registrar General's classification. 5 classes: 1. professional 2. managerial, other professionals 3. non-manual skilled, skilled manual 4. semi-skilled workers 5.unskilled workers 6. Full/part time students or volunteers with no paid employment	Age 5. NS-SEC 5 classes: 1. Higher managerial/admin/professional 2. intermediate 3. small employers/self-employed 4. lower supervisory and technical occupations 5. semi-routine and routine	Composite variable, with a 4th category for unemployment: BCS1970: Professional & Managerial Skilled Semi-skilled and unskilled Unemployed

	Note: students/volunteers were categorised as unemployed as they have no paid employment.	This 5-class version was collapsed into a 3-class version, as shown here: https://www.ons.gov.uk/methodology/classificationsandstandards/otherclassifications/thenationalstatisticssocioeconomicclassificationnssecrebasedonsoc2010#classes-and-collapses	MCS2001: Higher managerial Intermediate Routine Unemployed *Note: The convention used in the MCS2001 was used for the occupational status variables from both cohorts, for ease.*
Parental education: highest educational qualification (highest household level)	No qualifications Vocational qualifications O levels A-levels State registered nurse Certificate of education Degree +	None of these qualifications GCSE grades D-G O level/GCSE grades A-C A/AS/ S Levels Diplomas in higher education First degree Higher degree Other academic qualifications (incl.overseas)	No qualifications/low level qualifications O levels/GCSE grades A*-C A levels/earning a degree – post 16 education university level qualifications
Family Income	Weekly Income Bands (midpoint for each band) (Age 10) Under £35 pw (£17) £35 - £ 49 pw (£42) £50-£99 pw (£74.50) £100 - £149 pw (124.50)	Annual Income Bands (midpoint for each band) (Age 11) < £ 3,000 (£1500) £3,000- £6,999 (£5000)	OECD equivalisation was applied to the midpoint of each income band in each cohort separately, and these equivalized values were converted into quintiles to give OECD equivalised quintiles:

	£150 - £199 pw (174.50)	£ 7,000 - £ 10,499 (£8750)	
	£200 - £ 249 pw (224.50)	£ 10,500 - £ 12,499 (£11500)	Quintile 1 (Most Deprived)
	> £250 pw (£275)	£ 12,500 - £ 13,999 (£13250)	Quintile 2
		£ 14,000 - £ 14,999 (£14500)	Quintile 3
		£ 15,000 - £ 19,499 (£17250)	Quintile 4
		£ 19,500 - £ 23,499 (£21500)	Quintile 5 (Least Deprived)
		£ 23,500 - £ 27,499 (£25500)	
		£ 27,500 - £ 30,499 (£29000)	
		£ 30,500 - £ 34,499 (£32500)	
		£ 34,500 - £ 39,999 (£37250)	
		£ 40,500 - £ 47,999 (£44250)	
		£ 48,000 - £ 53,999 (£51000)	
		£ 54,000 - £ 62,999 (£58500)	
		£ 63,000 - £ 82,999 (£73000)	
		£ 83,000 - £ 114,999 (£99000)	
		£ 115,000 - £ 149,999 (£132500)	
		more than 150,000 (£150000)	
Ethnicity	European UK	White	Categorical measures collapsed into 0=White, 1=Minority
	European Other	Mixed	
	West Indian	Indian	
	Indian-Pakistani	Pakistani and Bangladeshi	
	Other Asian	Black or Black British	
	African	Other Ethnic group (incl. Chinese, Other)	
	Other		

Language spoken at home	English Welsh-Gaelic Hindi-Urdu Greek-Turkish Chinese-Oriental African Language European Language	Yes - English only Yes - English and other language(s) No - other language(s) only	Categorical measures collapsed into 0= Monolingual English 1= Other language

Results

Which SEC Measures Predict Child Vocabulary?

As can be seen in Table 2, for every SEC measure, the mean vocabulary score is greater with each increase in SEC group, with the highest mean vocabulary scores in the highest SEC group.

To assess the unique contribution of each predictor at each age, a model with all five SEC predictors was compared to models with each predictor removed in turn. Improvements in fit were assessed using model comparisons for imputed data, using the method of Meng and Rubin (Meng & Rubin, 1992). This drop-one analysis revealed that caregiver education, income, wealth, and occupational status accounted for significant unique variance in vocabulary at all ages (see Appendix G). Neighbourhood statistics accounted for significant variance in vocabulary at ages 3, 5 and 11.

Figure 1 presents partial R^2 values indicating the proportion of variance explained by each SEC predictor, above that of potential confounding variables (sex, ethnicity, and whether English is spoken as an Additional Language (EAL) in the home). Caregiver education explains the largest proportion of variance in vocabulary at each age (between 6.4% and 8.5% of variance), closely followed by income and occupational status, and at ages 11 and 14, wealth. Relative neighbourhood deprivation consistently contributes the least variance in vocabulary scores, regardless of age.

Reducing individual indicators to a single composite factor may afford us an efficient way of communicating and understanding inequalities in vocabulary but we do not yet know whether such composites explain more variance than certain SEC indicators considered alone, and/or are equivalent to models with each predictor considered separately. Confirmatory factor analysis was therefore used to create a composite variable of SEC (see Appendix F), which was then included as the predictor in an adjusted model predicting language at ages 3, 5, 11 and 14. Regardless of age, compared to each individual measure, the composite factor was a better fit to the data (see Table S4 in Appendix H), and explained 7.4-10.2% of variance in language across ages. However, a model with each SEC measure included simultaneously explained more variance than a model with just the composite measure and control variables (see Table S5 in Appendix H). This indicates that if one needs to identify a single variable for use in analyses, then a composite variable would be a better choice than any of the original individual predictors. In the absence of such a constraint, including a set of multiple predictors would be preferable.

Table 2: ***Means (±SD) and 95% CIs for language scores in each SEC group at each age (MCS2001 cohort)***

	Proportion (%) or Mean(±SD) *[95% CIs]*			
SEC Indicator	*Age 3 Vocabulary*	*Age 5 Vocabulary*	*Age 11 Vocabulary*	*Age 14 Vocabulary*[1]
Parent Education				
Parent education (NVQ1)	45.24(10.28) [44.61;45.87]	49.78(10.51) [49.14;50.43]	54.97(10.14) [54.35;55.6]	6.12(2.38) [5.97;6.27]
Parent education (NVQ2)	47.91(10.63) [47.59;48.23]	52.79(10.29) [52.48;53.1]	56.83(9.9) [56.53;57.12]	6.53(2.35) [6.46;6.6]
Parent education (NVQ3)	49.62(10.64) [49.23;50.01]	54.24(10.14) [53.86;54.61]	58.36(9.35) [58.01;58.7]	6.81(2.43) [6.72;6.9]
Parent education (NVQ4)	52.35(10.74) [52.07;52.63]	57.54(10.18) [57.28;57.81]	60.76(8.97) [60.53;60.99]	7.57(2.65) [7.5;7.64]
Parent education (NVQ5)	53.47(11.47) [52.82;54.11]	59.56(10.48) [58.97;60.14]	63.26(8.66) [62.77;63.74]	8.53(2.9) [8.37;8.69]
Parent education (none of these/overseas)	41.3(11.55) [40.79;41.8]	46.4(11.66) [45.9;46.91]	54.11(10.9) [53.64;54.58]	5.96(2.27) [5.86;6.06]
Income				
Income (Quintile 1)	44.26(11.49) [43.9;44.62]	49.45(11.3) [49.1;49.8]	55.7(10.62) [55.37;56.03]	6.28(2.35) [6.2;6.35]
Income (Quintile 2)	47.31(11.09) [46.99;47.64]	52.19(10.71) [51.88;52.5]	57.05(9.83) [56.76;57.33]	6.67(2.46) [6.6;6.75]
Income (Quintile 3)	51.18(10.65) [50.83;51.54]	55.97(10.18) [55.63;56.31]	59.05(9.35) [58.74;59.36]	7.08(2.54) [7;7.17]
Income (Quintile 4)	52.58(10.38) [52.22;52.94]	57.44(10.06) [57.1;57.79]	60.37(9.21) [60.05;60.69]	7.51(2.69) [7.42;7.61]
Income (Quintile 5)	53.65(10.32) [53.19;54.12]	59.48(9.78) [59.04;59.92]	62.64(8.46) [62.26;63.02]	7.99(2.79) [7.86;8.12]
Wealth				
Wealth (Quintile 1)	46.5(11.05) [46.19;46.82]	51.55(10.68) [51.25;51.86]	56.09(10.18) [55.8;56.38]	6.52(2.44) [6.45;6.59]

Wealth (Quintile 2)	46.71(11.29) [46.23;47.19]	51.49(11.11) [51.02;51.96]	56.56(10.15) [56.13;56.99]	6.48(2.4) [6.38;6.58]
Wealth (Quintile 3)	49.63(11.2) [49.26;50.01]	54.31(10.76) [53.95;54.67]	58.64(9.51) [58.32;58.96]	6.93(2.5) [6.85;7.02]
Wealth (Quintile 4)	50.75(11.18) [50.37;51.12]	55.68(10.75) [55.32;56.04]	59.59(9.58) [59.27;59.91]	7.16(2.57) [7.08;7.25]
Wealth (Quintile 5)	52.54(10.99) [52.17;52.91]	58.09(10.59) [57.74;58.45]	61.49(8.96) [61.19;61.79]	7.78(2.8) [7.69;7.88]
Occupational Status				
Occupational Status (Unemployed)	44.18(11.07) [43.82;44.54]	48.91(10.9) [48.56;49.27]	55.03(10.61) [54.69;55.38]	6.21(2.4) [6.13;6.29]
Occupational Status (Routine)	47.33(11.09) [46.99;47.67]	52.21(10.7) [51.88;52.54]	56.82(9.92) [56.52;57.13]	6.57(2.38) [6.5;6.65]
Occupational Status (Intermediate)	50.12(10.97) [49.74;50.5]	54.67(10.63) [54.3;55.04]	58.7(9.42) [58.38;59.03]	6.88(2.46) [6.8;6.97]
Occupational Status (higher managerial)	52.75(10.64) [52.48;53.01]	58.28(9.96) [58.03;58.53]	61.28(8.87) [61.06;61.5]	7.74(2.71) [7.67;7.8]
Relative Neighbourhood Deprivation				
Relative neighbourhood deprivation (most deprived)	43.7(11.64) [43.28;44.13]	48.69(11.2) [48.27;49.1]	54.91(10.6) [54.52;55.3]	6.27(2.39) [6.18;6.36]
Relative neighbourhood deprivation (10 - < 20%)	45.77(11.82) [45.3;46.25]	50.54(10.97) [50.09;50.98]	57.07(10.08) [56.67;57.48]	6.59(2.43) [6.49;6.69]
Relative neighbourhood deprivation (20 - < 30%)	48.01(11.1) [47.53;48.5]	53.13(10.6) [52.66;53.59]	57.64(9.94) [57.2;58.07]	6.74(2.54) [6.63;6.85]
Relative neighbourhood deprivation (30 - < 40%)	49.07(11.21) [48.54;49.61]	53.77(10.53) [53.27;54.27]	58.38(10.08) [57.9;58.86]	6.88(2.58) [6.76;7]
Relative neighbourhood deprivation (40 - < 50%)	49.56(10.97) [49;50.12]	54.49(10.89) [53.94;55.04]	58.38(9.12) [57.92;58.84]	6.95(2.53) [6.82;7.08]
Relative neighbourhood deprivation (50 - < 60%)	50.5(10.92) [49.93;51.06]	55.55(10.47) [55.01;56.1]	58.89(9.92) [58.37;59.4]	7.04(2.54) [6.91;7.17]

Relative neighbourhood deprivation (60 - < 70%)	51.48(10.58) [50.88;52.08]	56.35(10.37) [55.76;56.94]	60.16(9.96) [59.59;60.72]	7.25(2.7) [7.09;7.4]
Relative neighbourhood deprivation (70 - < 80%)	52.14(10.49) [51.56;52.72]	57.49(10.57) [56.91;58.08]	60.15(9.03) [59.65;60.65]	7.5(2.67) [7.35;7.65]
Relative neighbourhood deprivation (80 - < 90%)	52.19(10.33) [51.64;52.73]	57.55(10.2) [57.01;58.09]	60.16(9.08) [59.68;60.64]	7.48(2.57) [7.34;7.61]
Relative neighbourhood deprivation (least deprived)	53.61(9.94) [53.09;54.13]	58.93(9.55) [58.43;59.43]	61.45(8.68) [61;61.9]	7.75(2.79) [7.6;7.89]

[1]Note: different standardised vocabulary tests were used at different ages, hence the lower mean score at 14 years.

Does the relationship between SEC and child vocabulary change over developmental time from age 3 to 14 years?

Figure 2 shows the relationships between each SEC indicator and vocabulary at each age (coefficients and 95% CIs plotted; see also Table S6, Appendix I). Because vocabulary scores were converted into *z* scores, the coefficients indicate the change in vocabulary in units of standard deviation (SD) associated with different levels of each predictor. A steeper slope indicates greater inequalities. Inequalities in vocabulary size are consistently narrowest at age 3, and widen by age 5. They then persist throughout childhood and into adolescence, regardless of the SEC indicator used. The relation between SEC and age 14 vocabulary displays a discontinuity not seen for the other ages, with the line appearing shallow for the lower SEC groups and steeper between the higher SEC groups. It is nonetheless clear that across childhood, inequalities in vocabulary have not substantially changed in this cohort; gaps in vocabulary size have not narrowed over time.

Given that the SEC measures used in the above analyses were collected when cohort members were aged 3, it is plausible that this pattern of results is due to the proximity of the SEC measures to the developmental stage at which vocabulary was measured. Therefore, we conducted a sensitivity analysis with age 14 SEC indicators predicting age 14 vocabulary. Overall, despite some inequalities appearing to be wider based on age 14 SEC measures, the proximity of the SEC measure to age 14 vocabulary does not affect the main pattern of results (see Appendix J).

Does the relationship between SEC and child vocabulary change with historical time?

The caregivers of children in the MCS2001 cohort are noticeably different to those of the BCS1970 cohort when compared on the basis of the SEC measures available for both cohorts. More parents of the BCS1970 cohort held no or low-level qualifications compared to parents of the MCS2001 cohort (which is to be expected given changes in the age of compulsory schooling; see Table 3). Furthermore, proportionally more parents from the BCS1970 cohort were in intermediate occupations, whereas more parents from the MCS2001 cohort were in either routine or higher managerial occupations (which is expected given that the UK is becoming more of an hourglass economy; see Table 3; Holmes & Mayhew, 2012). For all SEC measures, the mean vocabulary score was greater with each increase in SEC group in both cohorts, with a higher mean score in the highest SEC groups (see Table S9, Appendix K).

As can be seen in Figure 3, vocabulary scores generally increased with SEC regardless of indicator and cohort (also see Table S10, Appendix K). The overall picture is thus one of continuity of social inequality across the generations. Nonetheless, compared to their BCS1970 counterparts, MCS2001 cohort members whose parents had university level qualifications were at a clearer advantage in terms of their language ability in early childhood and adolescence. In contrast, inequalities in vocabulary based on occupational status and income are wider for the BCS1970 cohort at all ages, as indicated by the steeper slopes for this cohort. As can be seen from partial R^2 values (Figure 4), inequalities are substantial in both cohorts. There is no evidence of a decrease in SEC inequalities over the 30-year period and there is even some evidence that inequalities may have widened in early childhood, with SEC indicators explaining more variance in the MCS2001 cohort for this age point. Whereas for the BCS1970 cohort SEC indicators explained most variance in late childhood, for the contemporary MCS2001 cohort, SEC indicators explained most variance in early childhood.

To examine whether our findings were robust to changes in the distribution of education and occupation measures or to the ethnic composition of the UK during the period separating the BCS1970 and MCS2001 cohorts, we conducted two sensitivity checks. First, highest household occupational status and highest household educational attainment were converted to Ridit scores to aid comparability across cohorts (see Appendix L; Donaldson, 1998). Second, we restricted our analyses to those of a White ethnicity only (see Appendix M). Neither analysis resulted in a change in the pattern of results observed.

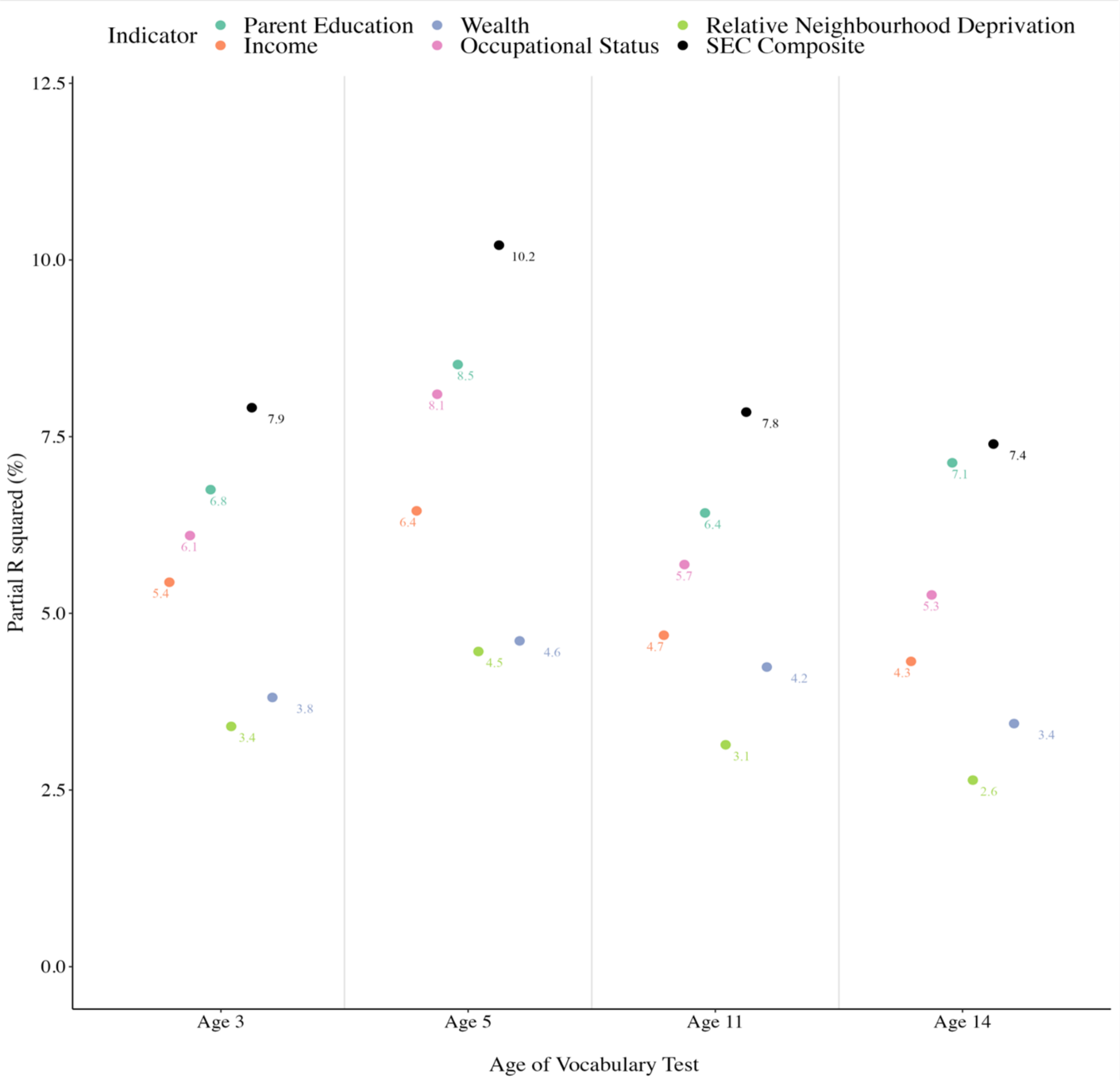


Figure 1. ***Variance explained by SEC indicators in predicting vocabulary in MCS2001 cohort.*** Partial R^2 values for separate models predicting vocabulary at ages 3, 5, 11 and 14, for 5 separate SEC indicators and a composite SEC indicator. Models adjusted for potential confounding variables of sex, ethnicity and English as an additional language (EAL).

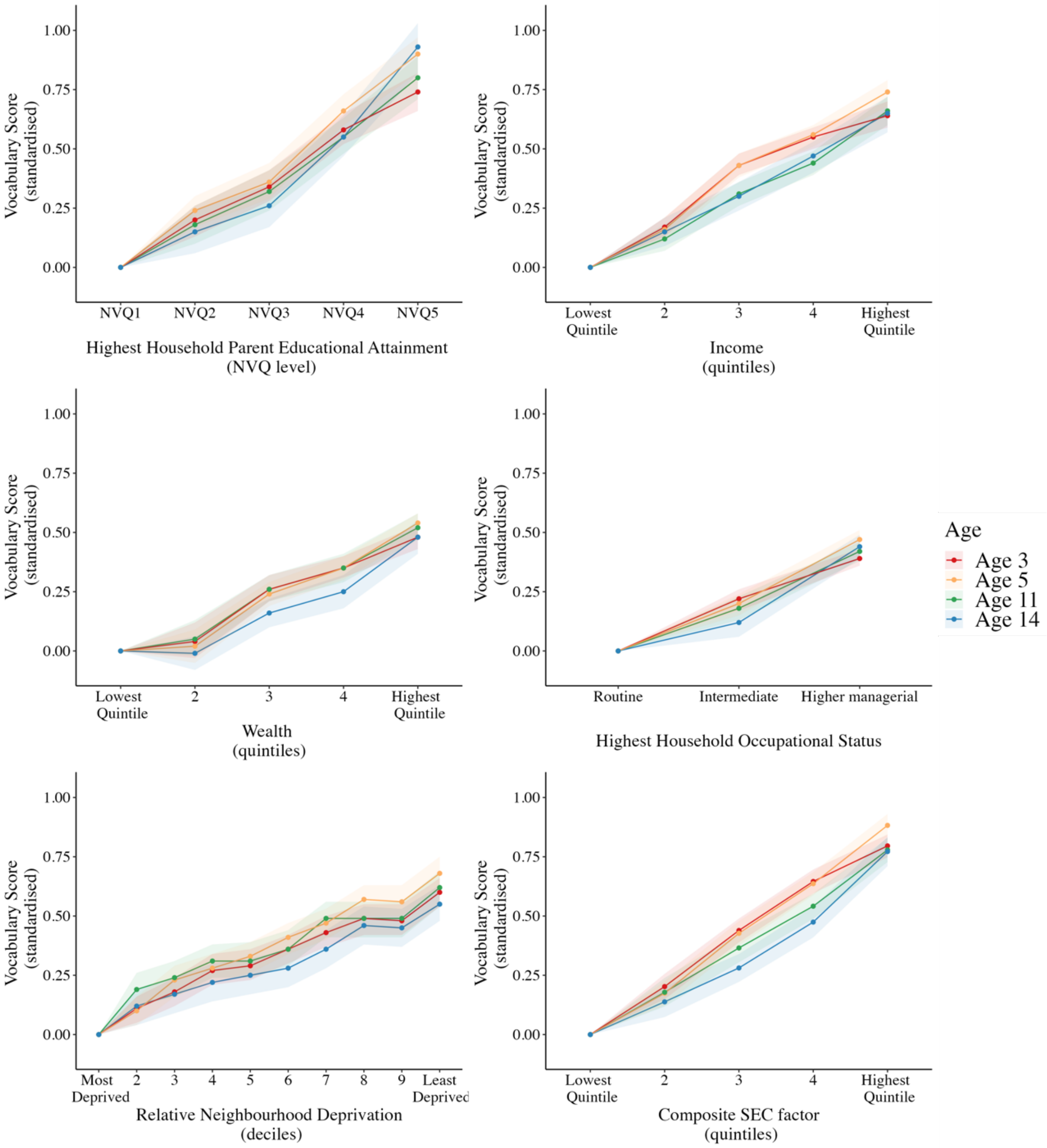
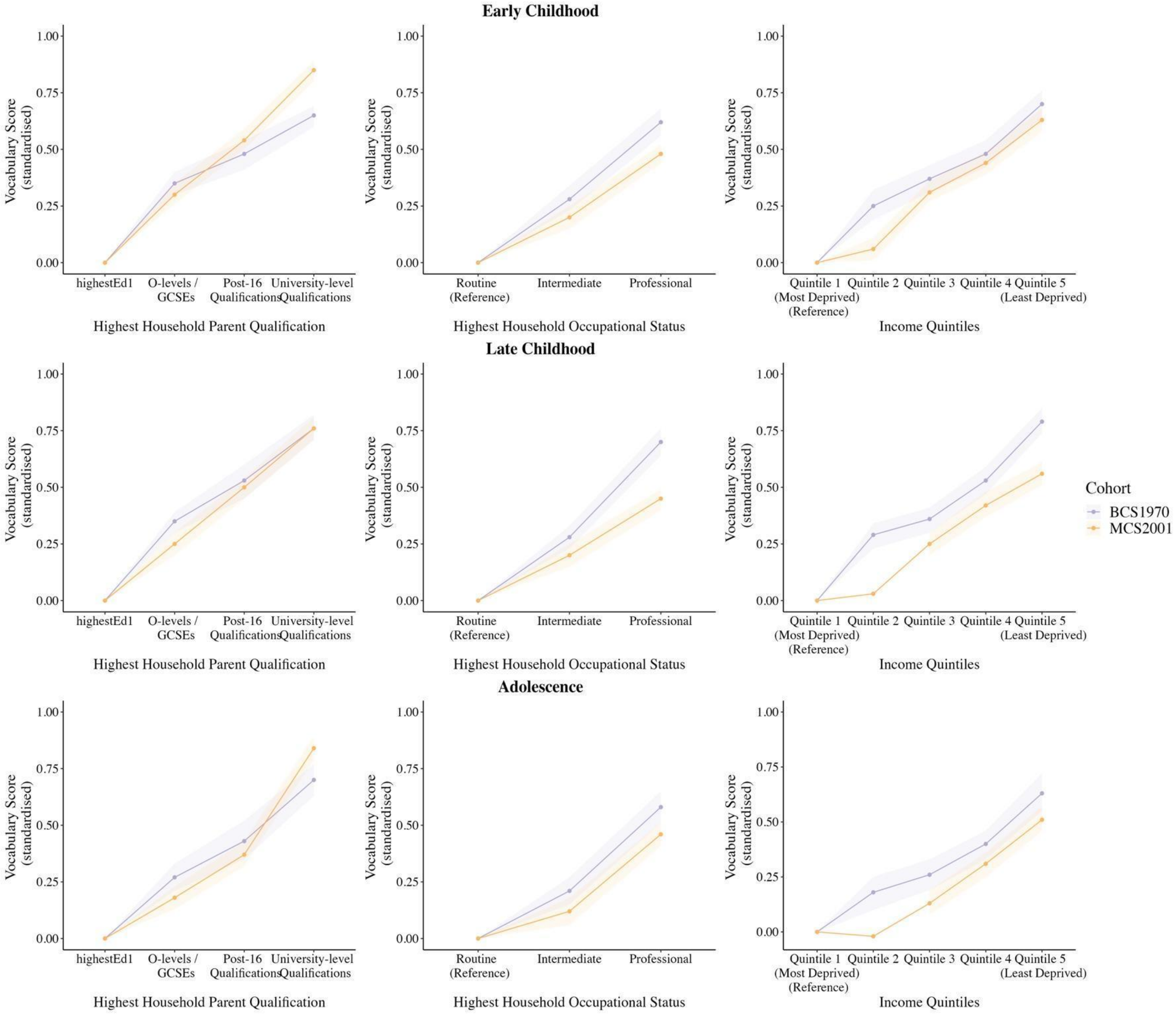


Figure 2: ***Associations between SEC indicators and vocabulary at ages 3, 5, 11 and 14 in the MCS2001 cohort.*** β coefficients and 95% confidence intervals for vocabulary at ages 3, 5, 11 and 14,

plotted as a function of each SEC indicator. Coefficients adjusted for potential confounding variables of sex, ethnicity, and English as an additional language (EAL).

Table 3: ***Descriptive Statistics in MCS2001 and BCS1970 for the cross-cohort comparison***

	Proportion (%) or Mean(±SD) [95% CIs]	
Variable	*BCS1970 (N = 14,851)*	*MCS2001 (N = 16,020)*
Demographics		
Sex (Male)	50.55	51.33
Sex (Female)	49.45	48.67
Ethnicity (White)	93.52	86.03
Ethnicity (Minority)	6.48	13.97
Language Status (English only)	94.97	88.64
Language Status (English as Additional Language)	5.03	11.36
Socioeconomic Circumstances		
Parent Education (no/low level)	54.49	21.14
Parent Education (O-levels/GCSEs grades A*-C)	20.23	32.1
Parent Education (ost-16 quals)	7.66	21.85
Parent Education (university level quals)	17.62	24.92
Income Quintile 1	21.31	19.67
Income Quintile 2	19.81	19.58
Income Quintile 3	20.84	20.44
Income Quintile 4	20.68	20.07
Income Quintile 5	17.36	20.24
Occupational Status (routine)	14.32	22.47
Occupational Status (intermediate)	50.88	18.98
Occupational Status (higher managerial)	33.63	38.76



Occupational Status (unemployed)	1.16	19.78

Descriptive statistics combined across 25 imputed datasets. Descriptive statistics are sample and attrition weighted (MCS2001 cohort) and attrition weighted (BCS1970 cohort)

Figure 3: *Associations between SEC and language ability in the MCS2001 and BCS1970 cohorts in early childhood, late childhood, and adolescence.* Vocabulary in early childhood (top), late childhood (middle) and adolescence (bottom), plotted as a function of highest household parent education (left), highest household occupational status (middle), and income (right) in two cohorts. Data are β coefficients and 95% confidence intervals. Coefficients adjusted for potential confounding variables (sex, ethnicity, English as an additional language and age at time of vocabulary test).

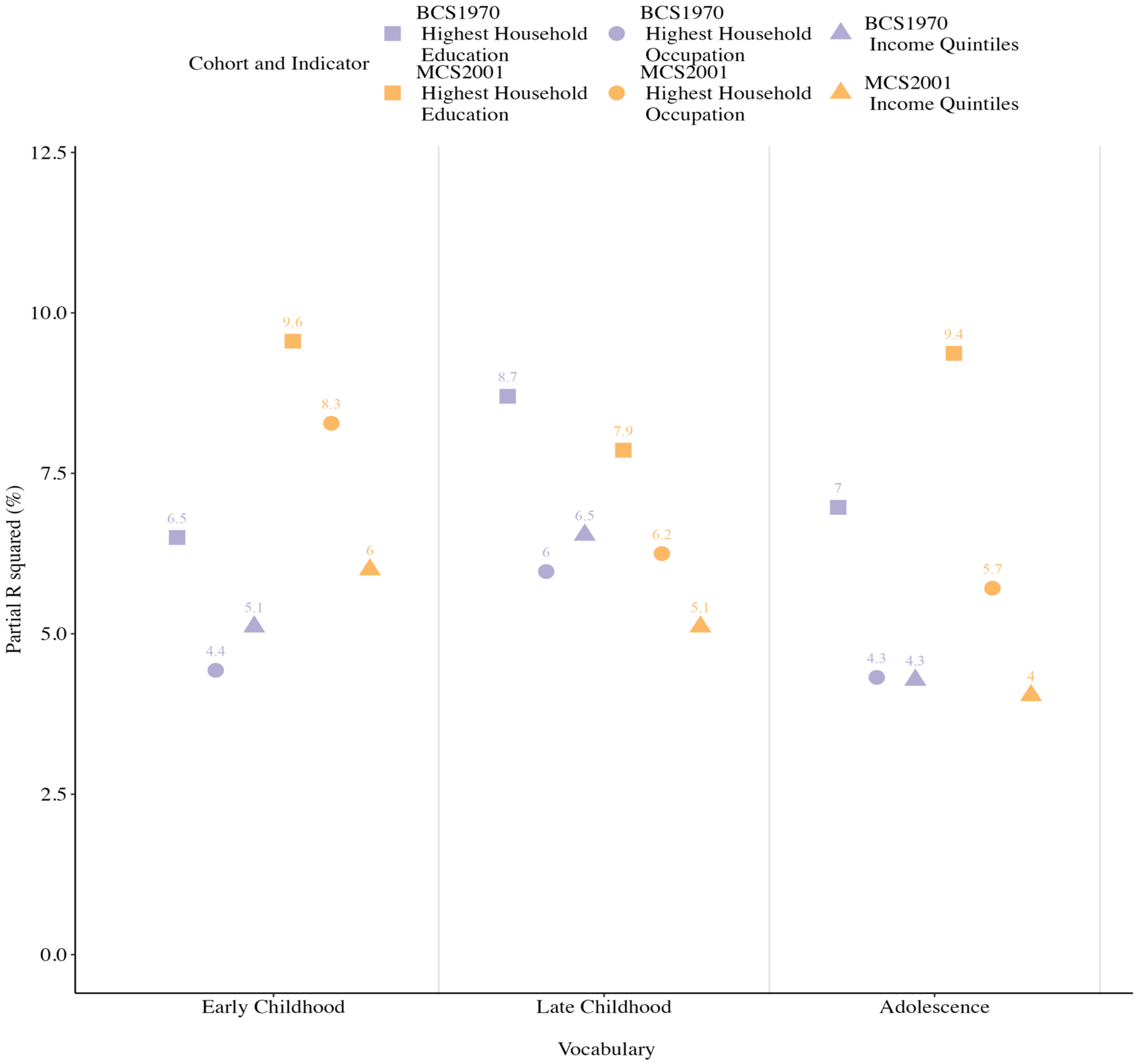


Figure 4: Variance in language explained by SEC indicators in the MCS2001 and BCS1970 cohort. Partial R^2 values (having adjusted for potential confounders of sex, ethnicity, English as additional language and age at time of vocabulary test) for highest household education and highest household

occupational status predicting vocabulary in early childhood, late childhood, and adolescence.

Discussion

Using two UK national birth cohorts, we analysed the relation between multiple SEC indicators and vocabulary across childhood and across generations, and found that (i) all SEC measures predict unique variance at most timepoints and there is generally a monotonic step up in child language for each step up on any given SEC measure. Parent education has the greatest predictive value (closely followed by income, wealth, and occupation) and neighbourhood deprivation the least; (ii) inequalities persist from ages 3 to 14 years, with SEC indicators explaining the most variance in vocabulary scores at 5 years, and an accelerated increase in vocabulary at the higher ends of the socio-economic scale at 14 years; and (iii) across three decades, observed inequalities have generally been stable, but the advantage associated with having parents with higher levels of education has increased.

Overall, the SEC predictor that explains the most variance in child vocabulary across development is caregiver education. However, income, wealth, and occupational status also uniquely predicted large amounts of variance. For all of these indicators, a step up from each level to the next was associated with a substantial step up in vocabulary. This pattern of monotonic increase occurred for all SEC indicators. Thus while most research exploring differences in child language, and in the quality and quantity in child directed speech, tends to compare higher and lower SEC groups (Fernald, Marchman, & Weisleder, 2013; Hart & Risley, 1995; Hirsh-Pasek, 2015; McGillion, Pine, Herbert, & Matthews, 2017; Rowe, 2012; Schwab & Lew-williams, 2016), our findings suggest differences exist across the range of the SEC measures, rather than just between those at the top and bottom of the distribution. Each of these SEC indicators deserve particular attention in the effort to unpick *why* SEC is related to child vocabulary so as to be able to find mechanisms for effective interventions. Caregiver education has been argued to be the most relevant SEC marker for child development (Hoff, 2013; Hoff et al., 2012) because it is associated with caregiver-child interactions and parent knowledge about development (Rowe, 2012, 2018). Parent vocabulary mediates the relation between parent education and child vocabulary ability (Sullivan, Moulton & Fitzsimons, 2021), as well as mediating the relationship between the home learning environment and vocabulary. For example, parents with strong language skills are more likely to participate in reading with their child and may also be more successful in engaging their children in such activities, compared to parents with poor language skills (Sullivan, Ketende & Joshi, 2013). The role of genetics should also be considered here, as language ability is observed to be partly heritable (Chow & Wong, 2021). Prising apart the relative influence of heredity and culture is challenging, given the interplay between the two (Scarr & Mccartney, 1983; Harden, 2021): caregivers and infants with different genetic profiles shape learning environments differently to one another. Unravelling this will require rich datasets that include information regarding interaction dynamics.

While income explained about 6% of unique variance in children's vocabulary, family wealth explained less (about 3-4%), particularly early in childhood. Income is often assumed to affect vocabulary outcomes through the provision of learning resources (Duncan et al., 2017; Washbrook & Waldfogel, 2011). Wealth is usually operationalised as total assets net of outstanding total debt (Killewald et al., 2017), and while one might assume this would act in a similar way to income, it may only become a predictor of outcomes in late adolescence-early adulthood, for example through access to quality secondary education in expensive neighbourhoods (Department for Education, 2017a; Machin, 2011), or financial assistance with higher education (Moulton, Goodman, Nasim, Ploubidis, & Gambaro, 2021; Pfeffer, 2018). Whereas in the UK, most wealth is concentrated in housing (with financial wealth only prominent at the top of the distribution), in the US, financial wealth is more common (Cowell, Karagiannaki, & McKnight, 2019; Office for National Statistics, 2019). International comparisons of the relative predictive value of different SEC indicators across many different countries, alongside qualitative studies, have the potential to shed light on the mechanisms via which these SEC indicators are likely affecting language acquisition and inequalities.

In the contemporary British cohort, inequalities in language ability widen between the ages of 3 and 5. This supports arguments for testing early interventions that seek to avoid inequalities becoming entrenched before children access formal schooling. There is also a clear advantage among 14-year-olds of having parents with a higher level of education. By this age, some adolescents may have vocabulary abilities exceeding those of their parents. Exposure to language occurs in increasingly diverse settings throughout the school years, including via interactions with peers, teachers, and written sources such as books and the internet (Sullivan et al., 2021). As children progress through school, vocabulary development (at least as measured by standardised tests) becomes more dependent on exposure to new words through reading, rather than oral language exposure (Elleman, Oslund, Griffin, & Myers, 2019). It is plausible that these sources of input are influenced by SEC. For example, the availability of and engagement with books and vocabulary-rich online content may be higher among higher SEC children (Maas, Emig & Seelmann, 2013). Children from disadvantaged backgrounds may require more support to acquire particular seams of vocabulary (Sullivan et al., 2021) and yet the type of school attended and the level of support available may differ based on SEC. For example, higher SEC children are more likely to attend private or higher quality schools than their lower SEC counterparts (Dearden, Ryan, & Sibieta, 2011), and parents of children at high performing schools are more likely to invest in educational materials and support, such as books and private tuition (Attanasio, Boneva, & Rauh, 2018). There are also SEC disparities in the amount of homework support adolescents receive at home, not only through tuition, but also in terms of additional hours spent on schoolwork (Jerrim, 2017). While universal education aims to address inequalities in educational opportunity in the UK, when it comes to vocabulary, disparities clearly persist throughout formal schooling. Further support across the lifespan and particularly in the early years and during adolescence is likely necessary to improve educational outcomes and open up employment opportunities (Deloitte, 2016).

Finally, cross-cohort comparisons suggest that inequalities in childhood language are generally similar across generations, despite decades of policy to reduce these inequalities. Nonetheless, there were some differences between the two cohorts: occupational status is becoming less valuable as a predictor, while parental university level qualifications are more clearly associated with better early child and adolescent language in contemporary society. Family income appears to be a slightly stronger predictor of early childhood language in the MCS2001 cohort, but a stronger predictor of late childhood and adolescent language in the BCS1970 cohort. It is possible that these measures are changing in the extent to which they are reliable indicators of the proximal causal factors that explain language learning (such as the caregiving / cultural environment and genetic factors). For example, the move to a more hour-glass shaped economy might mean that occupational status no longer differentiates households' social milieu as well as it once did. Likewise, while many once left the educational system even when they had the academic potential to go on, now with more opportunity to stay in education longer, this measure might better differentiate families along the lines of cognitive ability and educational aspiration. Finally, in the US, financial investments in children increased at the top of the income distribution with the rise of income inequality between 1970 and 2000 (Kornrich & Furstenberg, 2013); it is possible that corresponding increases in parental investments in children have also occurred in the UK, perhaps increasing the importance of income as a predictor of early childhood vocabulary in the MCS2001 cohort compared to the BCS1970 cohort. Alternatively, it might be that the relative importance of the various proximal causal mechanisms themselves is changing with time.

Limitations and strengths.

There are some limitations to our analyses that should be kept in mind when interpreting our results. First, although our cross-cohort comparison has provided insight into socioeconomic inequalities in vocabulary across historical time, and despite extensive efforts to harmonise our variables, historical and societal changes, particularly regarding occupational status and parent education, make it difficult to definitively compare results across the two cohorts, and such differences should be kept in mind when interpreting results. Nonetheless, when we conducted a sensitivity analysis to address this, using Ridit scores as a means of standardising SEC indicators, this revealed a similar pattern of results.

Second, it should be recognised that the vocabulary measures used at each age were necessarily different, meaning we could not assess within-child change in vocabulary scores throughout childhood. However, our focus was on the *extent* of inequalities at each age, and by using a standardised score, we were able to make comparisons that reflect population distributions in these language outcomes.

Third, while vocabulary is the most commonly used measure of language ability in research, especially with regards to inequalities, and is highly correlated with other aspects of language ability (Fenson et al., 1994; Fricke et al., 2017; Hulme, Snowling, West, Lervåg & Melby-Lervåg, 2020), a drawback of the exclusive use of standardised vocabulary tests is that they are potentially inherently biased against children experiencing social disadvantage, because the items included are more likely to occur in higher SEC settings. There has long been debate about how to separate out children's 'inherent potential' for learning language from the language ability they have in virtue of experience. Traditionally, this has been of interest to clinical researchers of speech, language, and communication pathologies, who have been interested in whether a child has a language delay due to relative lack of exposure to accessible linguistic interactions and/or due to an underlying difficulty with learning and/or processing language (e.g., Campbell et al., 1997) - since a clinician's therapeutic response may differ according to aetiology. Calls for the development and adoption of language measures that are sensitive to cultural variation in language experiences continue in the context of debates about how the observed relation between socio-economic disadvantage and language development plays out (Pace et al., 2017).

To demonstrate this limitation, we might outline three (not mutually exclusive) possible scenarios under which a child might perform poorly on a vocabulary test. First, we might consider a child who struggles to learn and process language (and who, in the absence of other known causes, may have a diagnosis of Developmental Language Disorder, with subsequent specialist speech and language therapy and educational support adapted to the specific challenges they face). Second, we might consider a child who has substantial linguistic experience and ability, but whose vocabulary has less overlap with that assessed by a standardised tests than children in the norming sample, due to cultural or socioeconomic differences. Despite having some linguistic strengths, this lack of overlap may still have a functional impact on the child, since this difference may play through in the educational system, making it harder to achieve grades that open doors to future social and economic opportunities. For example, it has long been argued that children from lower SEC backgrounds may have strengths in terms of their discourse skills, compared to middle class children, which are not captured by standardised tests (Heath, 1983; Hoff, 2013; Rogoff et al., 2017). Finally, we might consider a child who has had relatively little accessible linguistic experience, and as a consequence has lower language ability, but this difference is not associated with a skew in the types of language items being assessed on a standardised measure (as was the case for the second child). This is also likely to have a functional impact on the child, but one that cannot be as easily addressed in terms of changing the way standardised tests are normed (or indeed in terms of changes to curriculum, teaching methods, or educational assessment).

The standardised vocabulary measures employed as a proxy for general language ability cannot distinguish between these three types of children (or the more messy reality of several interacting factors contributing to differences in vocabulary assessment outcomes). However, whatever the source of a child's relative difficulty on a standardised test of vocabulary, these tests reflect skills that (rightly or wrongly) are likely important for accessing education (and are known to predict educational outcomes), thus understanding the relation between vocabulary measures and SEC remains important.

Finally, as with any longitudinal analysis, missing data had to be accounted for. Less advantaged individuals tend to be underrepresented in subsequent sweeps of cohort studies (Elliott & Shepherd, 2006; Mostafa & Wiggins, 2014). Further, a teachers strike in 1986 resulted in large amounts of missing data for the adolescent vocabulary measure in the BCS1970 (63.92%). To address this, our analyses were attrition weighted and we used multiple imputations with a rich set of auxiliary indicators to account for missing data, which is considered to be the best approach for appropriately dealing with such missingness (Little & Rubin, 2002). Despite these limitations, the strengths of this research lie in the use of large, nationally representative birth cohort studies with rich information on childhood SEC and researcher-collected, gold standard language measures throughout childhood. Although findings are generalisable to the United Kingdom and hold relatively stable across generations, they may not be generalisable beyond the UK.

Implications

The current findings have several important implications. First parent education level, income, wealth, and occupational status all explain substantial unique variance in child language. This suggests it is well worth testing the causal effects of supporting caregiver education (through lifelong learning) and/or caregiver understanding, motivation, and confidence in supporting child language development (through parenting support). Equally, it is worth testing the effect of reducing poverty – defined as low income relative to a norm (see the Baby's First Years project in the US for a move in this direction: Baby's First Years, 2018). Despite efforts to reduce poverty in the UK, it is ever-present: 22% of the UK population and 30% of children were living in relative poverty (after housing costs) in 2018-19 (Francis-Devine, 2020). Beyond political choices regarding wealth redistribution, educational attainment is claimed to be the key factor causing poor children to become poor adults (DWP, 2014). Since language is the foundation for reading ability and success in education (Public Health England, 2020), and our cross-cohort comparison revealed inequalities in vocabulary are persistently wide across time, targeting these sustained inequalities is assumed to be important in reducing the intergenerational transmission of poverty (Joseph Rowntree Foundation, 2016).

Second, since inequalities in vocabulary widen markedly between the ages of 3 and 5, it remains important to target this age group. A two-pronged approach is likely necessary, whereby family support is provided at the same time as increasing the quality of provision in early years settings (Department for Education, 2017b; Gambaro, Stewart, & Waldfogel, 2015). Regarding the first prong, we need to test ways of creating sustained support for families that leads to lasting cognitive benefits (e.g. testing the BBC's UK-wide Tiny Happy People programme; Tiny Happy People, 2021; Matthews et al., 2023). For the second prong, we need to test ways of improving the consistency and quality of pre-school education to help inequalities becoming entrenched before entry to formal schooling. Quality pre-school provision benefits language development (Becker, 2011; Schmerse, 2020) and is an important factor in supporting later educational attainment, particularly for disadvantaged SEC children (Department for Education, 2015). The introduction in the UK of the National Childcare Strategy in 1998 has made early years education a focus of policy making, particularly with respect to the availability, affordability and quality of education (Department for Education, 2017c). However, quality is inconsistent across different early years settings (Gambaro et al., 2015), such that it is now included in the Ofsted Education Inspection Framework (Ofsted, 2019).

Third, inequalities in vocabulary remain wide throughout childhood and the relative advantage of having parents with higher levels of education accelerates in adolescence as children near the point of being able to leave the education system. However, most language assessments and interventions do not go beyond the early years (Bercow, 2018). Since language skill is important for accessing many employment opportunities, not to mention taking part in wider activities and accessing services, seeking out effective ways to support adolescent language development is important (Bercow, 2018; Spencer et al., 2012).

Fourth, the fact that inequalities generally persist over historical time might be taken to support proposals that interventions to lift the language skills of more disadvantaged children need to be ambitious and scaled up considerably while remaining acceptable to those they are intended to support (Greenwood, Schnitz, Carta, Wallisch, & Irvin, 2020; List, Pernaudet, & Suskind, 2021; Wake et al., 2012). One cause for optimism on this front is that a recent large-scale evaluation has found that the Nuffield Early Language Intervention (NELI) is effective in promoting language skills of children entering formal education in England (West et al., 2021). However, another recent evaluation of a prominent UK intervention, Sure Start, suggests it benefitted child physical health (for example, reduced hospitalisations) - and did so most for those living in disadvantaged areas (Cattan, Conti, Ginja, & Farquharson, 2019) - but the benefits for cognitive outcomes are less clear (Melhuish, Belsky, & Leyland, 2010), perhaps because of a struggle to reach populations who stood to derive the maximum benefit (Law, Parkin, & Lewis, 2012). The current analyses suggest that to have a chance of making a difference, we would need to test a multi-pronged approach, implemented at a meaningful scale, for the long term and in a manner acceptable to children and their families, so as to reap sustained benefits and see the next generation of children reach their potential.

Conclusion

To sum up, the substantial individual differences we observe in child and adolescent language are explained by several SEC indicators each making their own unique contribution, most notably caregiver education, income, wealth, and occupational status. Inequalities are generally stable over developmental and historical time, and are monotonic, with each step up in SEC predicting a step up in language. The current evidence suggests a need to focus on the widening of inequalities as children enter compulsory education and as they prepare to leave it. This supports calls to test the effects of reducing poverty, increasing caregiver lifelong learning, improving early parenting support, improving quality of preschool education, and sustaining educational support throughout adolescence. Tests would need to provide evidence of both causal efficacy and acceptability to those they are intended to help. To succeed on both these fronts, the current evidence suggests we need to be ambitious.

References

Attanasio, O., Boneva, T., & Rauh, C. (2018). Parental Beliefs about Returns to Different Types of Investments in School Children. *Journal of Human Resources,* 0719-10299R1. https://doi.org/10.3368/jhr.58.2.0719-10299r1

Baby's First Years. (2018). https://www.babysfirstyears.com

Bann, D., Johnson, W., Li, L., Kuh, D., & Hardy, R. (2018). Socioeconomic inequalities in childhood and adolescent body-mass index, weight, and height from 1953 to 2015: an analysis of four longitudinal, observational, British birth cohort studies. *The Lancet Public Health*. https://doi.org/10.1016/S2468-2667(18)30045-8

Becker, B. (2011). Social disparities in children's vocabulary in early childhood. Does pre-school education help to close the gap? *British Journal of Sociology, 62*(1), 69–88. https://doi.org/10.1111/j.1468-4446.2010.01345.x

Beddington, J., Cooper, C. L., Goswami, U., Huppert, F. A., Jenkins, R., Jones, H. S., Tom, B. L., Sahakian, B. J., & Thomas, M. (2008). The mental wealth of nations. *Nature, 455,* 1057–1060.

Bennetts, S. K., Love, J., Bennett, C., Burgemeister, F., Westrupp, E. M., Hackworth, N. J., Mensah, F. K., Levickis, P., & Nicholson, J. M. (2022). Do neighbourhoods influence how parents and children interact? Direct observations of parent-child interactions within a large Australian study. *Children and Youth Services Review,* 106704. https://doi.org/10.1016/j.childyouth.2022.106704

Bercow, J. (2018). Bercow : Ten Years On. *Royal College of Speech and Language Therapists, March,* 1–25. https://www.bercow10yearson.com/wp-content/uploads/2018/04/Bercow-Ten-Years-On-Summary-Report-.pdf

Bohlman, E. (2018). *ridittools: Useful Functions for Ridit Analysis*. https://cran.r-project.org/package=ridittools

Bolton, P. (2012). Education: historical statistics. *House of Commons Library, SN/SG/4252*(November), 20. http://www.parliament.uk/briefing-papers/SN00620.pdf

Brimer, M. A., & Dunn, L. M. (1962). *English Picture Vocabulary Test: Educational Evaluation Enterprises. English version of the Peabody Picture Vocabulary Test (PPVT; Dunn, 1959).*

Bronfenbrenner, U. (1979). *The Ecology of Human Development: Experiments by Nature and Design*. Harvard University Press.

Butler, N., Despotidou, S., & Shepherd, P. (1981). *1970 British Cohort Study: Ten Year Follow-Up*.

Campbell, T., Dollaghan, C., Needleman, H., & Janosky, J. (1997). Reducing Bias in Language Assessment. *Journal of Speech, Language, and Hearing Research, 40*(3), 519-525. https://doi.org/doi:10.1044/jslhr.4003.519

Carmines, E. G., & McIver, J. P. (1981). Analysing Models with Unobserved Variables: Analysis of Covariance Structures. In G. W. Bohrnstedt & E. F. Borgatta (Eds.), *Social Measurement: Current Issues* (pp. 65–115). Sage, Beverly Hills.

Cattan, S., Conti, G., Ginja, R., & Farquharson, C. (2019). *The health effects of Sure Start*. Institute for Fiscal Studies. http://www.ifs.org.uk

Chaplin Gray, J., Gatenby, R., & Simmonds, N. (2009). *Millennium Cohort Study Sweep 3 Technical Report.*

Chow, B. W., & Wong, S. W. L. (2021). What does genetic research tell us about the origins of language and literacy development? A reflection on Verhoef et al. (2020). *Journal of Child Psychology and Psychiatry, 62*(6), 739–741. https://doi.org/10.1111/jcpp.13399

Closs, S. J. (1986). *APU vocabulary test (multiple choice format, 1986)*. Kent: Hodder and Stoughton Educational Ltd.

Conger, R. D., & Donnellan, M. B. (2007). An Interactionist Perspective on the Socioeconomic Context of Human Development. *Annual Review of Psychology, 58*(1), 175–199. https://doi.org/10.1146/annurev.psych.58.110405.085551

Connelly, R., & Platt, L. (2014). Cohort profile: UK Millennium Cohort Study (MCS). *International Journal of Epidemiology, 43*(6), 1719–1725. https://doi.org/10.1093/ije/dyu001

Cowell, F., Karagiannaki, E., & McKnight, A. (2019). The changing distribution of wealth in the pre-crisis US and UK: The role of socio-economic factors. *Oxford Economic Papers, 71*(1), 1–24. https://doi.org/10.1093/oep/gpy047

Dearden, L., Ryan, C., & Sibieta, L. (2011). What Determines Private School Choice? A Comparison between the United Kingdom and Australia. *Australian Economic Review, 44*(3), 308–320. https://doi.org/10.1111/j.1467-8462.2011.00650.x

Deloitte. (2016). *Talent for survival: Essential skills for humans working in the machine age.*

Department for Education. (2015). *How pre-school influences children and young people's attainment and developmental outcomes over time (DFE-RB455). June,* 1–50. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/455670/RB455_Effective_pre-school_primary_and_secondary_education_project.pdf.pdf

Department for Education. (2017a). *House prices and schools: do houses close to the best-performing schools cost more?* (Issue March). https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/600623/House_prices_and_schools.pdf

Department for Education. (2017b). *Statutory framework for the early years foundation stage.*

Department for Education. (2017c). *Study of Early Education: Good Practice in Early Education. January.*

Donaldson, G. W. (1998). Ridit scores for analysis and interpretation of ordinal pain data. *European Journal of Pain, 2*(3), 221–227. https://doi.org/10.1016/S1090-3801(98)90018-0

Dorling, D., Rigby, J., Wheeler, B., Ballas, D., Thomas, B., Fahmy, E., Gordon, D., & Lupton, R. (2007). *Poverty, wealth and place in Britain, 1968 to 2005.*

Doyle, O., Harmon, C. P., Heckman, J. J., & Tremblay, R. E. (2009). Investing in early human development: Timing and economic efficiency. *Economics and Human Biology, 7*(1), 1–6. https://doi.org/10.1016/j.ehb.2009.01.002

Duncan, G. J., & Magnuson, K. (2012). Socioeconomic status and cognitive functioning: Moving from correlation to causation. *Wiley Interdisciplinary Reviews: Cognitive Science, 3*(3), 377–386. https://doi.org/10.1002/wcs.1176

Duncan, G. J., Magnuson, K., & Votruba-Drzal, E. (2017). Moving beyond Correlations in Assessing the Consequences of Poverty. *Annual Review of Psychology, 68,* 413–434. https://doi.org/10.1146/annurev-psych-010416-044224

Dunn, L. M., Dunn, L. M., Bulheller, S., & Häcker, H. (1965). *Peabody Picture Vocabulary Test.* Circle Pines, MN: American Guidance Service.

DWP. (2014). *An evidence review of the drivers of child poverty for families in poverty now and for poor children growing up to be poor adults.*

Elleman, A. M., Oslund, E. L., Griffin, N. M., & Myers, K. E. (2019). A review of middle school vocabulary interventions: Five research-based recommendations for practice. *Language, Speech, and Hearing Services in Schools, 50*(4), 477–492. https://doi.org/10.1044/2019_LSHSS-VOIA-18-0145

Elliott, C. D., Murray, D. J., & Pearson, L. S. (1979). *British Ability Scales.* Slough: National Foundation for Educational Research.

Elliott, C. D., Smith, P., & McCulloch, K. (1996). *British Ability Scales Second Edition (BAS II) Early Years.* NFER-Nelson.

Elliott, J., & Shepherd, P. (2006). Cohort profile: 1970 British Birth Cohort (BCS70). *International Journal of Epidemiology, 35*(4), 836–843. https://doi.org/10.1093/ije/dyl174

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., Tomasello, M., Mervis, C. B., & Stiles, J. (1994). Variability in Early Communicative Development. *Monographs of the Society for Research in Child Development, 59*(5).

Fernald, A., Marchman, V. A., & Weisleder, A. (2013). SES differences in language processing skill and vocabulary are evident at 18 months. *Developmental science, 16*(2), 234–248. https://doi.org/10.1111/desc.12019

Fitzsimons, E., Agalioti-Sgompou, V., Calderwood, L., Gilbert, E., Haselden, L., & Johnson, J. (2017). *Millennium Cohort Study Sixth Survey 2015-2016 User Guide (First Edition). February.*

Francis-Devine, B. (2020). *Poverty in the UK: statistics - briefing paper 7096* (Issue 7096).

Fricke, S., Burgoyne, K., Bowyer-Crane, C., Kyriacou, M., Zosimidou, A., Maxwell, L., Lervåg, A., Snowling, M. J., & Hulme, C. (2017). The efficacy of early language intervention in mainstream school settings: a randomized controlled trial. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 58*(10), 1141–1151. https://doi.org/10.1111/jcpp.12737

Gambaro, L., Stewart, K., & Waldfogel, J. (2015). A question of quality: Do children from disadvantaged backgrounds receive lower quality early childhood education and care? *British Educational Research Journal, 41*(4), 553–574. https://doi.org/10.1002/berj.3161

Goodman, A., & Butler, N. (1986). *BCS70 - The 1970 British Cohort Study : The Sixteen-*

year Follow-up.

Greenwood, C. R., Schnitz, A. G., Carta, J. J., Wallisch, A., & Irvin, D. W. (2020). A systematic review of language intervention research with low-income families: A word gap prevention perspective. *Early Childhood Research Quarterly, 50,* 230–245. https://doi.org/10.1016/j.ecresq.2019.04.001

Griffiths, D., Lambert, P. S., & Tranmer, M. (2011). Multilevel modelling of social networks and occupational structure. *Applications of Social Network Analysis (ASNA), University of Zurich,* 4–7.

Grinstein-Weiss, M., Williams Shanks, T. R., & Beverly, S. G. (2014). Family assets and child outcomes: Evidence and directions. *Future of Children, 24*(1), 147–170. https://doi.org/10.1353/foc.2014.0002

Hansen, K. (2014). *Millennium Cohort Study: A Guide to the Datasets (Eighth Edition) First, Second, Third, Fourth and Fifth Surveys. February,* 1–102.

Harden, K. P. (2021). The genetic lottery: why DNA matters for social equality. Princeton University Press.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children.* Paul H Brookes Publishing.

Heath, S. B. (1983). *Ways with words.* Cambridge University Press.

Hirsh-Pasek, K., Adamson, L. B., Bakeman, R., Owen, M. T., Golinkoff, R. M., Pace, A., Yust, P. K. S., & Suma, K. (2015). The Contribution of Early Communication Quality to Low-Income Children's Language Success. *Psychological Science, 26*(7), 1071–1083. https://doi.org/10.1177/0956797615581493

Hoff, E. (2013). Interpreting the Early Language Trajectories of Children from Low SES and Language Minority Homes: Implications for Closing Achievement Gaps. *Developmental Psychology, 49*(1), 4–14. https://doi.org/10.1037/a0027238

Hoff, E., Laursen, B., & Bridges, K. (2012). Measurement and Model Building in Studying the Influence of Socioeconomic Status on Child Development. *The Cambridge Handbook of Environment in Human Development,* 590–606. https://doi.org/10.1017/cbo9781139016827.033

Holmes, C., & Mayhew, K. (2012). The Changing Shape of the UK Job Market and its Implications for the Bottom Half of Earners. In *Resolution Foundation* (Issue March).

https://www.isc.co.uk/media/3179/isc_census_2016_final.pdf

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hulme, C., Snowling, M. J., West, G., Lervåg, A., & Melby-Lervåg, M. (2020). Children's Language Skills Can Be Improved: Lessons From Psychological Science for Educational Policy. *Current Directions in Psychological Science, 29*(4), 372–377. https://doi.org/10.1177/0963721420923684

Institute of Child Health. (1975). *The 1970 birth cohort: 5 year follow up maternal self completion questionnaire.*

Jerrim, J. (2017). Private tuition and out of school study, new international evidence. In *The Sutton Trust* (Issue September).

Joseph Rowntree Foundation. (2016). *We can solve poverty in the UK.*

Killewald, A., Pfeffer, F. T., & Schachner, J. N. (2017). Wealth inequality and accumulation. *Annual Review of Sociology, 43,* 379–404. https://doi.org/10.1146/annurev-soc-060116-053331

Kornrich, S., & Furstenberg, F. (2013). Investing in Children: Changes in Parental Spending on Children, 1972–2007. *Demography, 50*(1), 1–23. https://doi.org/10.1007/s13524-012-0146-4

Law, C., Parkin, C., & Lewis, H. (2012). Policies to tackle inequalities in child health: Why haven't they worked (better)? *Archives of Disease in Childhood, 97*(4), 301–303. https://doi.org/10.1136/archdischild-2011-300827

Law, J., Charlton, J., & Asmussen, K. (2017). Language As a Child Wellbeing Indicator. In *Early intervention foundation.* Early Intervention Foundation. https://www.eif.org.uk/report/language-as-a-child-wellbeing-indicator\

Law, J., Charlton, J., Dockrell, J., Gascoingne, M., McKean, C., & Theakston, A. (2017). Early language development : Needs , provision , and intervention for preschool children from socio- economically disadvantaged backgrounds a report for the Education Endowment Foundation. In *Public Health England.* https://educationendowmentfoundation.org.uk/public/files/Law_et_al_Early_Language_Development_final.pdf

List, J. A., Pernaudet, J., & Suskind, D. L. (2021). Shifting parental beliefs about child development to foster parental investments and improve school readiness outcomes. *Nature Communications, 12*(1), 1–10. https://doi.org/10.1038/s41467-021-25964-y

Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: Wiley.

Maas, J. F., Ehmig, S. C., & Seelmann, C. (2013). Prepare for life: raising awareness for early literacy education; results and implications of the international conference of experts. Prepare for Life: International Conference of Experts, Leipzig, Germany. Leipzig: Stiftung Lesen.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*(2), 130–149. https://doi.org/10.1037/1082-989X.1.2.130

Machin, S. (2011). Houses and schools: Valuation of school quality through the housing market. *Labour Economics, 18*(6), 723–729. https://doi.org/10.1016/j.labeco.2011.05.005

Matthews, D., Bannard, C., Fricke, S., Levickis, P., Salter, G., Solaiman, K., Thornton, E., & Pine J. (2023) *Tiny Happy People Evaluation: Final Results.* BBC Education.

McGillion, M., Pine, J. M., Herbert, J. S., & Matthews, D. (2017). A randomised controlled trial to test the effect of promoting caregiver contingent talk on language development in infants from diverse socioeconomic status backgrounds. *Journal of child psychology and psychiatry, and allied disciplines, 58*(10), 1122–1131. https://doi.org/10.1111/jcpp.12725

Mclennan, D., Noble, S., Noble, M., Plunkett, E., Wright, G., & Gutacker, N. (2019). The English Indices of Deprivation 2019 - technical report. In *Ministry of Housing, Communities and Local Government.* https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/833951/IoD2019_Technical_Report.pdf

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology, 27,* 415–444.

Melhuish, E., Belsky, J., & Leyland, A. (2010). The impact of Sure Start Local Programmes on five year olds and their families. In *DfE.*

Meng, X. L., & Rubin, D. B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika, 79*(1), 103–111. https://doi.org/10.1093/biomet/79.1.103

Mostafa, T., & Wiggins, R. D. (2014). Handling attrition and non-response in the 1970 British Cohort Study. *CLS Working Paper 2014/2, June.*

Moulton, V., Goodman, A., Nasim, B., Ploubidis, G. B., & Gambaro, L. (2021). Parental Wealth and Children's Cognitive Ability, Mental, and Physical Health: Evidence From the UK Millennium Cohort Study. *Child Development, 92*(1), 115–123. https://doi.org/10.1111/cdev.13413

Moulton, V., McElroy, E., Richards, M., Fitzsimons, E., Northstone, K., Conti, G., Ploubidis, G. B. G. B., Sullivan, A., & O'Neill, D. (2020). *A guide to the cognitive measures in five British birth cohort studies.* (Issue August). London, UK: CLOSER. https://www.closer.ac.uk/cognitive-measures-guide

Neuman, S. B., Kaefer, T., & Pinkham, A. M. (2018). Double dose of disadvantage: Language experiences for low-income children in home and school. *Journal of Educational Psychology, 110*(1), 102–118. https://doi.org/10.1037/edu0000201

O'Neill, D., Kaye, N., & Hardy, R. (2020). *Data harmonisation.* CLOSER Learning Hub, London, UK: CLOSER.

Office for National Statistics. (2019). *Total wealth in Great Britain April 2016 to March 2018* (Issue April 2016).

Ofsted. (2019). *Early years inspection handbook for Ofsted registered provion.* www.gov.uk/government/publications/ofsted-

Osborn, A. F., Butler, N. R., & Morris, A. C. (1984). *The social life of Britain's five-year-olds: A report of the Child Health and Education Study.* Routledge & Kegan Paul.

Oxford University Press. (2018). *Why Closing the Word Gap Matters: Oxford Language Report.*

Pace, A., Luo, R., Hirsh-Pasek, K., & Golinkoff, R. M. (2017). Identifying Pathways Between Socioeconomic Status and Language Development. *Annual Review of Linguistics, 3*(1), 285–308. https://doi.org/10.1146/annurev-linguistics-011516-034226

Parsons, S. (2014). *Childhood cognition in the 1970 British Cohort Study* (Issue November). Centre for Longitudinal Studies.

Perkins, S. C., Finegood, E. D., & Swain, J. E. (2013). Poverty and language development: Roles of parenting and stress. *Innovations in Clinical Neuroscience, 10*(4), 10–19.

Pfeffer, F. T. (2018). Growing Wealth Gaps in Education. *Demography, 55*(3), 1033–1068. https://doi.org/10.1007/s13524-018-0666-7

Public Health England. (2020). *Best start in speech , language and communication : Guidance to support local commissioners and service leads.*

Regidor, E. (2004). Measures of health inequalities: Part 2. *Journal of Epidemiology and Community Health, 58*(11), 900–903. https://doi.org/10.1136/jech.2004.023036

Renard, F., Devleesschauwer, B., Speybroeck, N., & Deboosere, P. (2019). Monitoring health inequalities when the socio-economic composition changes: Are the slope and relative indices of inequality appropriate? Results of a simulation study. *BMC Public Health, 19*(1), 1–9. https://doi.org/10.1186/s12889-019-6980-1

Rogoff, B., Coppens, A. D., Alcalá, L., Aceves-Azuara, I., Ruvalcaba, O., López, A., & Dayton, A. (2017). Noticing Learners' Strengths Through Cultural Research. *Perspectives on Psychological Science, 12*(5), 876–888.
https://doi.org/10.1177/1745691617718355

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2). https://doi.org/10.18637/jss.v048.i02

Rowe, M. L. (2012). A longitudinal investigation of the role of quantity and quality of child-directed speech vocabulary development. *Child Development, 83*(5), 1762–1774.
https://doi.org/10.1111/j.1467-8624.2012.01805.x

Rowe, M. L. (2018). Understanding Socioeconomic Differences in Parents' Speech to Children. *Child Development Perspectives, 12*(2), 122–127.
https://doi.org/10.1111/cdep.12271

Rowe, M. L., & Weisleder, A. (2020). Language Development in Context. *Annual Review of Developmental Psychology, 2*(1), 201–223. https://doi.org/10.1146/annurev-devpsych-042220-121816

Rubin, D. B. (1984). *Multiple imputation for nonresponse in surveys.* New York, NY: John Wiley & Sons.

Scarr, S., & Mccartney, K. (1983). How People Make Their Own Environments : A Theory of Genotype → Environment Effects. *Child Development, 54*(2), 424–435.

Schmerse, D. (2020). Preschool Quality Effects on Learning Behavior and Later Achievement in Germany: Moderation by Socioeconomic Status. *Child Development, 91*(6), 2237–2254. https://doi.org/10.1111/cdev.13357

Schomaker, M., & Heumann, C. (2014). Model selection and model averaging after multiple imputation. *Computational Statistics and Data Analysis, 71*, 758–770. https://doi.org/10.1016/j.csda.2013.02.017

Schwab, J. F., & Lew-Williams, C. (2016). Language learning, socioeconomic status, and child-directed speech. *Wiley interdisciplinary reviews. Cognitive science, 7*(4), 264–275. https://doi.org/10.1002/wcs.1393

Spencer, S., Clegg, J., & Stackhouse, J. (2012). Language and disadvantage: A comparison of the language abilities of adolescents from two different socioeconomic areas. *International Journal of Language and Communication Disorders, 47*(3), 274–284. https://doi.org/10.1111/j.1460-6984.2011.00104.x

Sullivan, A. (2007). Cultural capital, cultural knowledge and ability. *Sociological Research Online, 12*(6), 1–14. https://doi.org/10.5153/sro.1596

Sullivan, A., & Brown, M. (2015). Reading for pleasure and progress in vocabulary and mathematics. *British Educational Research Journal, 41*(6), 971–991. https://doi.org/10.1002/berj.3180

Sullivan, A., Ketende, S., & Joshi, H. (2013). Social Class and Inequalities in Early Cognitive Scores. *Sociology, 47*(6), 1187–1206. https://doi.org/10.1177/0038038512461861

Sullivan, A., Moulton, V., & Fitzsimons, E. (2021). The intergenerational transmission of language skill. *The British Journal of Sociology, 72*(2), 207–232. https://doi.org/10.1111/1468-4446.12780

Tiny Happy People. (2021). https://www.bbc.co.uk/tiny-happy-people

Ullman, J. B. (2001). Structural equation modeling. In B. G. Tabachnick & L. S. Fidell (Eds.), *Using Multivariate Statistics* (4th ed., pp. 653–771). Allyn & Bacon: Needham Heights, MA, USA.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

Wake, M., Levickis, P., Tobin, S., Zens, N., Law, J., Gold, L., Ukoumunne, O. C., Goldfeld, S., Le, H. N. D., Skeat, J., & Reilly, S. (2012). Improving outcomes of preschool language delay in the community: Protocol for the Language for Learning randomised controlled trial. *BMC Pediatrics, 12*. https://doi.org/10.1186/1471-2431-12-96

Washbrook, E., & Waldfogel, J. (2011). *On your marks: Measuring the school readiness of children in low-to-middle income families. December*. http://www.resolutionfoundation.org/app/uploads/2014/08/On-your-marks.pdf

West, G., Snowling, M. J., Lervåg, A., Buchanan-Worster, E., Duta, M., Hall, A., McLachlan, H., & Hulme, C. (2021). Early language screening and intervention can be delivered successfully at scale: evidence from a cluster randomized controlled trial. *Journal of Child Psychology and Psychiatry and Allied Disciplines*. https://doi.org/10.1111/jcpp.13415

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine, 30*(4), 377–399. https://doi.org/10.1002/sim.4067

WHO. (2013). Health Inequality Monitoring. *World Health Organization*.

Data, code and materials availability statement

Data: The data used in this paper are held by the UK data service (UKDS; https://ukdataservice.ac.uk). Users of these datasets must agree to an End User License before accessing the data. The datasets can be accessed by creating an account, setting up a project, adding the relevant datasets to the project and agreeing to the End User License for each dataset to be downloaded. Because of restrictions put in place by the UKDS, we are unable to provide a direct link to the data. However, we have provided a note on the datasets used in the GitHub repository for this project (https://github.com/emmathornton/inequalities-vocabulary), which details each dataset required.
Code: All code for this paper can be found on GitHub: https://github.com/emmathornton/inequalities-vocabulary

Ethics statement

Analyses in this paper consist entirely of secondary data analysis of previously collected publicly-available data, therefore no ethical approval was sought.

Authorship and Contributorship Statement

ET, DM, PP and CB conceptualised the study and designed the analysis plan. ET and CB analysed the data, and ET produced data visualisations. ET wrote the first draft of the manuscript, and ET, DM, PP and CB revised the manuscript. All authors approved the final version of the manuscript and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Acknowledgements

We are grateful to the cohort members of the 1970 British Cohort Study, the Millennium Cohort Study and their families. We are also grateful to the Centre for Longitudinal Studies (CLS), for the data collection and management, and to the UK Data Service for making them available. It must be noted that neither CLS nor the UK Data Service have any responsibility or took any part in the analysis or interpretation of these data. We also thank Dr Vanessa Moulton (CLS, UCL) for their extremely helpful comments and advice on the inclusion of income in our cross-cohort comparison analyses, and on the conceptualisation of the wealth variable.

Appendix A

Appendix A contains the details of the vocabulary measures used.

MCS2001 cohort only analyses.

British Ability Scales II (BAS II): Naming vocabulary. Ages 3 & 5 (Elliott, Smith & McCulloch, 1996). This test consists of 36 items of coloured pictures of objects. Cohort members were asked to name each item. Progression through this test depends on performance, and poor performance may result in a different, easier set of items being administered. Cohort members were born over a 1.5 year period (September 2000-January 2002) and assessed over a range of months, so age at the time of testing may differ between cohort members. Therefore, we used t-scores (as published in the data), which are adjusted for item difficulty and age on BAS II age normed data. These were converted to z scores for analyses.

Age 3: At the age of 3, cohort members start the test at item 1. The test ended if the cohort member made five sequential errors. Item 16 was a "decision point" based on performance so far: if the cohort member had got 3 or more items wrong prior to item 16, the test was terminated. If not, the test continued to item 30, the next decision point, where the test was terminated if the cohort member had got 3 or more items wrong. If not, the test continued until item 36 (the end of the test)(Moulton, 2020).

Age 5: The assessment started from picture 12, as this is where children aged 5 start the test. Progression depended on the answers given by the cohort member and the test ended when the child made five sequential errors. However, if at the beginning of the test, the child has made five sequential errors and had less than three correct items, the assessment restarted at an earlier stage with easier items and more teaching items (Chaplin Gray, Gatenby, & Simmonds, 2009; Moulton, 2020). Therefore, MCS cohort members did not complete the same items, as progression through the test depends on their performance and poor performance may result in administration of an easier set of items.

British Ability Scales II (BAS II): Verbal similarities. Age 11. (Elliott, Smith & McCulloch, 1996). This is a measure of verbal reasoning and verbal knowledge. There were 37 items in total (although the first was a practice item and not counted in the final score). Three words were read out to the cohort member, usually by the interviewer, and cohort members had to name the category to which the three words belong (Moulton, 2020; see Figure S1 for examples). Cohort members started the test at age 16, as this is where children aged 11 start the test, and completed up to item 28 (the decision point, based on performance so far). At this point, if there are less than

3 incorrect answers, cohort members continue to item 33. If there are less than 3 correct answers, cohort members are rerouted to an earlier stage, and instead complete items 8-15. If there are five sequential errors and less than three correct items, the cohort members are rerouted to an earlier stage and again complete items 8-15. However, if these items are also too difficult, the test starts again from item 1(Hansen, 2014; Moulton, 2020).

Progression through this test depends on performance, and poor performance may result in a different, easier set of items being administered. Cohort members were born over a 1.5 year period (September 2000-January 2002) and assessed over a range of months, so age at the time of testing may differ between cohort members. Therefore, we used t-scores (as published in the data), which are adjusted for item difficulty and age on BAS II age normed data. These were converted to *z* scores for analyses.

Word Activity Task. Age 14 (Closs, 1986). This test is a measure of vocabulary and also assessed the understanding of meanings of words and word knowledge. Items were a subset of the items from the Applied Psychology Unit (APU) Vocabulary Test(Closs, 1986). Cohort members were given a list of 20 target words, each presented alongside 5 other words. Cohort members had to choose the word which meant the same, or nearly the same as the target word, from the 5 options. Items increased in difficulty throughout the test (Fitzsimons et al., 2017; Moulton, 2020). See Figure S2 for examples of items.

Figure S1. ***Example items from BAS II Verbal Similarities***

	(a)	*(b)*	*(c)*	*(d)* *(cohort member's answer)*
FIRST 3 ITEMS *(from item 16)*				
	Syrup	Toffee	Cake	_____________
	Water	Oil	Blood	_____________
	Jar	Bag	Box	_____________
LAST 3 ITEMS *(items 26-28)*				
	Fraud	Lie	Forgery	_____________
	Hurricane	Draught	Blizzard	_____________
	Siren	Beacon	Horn	_____________

Figure S2. ***Example Items from Word Activity Task***

	(a)	*(b)*	*(c)*	*(d)*	*(e)*
FIRST 5 WORDS					
QUICK	always	best	neat	sick	fast
TIDINGS	steps	reason	jetty	mountains	news
CONCEAL	advise	hide	gather	freeze	conciliate
UNIQUE	several	matchless	simple	ancient	absurd
DUBIOUS	tawny	obstinate	gloomy	muddy	doubtful
LAST 5 WORDS					
OBSOLETE	execrable	secret	innocuous	rigid	redundant
ERUDITE	learned	spasmodic	superfluous	pathetic	spurious
PROSAIC	commonplace	flowery	laudable	poetical	spacious
ASCETIC	artistic	dissolute	austere	antipathetic	charlatan
PUSILLANIMOUS	loud	living	timid	averse	correct

Cross Cohort Comparison

Early Childhood Language Ability

MCS2001 Age 5: BAS II Naming vocabulary (Elliott, Smith & McCulloch, 1997). Details of this test can be found above. The difference here is that in order to aid comparability to BCS1970 data, we here used the ability scores, which are just adjusted for item difficulty and account for the items that the cohort member completed. We adjusted for age in months at the time of the test, instead of using the t-scores available in the data, which are adjusted for age based on BAS II age norms.

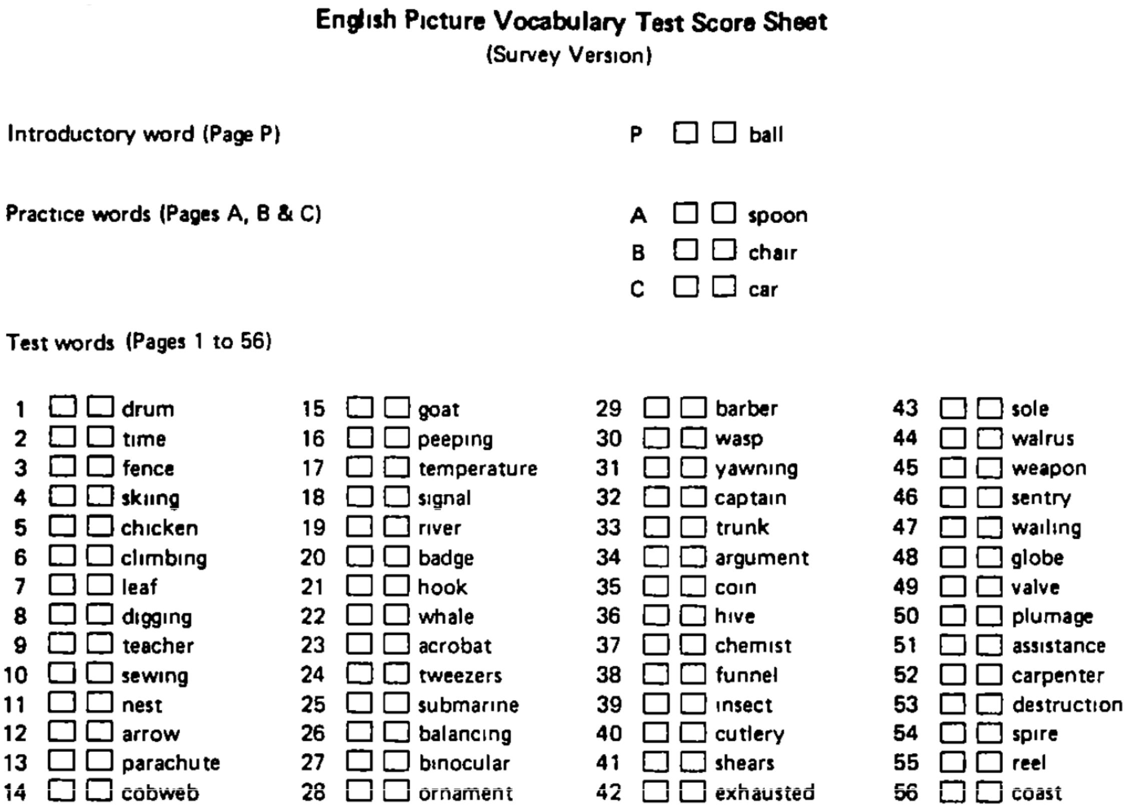
BCS1970 Age 5: English Picture Vocabulary Test (EPVT; Brimer & Dunn, 1962). This test is a UK version of the Peabody Picture Vocabulary Test (Dunn, Dunn, Bullheller & Häcker, 1965). Cohort members were shown 56 sets of four diverse images and a specific word associated with each set of four images. They were asked to select one picture that matched the presented word and were awarded one point for every correct response. The items became increasingly difficult as the test progressed, and the test stopped when the child made five errors in a set of eight items (Parsons, 2014); the 5th wrong answer in a set of 8 sequential items was the ceiling item. Each cohort member's score was the number of correct responses reached before the ceiling item, or (for cohort members who completed the final item of the test without making 5 mistakes in 8 consecutive items), the number of correct responses at the end of the test. Some children did not have a base item, meaning they did not correctly answer 5 of the first 8 items; these children were given a score of 0. Details on the scoring of this vocabulary measure and the SPSS syntax used can be found in appendix 3 of "Childhood Cognition in the 1970 British Cohort Study" (Parsons, 2014).

Scores in the current sample ranged from 0- 56, with higher scores indicating a better language ability. The EPVT has been reported to have a reliability coefficient of .96 (Osborn, Butler, & Morris, 1984). The BCS data does not contain item level responses for the EPVT, only the raw total score, therefore we cannot report the alphas for our analysis sample. However, the items administered in this test were obtained from the British Library to ensure that the procedure and items administered were comparable to other vocabulary tests. Target words can be found in Figure S3 (which are taken from the Age 5 Test Booklet, see here: https://cls.ucl.ac.uk/wp-content/uploads/2017/07/BCS70_age5_test_booklet.pdf). An example of the 4 pictures administered to cohort members could be a drawing of a spider, whale (target), bird and giraffe.

Figure S3. ***English Picture Vocabulary Test Items***
Late Childhood Language Ability

MCS2001 Age 11: BAS II verbal similarities (Elliott, Smith & McCulloch, 1997). Details of this test can be found above. The difference here is that in order to aid comparability to BCS1970 data, we here used the ability scores, which are just adjusted for item difficulty and account for the items that the cohort member completed. We adjusted for age in months at the time of the test, instead of using the t-scores available in the data, which are adjusted for age based on BAS II age norms.

BCS1970 Age 10: BAS word similarities (Elliot, Murray & Pearson, 1979). This test was made up of 21 items, each of which consisted of three words. The teacher read these sets of items out loud and cohort members had to a) name another word that was consistent with the three words in the item and b) state how the words were related. In order to receive a point, cohort members had to correctly answer both parts of the question (Moulton, 2020; Parsons, 2014). If they only answered one part correctly, cohort members received a score of 0 for that item. When the cohort member failed to give the correct group name and an example for four sequential items,



English Picture Vocabulary Test Score Sheet
(Survey Version)

Introductory word (Page P)
P □ □ ball

Practice words (Pages A, B & C)
A □ □ spoon
B □ □ chair
C □ □ car

Test words (Pages 1 to 56)

1 □ □ drum
2 □ □ time
3 □ □ fence
4 □ □ skiing
5 □ □ chicken
6 □ □ climbing
7 □ □ leaf
8 □ □ digging
9 □ □ teacher
10 □ □ sewing
11 □ □ nest
12 □ □ arrow
13 □ □ parachute
14 □ □ cobweb
15 □ □ goat
16 □ □ peeping
17 □ □ temperature
18 □ □ signal
19 □ □ river
20 □ □ badge
21 □ □ hook
22 □ □ whale
23 □ □ acrobat
24 □ □ tweezers
25 □ □ submarine
26 □ □ balancing
27 □ □ binocular
28 □ □ ornament
29 □ □ barber
30 □ □ wasp
31 □ □ yawning
32 □ □ captain
33 □ □ trunk
34 □ □ argument
35 □ □ coin
36 □ □ hive
37 □ □ chemist
38 □ □ funnel
39 □ □ insect
40 □ □ cutlery
41 □ □ shears
42 □ □ exhausted
43 □ □ sole
44 □ □ walrus
45 □ □ weapon
46 □ □ sentry
47 □ □ wailing
48 □ □ globe
49 □ □ valve
50 □ □ plumage
51 □ □ assistance
52 □ □ carpenter
53 □ □ destruction
54 □ □ spire
55 □ □ reel
56 □ □ coast

the test was terminated. Items became progressively harder throughout the test. Details on the scoring of this vocabulary measure and the SPSS syntax used can be found in appendix 3 of "Childhood Cognition in the 1970 British Cohort Study" (Parsons, 2014).

Adolescent language ability

MCS2001 Age 14: Word activity task (Closs, 1986). Details of this test can be found above. We adjusted for age in months at the time of the test, to account for the fact that cohort members were different ages in the MCS2001 and BCS1970 cohorts at the adolescent time point. Items from this test were a subset of the test administered to BCS1970 cohort members when they were aged 16.

BCS1970 Age 16: Vocabulary test (Closs, 1986). This test consisted of 75 items: an item consisted of a target word, presented with a multiple-choice list, from which cohort members had to select a word that meant the same as the target word (Moulton, 2020; Parsons, 2014). Items were progressively harder throughout the test (see Figure S4 for examples). Details on the scoring of this vocabulary measure and the SPSS syntax used can be found in appendix 3 of "Childhood Cognition in the 1970 British Cohort Study"(Parsons, 2014).

Figure S4. ***Example Items from the Vocabulary Test***

	(a)	*(b)*	*(c)*	*(d)*	*(e)*
FIRST 5 WORDS					
BEGIN	ask	start	plain	over	away
AID	help	contrive	assent	manage	hurry
FOREST	grass	wood	sleep	grind	judge
QUICK	always	best	neat	sick	fast
REWARD	notice	golden	prize	stable	Marine
LAST 5 WORDS					
UBIQUITOUS	omnipresent	perdition	adduce	muddy	viscous

PROSAIC	commonplace	flowery	laudable	poetical	spacious
ASCETIC	artistic	dissolute	austere	antipathetic	charlatan
APOSTATE	insufferable	monastic	exegesis	renegade	vicious
PUSILLANIMOUS	loud	living	timid	averse	correct

Figure adapted from Childhood Cognition in the 1970 British Cohort Study, page 29 (Parsons, 2014). Full list of items can be found in the age 16 guide to BCS1970 data(Goodman & Butler, 1986)

Appendix B

Appendix B contains the methods and results for the preliminary analysis of the parent education variable, to determine whether maternal education or highest household education should be used.

Rationale

Previous research often uses maternal education as an indicator of parent education. We consider household SES for all of our other indicators. We therefore conducted a preliminary analysis to determine which measure of parent education predicted the most variance in our outcomes (language at ages 3, 5, 11 and 14). We stated in our preregistration (https://osf.io/482zw/) that we would use the measure of parent education that predicted the most variance in our outcome variables in our main analyses.

Method

Measures

Language ability. At ages 3 and 5, cohort members completed the naming vocabulary subscale of the BAS II. At age 11, cohort members completed the verbal similarities subscale of the BAS II. At age 14, cohort members completed a Word Activity Task. Please refer to the main manuscript and Appendix A for details.

NVQ. When cohort members were aged 3, highest NVQ level was used (both academic and vocational qualifications derived into NVQ levels 1-5, with level 5 equating to higher qualifications). Highest household NVQ was derived from mother and fathers NVQ levels. We considered highest household, mother's and father's NVQ levels as separate predictors.

Analysis plan

Following multiple imputation (see manuscript), we conducted a series of multiple linear regressions: we predicted language at each age with 3 separate regression models, with highest household NVQ level, mother's NVQ level and father's NVQ level as predictors in separate models, in turn. We controlled for gender, ethnicity and whether English was spoken as an additional language in the home.

Results

Table S1 shows results for separate models (one with highest household NVQ level, one with mother's NVQ level and one with father's NVQ level) predicting language at ages 3, 5, 11 and 14. As can be seen from Table S1, highest household NVQ consistently predicted the most variance in language at each age. Therefore, we use a measure of highest household NVQ as an indicator of parent's education in our analyses.

Table S1. ***Partial R2 values for NVQ variables***

	Partial R^2 (%)			
	Age 3	Age 5	Age 11	Age 14
Highest household NVQ	6.81	8.53	6.45	7.16
Mother's NVQ	6.71	8.38	5.83	6.84
Father's NVQ	4.8	6.22	5.22	5.9

Appendix C

Appendix C contains plots showing the extent of missing data in each of our analyses.

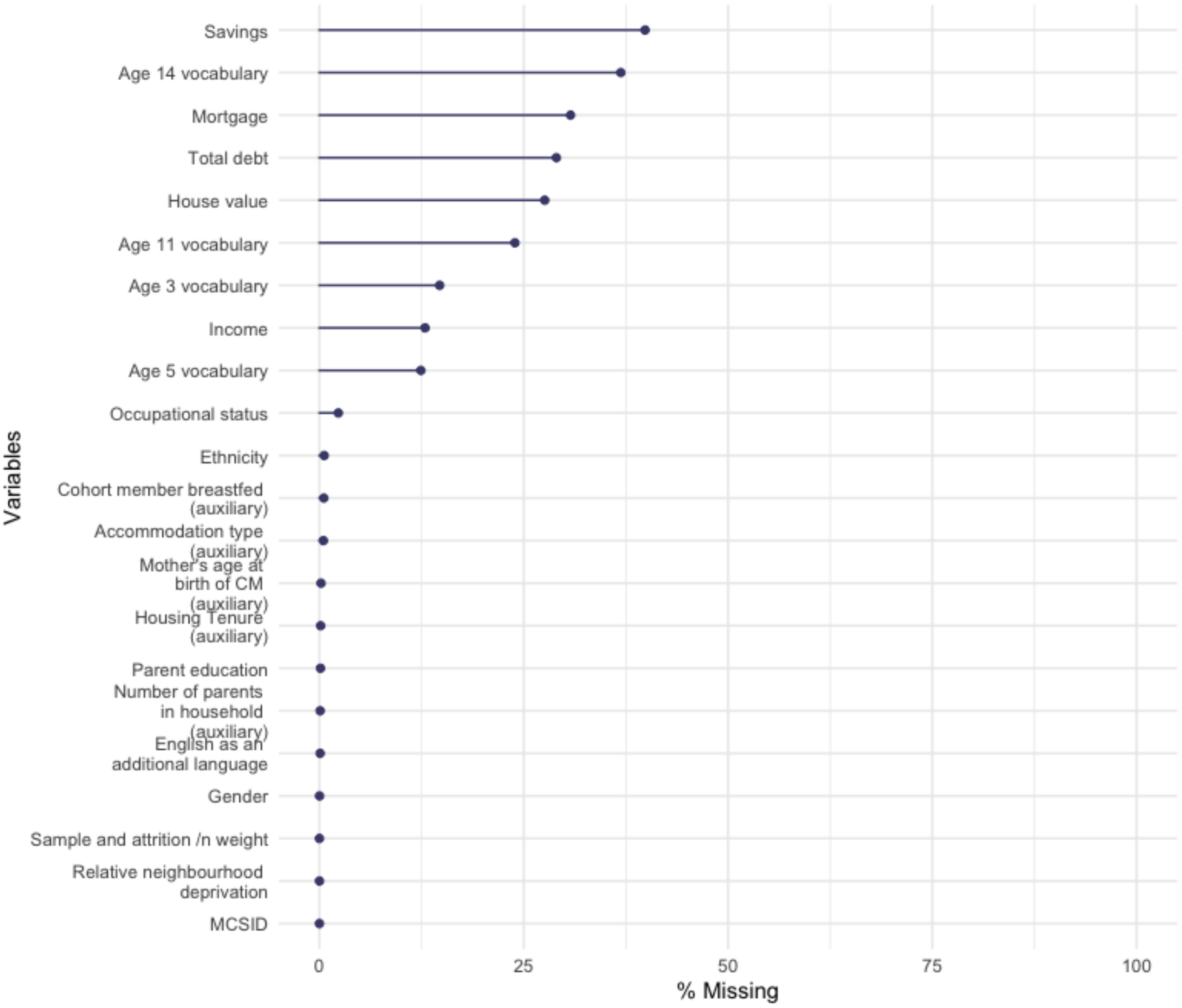


Figure S5. ***Proportion of missing data in the analytical sample used in RQ1-3 (MCS2001, N = 17,070)***

Figure S6. Proportion of missing data in the analytical sample used in the cross cohort comparison (MCS2001, N = 16,020)

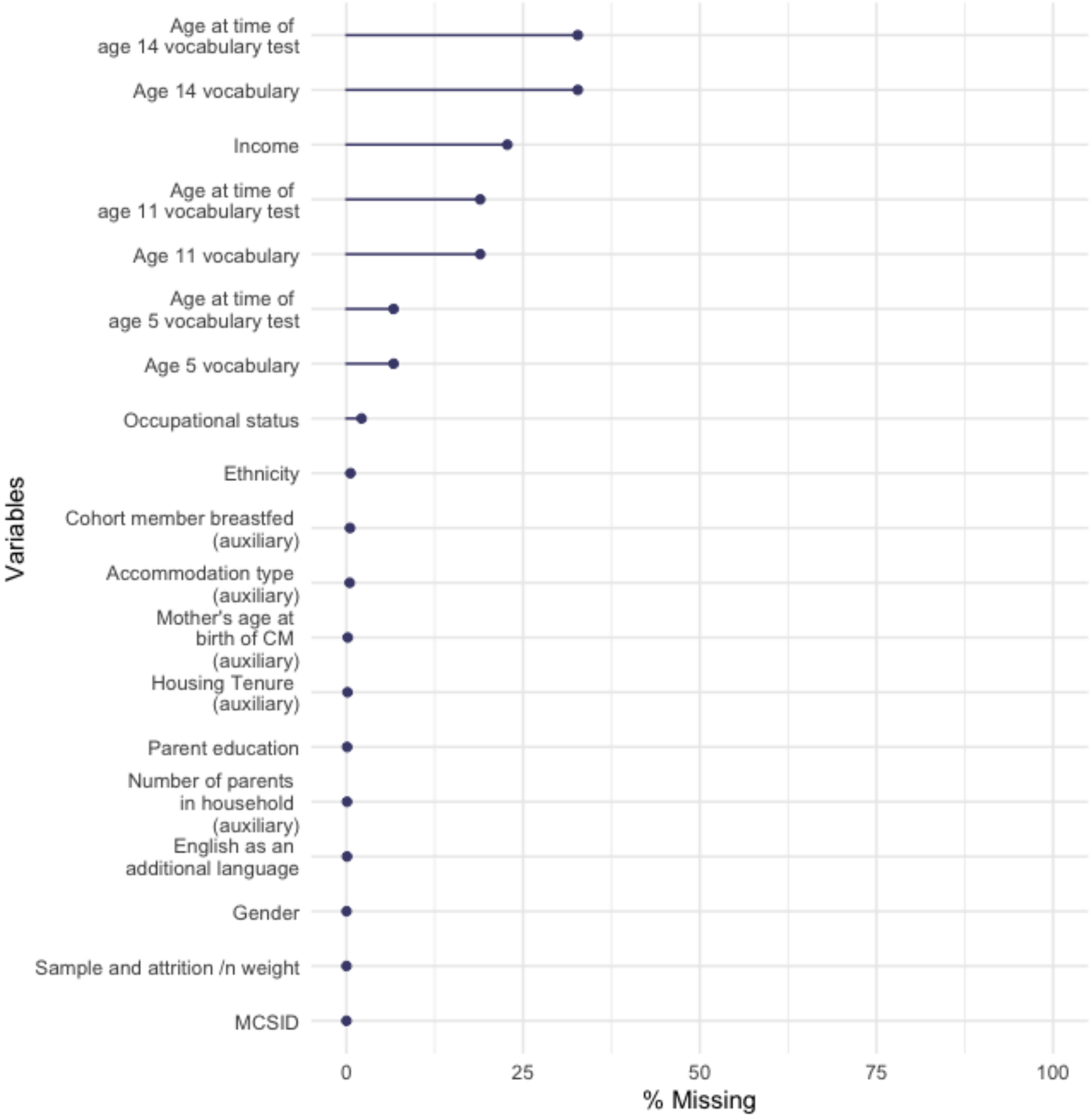
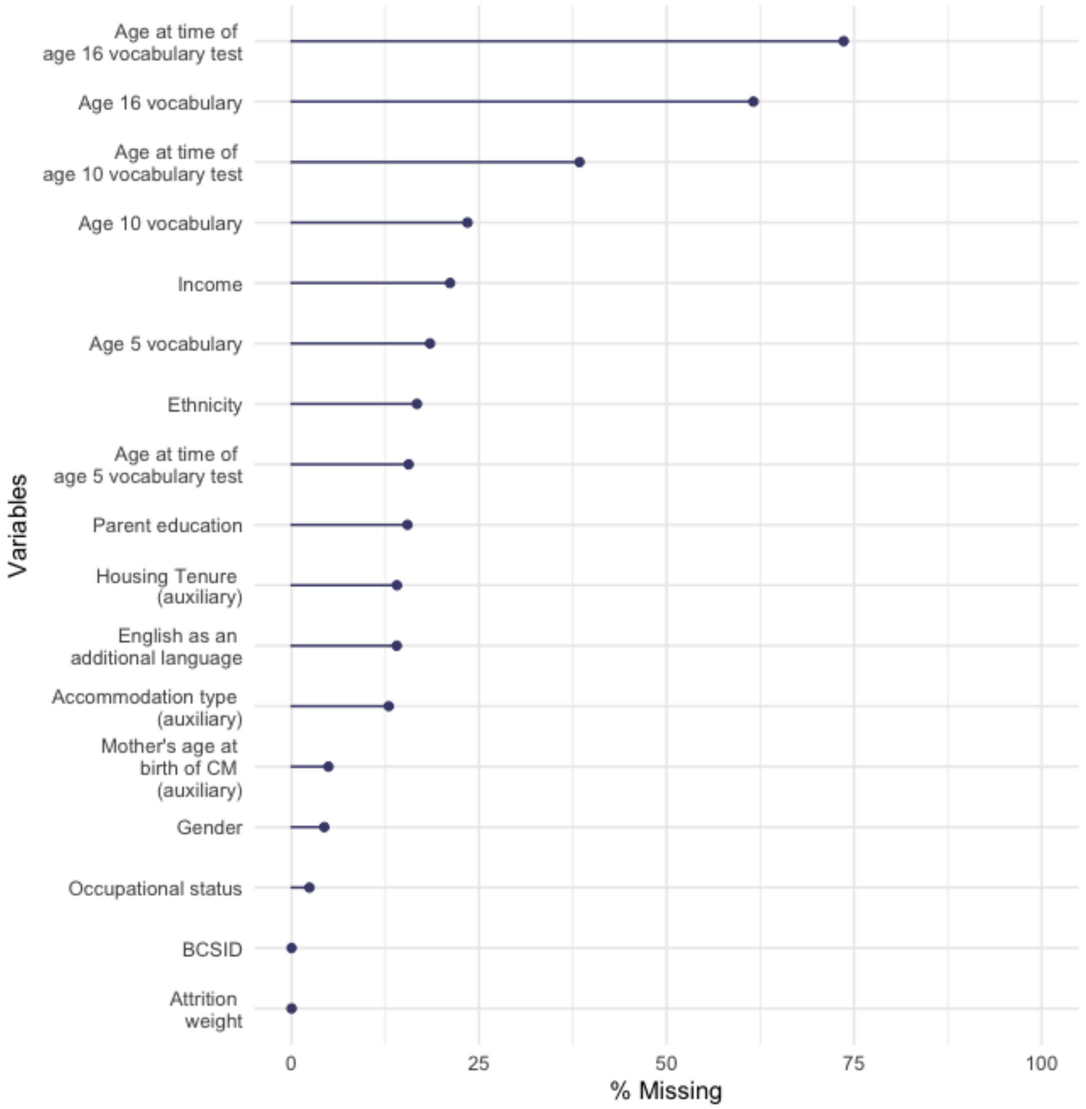


Figure S7. Proportion of missing data in the analytical sample used in the cross cohort comparison (BCS1970, N = 14,851)



Appendix D

Appendix D contains the details for the creation of the attrition weight in the BCS1970.

Procedure

1. Generate a response variable, whereby 1=response and 0=missing
2. Compile predictor variables (detailed below). Where data was missing for these, single imputation was used (random imputation, where impute random values sampled from the non-missing values of the variable)
3. Logistic regression, where response variable is the outcome, and predictor variables are variables deemed to predict missingness (detailed below)
4. Obtain predicted probabilities from the logistic regression
5. The weight variable is the inverse of these probabilities (ie predicted value/1
6. Apply a constant to the weight (weight/1.38)

A weight was created for those who were missing at age 5, those who were missing at age 10 and those who were missing at age 16. This is because although some people may have been missing at age 5, they could have returned by age 10, or they may have participated at age 5, but not age 10. These three weights were then combined into one weight variable, where the weight for age 5 response was used, if this was missing, the age 10 weight was used and if both of these were missing, the age 16 weight was used.

The mean of the final weight variable was 0.9, with a standard deviation of 0.16. The range was 0.83 to 3.82.

Predictor variables

The decision on which variables to include as predictors of response were made following the guides to the BCS datasets (Butler, Despotidou, & Shepherd, 1981; Goodman, 1986; Institute of Child Health, 1975).

Variables predicting response at the age 5 sweep:
From the birth data:

- Whether the cohort member was born to a teenage mother
- Whether the mother had high parity (defined as ≥5 pregnancies of ≥ 20 weeks of gestation)
- Whether the mother was a heavy smoker (defined as ≥15 a day)
- Marital status of mother at birth of cohort member (0=married, 1=single)
- Gender
- Father’s social class

- Mother’s social class

Variables predicting response at the age 10 sweep:
From the birth data:

- Gender
- Parents born outside of Britain
- Age mother and father left full time education
- Whether the cohort member was born to a teenage mother
- Whether the mother was a single mother at birth
- Father unemployed
- Whether the cohort member was a twin
- Mother aged 40+ at child’s birth

From the age 5 data:

- Child’s ethnic group
- Parents with no qualifications
- Separation of mother and cohort member as a baby for 1 month or more
- Father’s social class
- Low birthweight (<5lb)
- Family moved 3 or more times since 1970
- Crowded accommodation (>1 person per room = crowded)
- Whether living in private rented accommodation
- Social rating of the neighbourhood (1=poor, 0=not poor)

Variables predicting response at the age 16 sweep:

- Gender
- Father’s social class
- Region

Appendix E

Appendix E contains the comparisons of the analytical sample with the full cohort samples.

Table S2. ***Full cohort sample vs analytical sample: RQ 1-3, ~2001 born cohort sample only***

	Proportion (%) or Mean(±SD) [95% CIs]	
Variable	*Whole Cohort (N= 19243)*	*Analytical Sample (N=17,070)*
Vocabulary		
Age 3 (Naming Vocabulary Score)	49.9(±11.13) [49.72;50.08]	49.33(±11.38) [49.16;49.5]
Age 5 (Naming Vocabulary Score)	54.67(±10.97) [54.5;54.85]	54.38(±11.05) [54.21;54.54]
Age 11 (Word Similarities Score)	58.8(±9.76) [58.64;58.97]	58.55(±9.88) [58.4;58.7]
Age 14 (Word Activity Task Score)	7.15(±2.63) [7.1;7.2]	7.01(±2.61) [6.97;7.05]
Demographics		
Sex (Male)	50.95	50.95
Sex (Female)	49.05	49.05
Ethnicity (White)	85.98	85.97
Ethnicity (mixed)	3.33	3.33
Ethnicity (Indian)	1.91	1.91
Ethnicity (Pakistani & Bangladeshi)	4.47	4.48
Ethnicity (Black/ Black British)	3.05	3.05

Ethnicity (other incl. Chinese)	1.27	1.26
EAL (English only)	88.5	88.49
EAL (English and another language)	9.01	9.02
EAL (only another language)	2.49	2.49
Socioeconomic Circumstances		
Parent Education (NVQ1)	5.75	5.75
Parent Education (NVQ2)	25.3	10.23
Parent Education (NVQ3)	15.97	25.3
Parent Education (NVQ4)	35.38	15.97
Parent Education (NVQ5)	7.37	35.37
Parent Education (None of these/overseas qualifications)	10.23	7.37
Income Quintile 1	20	21.28
Income Quintile 2	24.46	25
Income Quintile 3	21.53	21.13
Income Quintile 4	20.79	19.97
Income Quintile 5	13.22	12.62
Occupational Status (routine)	22.16	26.73
Occupational Status (intermediate)	18.99	12.26
Occupational Status (higher managerial)	39.04	17.91

Occupational Status (unemployed)	19.81	20.02
Wealth Quintile 1		23.08
Wealth Quintile 2		22.43
Wealth Quintile 3		19.08
Wealth Quintile 4		38.69
Wealth Quintile 5		19.8
Relative Neighbourhood Deprivation (most deprived decile)	12.96	12.96
Relative Neighbourhood Deprivation (10 - <20%)	10.84	10.84
Relative Neighbourhood Deprivation (20 - <30%)	10.32	10.32
Relative Neighbourhood Deprivation (30 - <40%)	9.11	9.11
Relative Neighbourhood Deprivation (40 - <50%)	9.73	9.73
Relative Neighbourhood Deprivation (50 - <60%)	9.73	9.73
Relative Neighbourhood Deprivation (60 - <70%)	8.77	8.77
Relative Neighbourhood Deprivation (70 - <80%)	9.02	9.02
Relative Neighbourhood Deprivation (80 - <90%)	9.55	9.55
Relative Neighbourhood	9.96	9.96

Deprivation (least deprived decile)	

Note: wealth variable compiled after imputation of house value, mortgage, savings and debts and then split to quintiles, therefore cannot calculate proportions before imputation for full sample. Means (±SD) and proportions for analytical sample are pooled across 25 imputed datasets. All descriptives are sample and attrition weighted.

Table S3. ***Full sample vs analytical sample comparisons for RQ4: cross-cohort comparison***

	BCS1970 cohort		MCS2001 cohort	
	Full Cohort Sample (N=17,196)	Analytical Sample (14,851)	Full Cohort Sample (N=19,243)	Analytical Sample (N=16,020)
Language				
Early childhood	35.3(±10.81) [35.11;35.49]	34.74(±11.19) [34.56;34.92]	108.42(±15.89) [108.17;108.68]	107.98(±16.09) [107.73;108.23]
Late childhood	12.06(±2.61) [12.01;12.11]	12.03(±2.64) [11.99;12.07]	120.64(±16.52) [120.36;120.93]	120.18(±16.83) [119.92;120.44]
Adolescence	42.49(±12.65) [42.16;42.82]	41.51(±13.23) [41.3;41.72]	7.13(±2.63) [7.08;7.18]	7(±2.6) [6.96;7.04]
Potential confounders				
Sex (male)	51.42	50.55	51.33	51.33
Sex (female)	48.58	49.45	48.67	48.67
Ethnicity (white)	95.83	93.52	86.03	86.03
Ethnicity (minority)	4.17	6.48	13.97	13.97
English as an additional language (no)	96.86	94.97	88.64	88.64

English as an additional language (yes)	3.14	5.03	11.36	11.36
SES predictors				
Parent education (no/low level)	54.13	54.49	21.13	21.14
Parent education (O-levels/GCSEs grades A*-C)	21.08	20.23	32.1	32.1
Parent education(post-16 quals)	7.76	7.66	21.85	21.85
Parent education (university level quals)	17.03	17.62	24.93	24.92
Income Quintile 1	20.94	21.31	18.09	19.67
Income Quintile 2	19.56	19.81	18.65	19.58
Income Quintile 3	21.28	20.84	20.32	20.44
Income Quintile 4	21.11	20.68	21.04	20.07
Income Quintile 5	17.11	17.36	21.9	20.24
Occupational status (routine)	0.55	14.32	19.78	22.47
Occupational status (intermediate)	15.21	1.16	22.27	19.78
Occupational status (higher managerial)	53.71	50.88	18.94	18.98

Occupational status (unemployed)	30.52	33.63	39.01	38.76

Means (±SD) and proportions for analytical sample are pooled across 25 imputed datasets. All descriptives are sample and attrition weighted (MCS2001 cohort only).

Appendix F

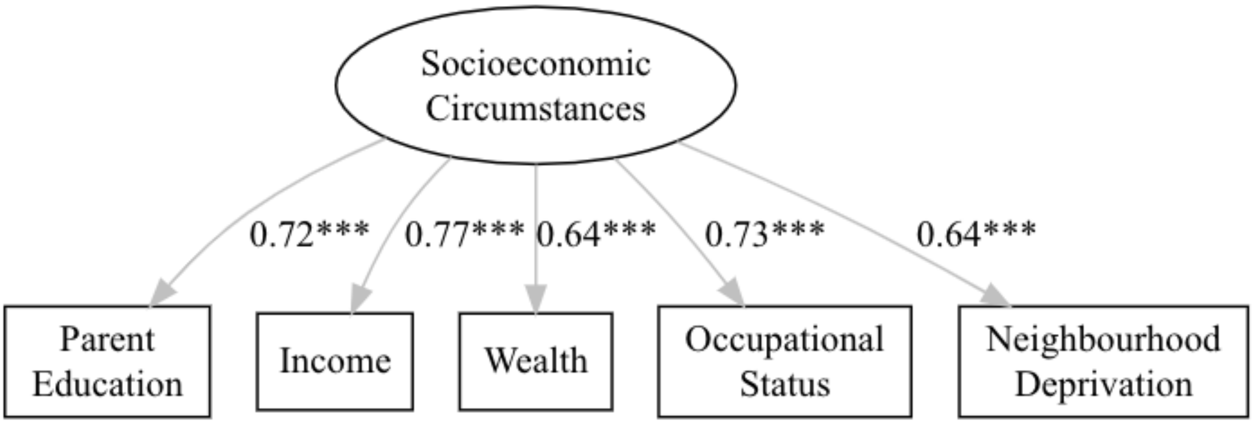
Appendix F contains the details for the confirmatory factor analysis of socioeconomic variables.

Using the lavaan package in R (Rosseel, 2012), a CFA was conducted to create a latent variable of SES. A robust weighted least squares estimator (WLSMV in the lavaan package) was used. This was due to the fact that maximum likelihood estimators are not currently supported for ordered data in the package. A latent variable factor score was then created for each individual imputed dataset, and regression models, where the factor score was the main predictor, were ran for each imputed dataset. The results of the regression models were then pooled. This procedure was conducted on separate regression models, where vocabulary at ages 3, 5, 11 and 14 were the outcome variables.

The latent variable was made up of highest household education, income, wealth, occupational status and relative neighbourhood deprivation. These variables were added to the CFA model in this order. Factor loadings can be found in Figure S1. Model fit was examined with the normed χ^2 (χ^2/df) statistic (Ullman, 2001), Comparative Fit Index (CFI) (Hu & Bentler, 1999), Root Mean Square Error of Approximation (RMSEA)(MacCallum, Browne, & Sugawara, 1996), Standardized Root Mean Square Residual (SRMR)(Hu, 1999) and Tucker Lewis Index (TLI)(Hu, 1999). Normed χ^2 statistics between 1 and 2 suggest a good model fit, and between 2 and 3 suggest an acceptable model fit (Carmines & McIver, 1981). CFI and TLI values of >.9 indicate an acceptable fit and >.95 indicate a good model fit (Hu, 1999). RMSEA values of 0.01 indicate an excellent model fit, 0.05 indicates a good fit and 0.08 indicates an acceptable model fit (MacCallum, 1996). Finally, SRMR values <.08 are indicative of a good fit (Hu, 1999). Robust fit indices are reported.

The model converged on 25 imputed datasets. Estimates were pooled across the 25 imputed datasets, using Rubin's rules (Rubin, 1984). The normed χ^2 statistic indicated a poor model fit (normed χ^2 (χ^2/5)) = 20.39. The remaining fit indices indicated the model was a good fit to the data (RMSEA = 0.034; SRMR = 0.023; CFI = 0.996; TLI= 0.993). Standardised factor loadings indicate that all variables loaded onto the latent construct (see Figure S8).

Figure S8. ***Factor Loadings for CFA***



Appendix G

Appendix G contains the model comparisons for the main analysis.

Model comparisons were conducted to determine whether each SES predictor contributed unique variance in language ability at each age; a model with all indicators included simultaneously was compared to a model with each removed in turn). If the five-predictor model was a better fit to the data than the four-predictor model following the removal of an SES indicator, then the SES variable that was dropped can be said to account for significant variance in language ability at that age.

Age 3. Parent education (Dm(5, 4519.02)= 47.08, *p*<.001), income (Dm(4, 2541.26)= 14.62, *p*<.001), wealth (Dm(4, 415.26) = 5.16, *p* <.001), occupational status (Dm(3, 1421.67)= 17.07, *p*<.001) and relative neighbourhood deprivation (Dm(9, 8022.27)= 2.42, *p*=.009) all accounted for significant variance in language ability at age 3.

Age 5. Parent education (Dm(5, 3051.86)= 51.42, *p*<.001), income (Dm(4, 1458.42)= 10.01, *p*<.001), wealth (Dm(4, 481.19) = 4.39, *p* = .002), occupational status (Dm(3, 2602.84)= 35.08, *p*<.001) and relative neighbourhood deprivation (Dm(9, 7731.82)= 3.63, *p*<.001) all accounted for significant variance in language ability at age 5.

Age 11. Parent education (Dm(5, 1308.32)= 30.99, *p*<.001), income (Dm(4, 861.01)= 7.33, *p*<.001), wealth (Dm(4, 352.28) = 8.57, *p* <.001), occupational status (Dm(3, 473.5)= 11.99, *p*<.001) and relative neighbourhood deprivation (Dm(9, 2628.53)= 2.97, *p* = .002) all accounted for significant variance in language ability at age 11.

Age 14. Parent education Dm(5, 690.38)= 41.28, *p*<.001), income (Dm(4, 494.82)= 4.05, *p* = .003), wealth (Dm(4, 316.61)= 4.08, *p*=.003) , occupational status (Dm(3, 382.10)= 9.02, *p*<.001) all accounted for significant variance in language ability at age 14. Relative neighbourhood deprivation did not account for significant variance (Dm(9, 1702.14)= .83, *p*=.589).

Appendix H

Appendix H contains the AIC values for the main analysis.

Method

AIC values were used to determine whether a model that condenses multiple SES indicators into a single composite factor is a better fit to the data than a model that includes all of these predictors simultaneously. This was to assess how a composite measure of overall socioeconomic position performs relative to individual measures and all indicators included simultaneously. The model with the lowest AIC value is the "best model" and the ΔAIC is the difference between the AIC of each of the remaining models and the AIC of the best model. The ΔAIC values are used to infer the level of support for each remaining model (Fabozzi, Focardi, Rachev & Arshanapalli, 2014). The rules of thumb for interpreting the ΔAIC values are: <2 indicates that the candidate model is almost as good as the best model; values 4-7 indicate considerably less support for the candidate model and >10 indicates that there is no support for this model being the best fit to the data (Fabozzi et al, 2014; Burnham & Anderson, 2002). AIC values are needed here as the models are not nested, therefore the drop one analyses previously used are not applicable. There are also differing numbers of predictors between the composite model and a model containing all predictors simultaneously; AIC values take account of model complexity.

Results

Regardless of age, a model that included each SES indicator as separate predictors was the "best model" (indicated by the smallest AIC values) and the ΔAIC values for the composite model at all ages were greater than 10, lending no support for the composite factor being as good a fit to the data as the 'all predictors separately' model (see Table S4). Thus, it is better to include SES indicators separately when predicting language ability, even when the greater model complexity is taken account of, and there may be a reduction in the predictive accuracy of the model if we reduce the indicators to a composite measure. Compared to individual measures, however, the composite factor was a better fit to the data at all ages (see Table S5). Therefore, compared to individual indicators of SES a composite measure is better than any one measure, but including all as separate indicators provides the best fit to the data.

Table S4. ***AIC and ΔAIC values Individual SES predictors compared to composite factor***

	Mean AIC [95% CIs]							
Indicator	AIC	ΔAIC	AIC	ΔAIC	AIC	ΔAIC	AIC	ΔAIC
Parent Education	47295.78[47256.69;47334.88]	261.77	47721.07[47686.83;47755.32]	373.56	50165.43[50112.69;50218.17]	273.82	51195.47[51120.75;51270.19]	56.95
Income	47576.52[47538.05;47614.99]	542.51	48155.47[48119.85;48191.09]	807.96	50479.66[50423.63;50535.68]	588.05	51705.46[51629.54;51781.37]	566.94
Wealth	47921.55[47885.86;47957.25]	887.54	48533.14[48494.77;48571.52]	1185.63	50560.94[50502.53;50619.35]	669.33	51860.72[51783.21;51938.23]	722.2
Occupational Status	47432.03[47391.75;47472.31]	398.02	47805.53[47771.99;47839.08]	458.02	50294.69[50239.6;50349.78]	403.08	51533.4[51453.66;51613.14]	394.88
Neighbourhood Deprivation	48017.16[47976.83;48057.5]	983.15	48574.37[48539.55;48609.18]	1226.86	50768.7[50713.05;50824.35]	877.09	52014.12[51933.42;52094.82]	875.6
Composite	47034.01[46992.18;47075.84]	AIC*	47347.51[47310.99;47384.04]	AIC*	49891.61[49833.8;49949.43]	AIC*	51138.52[51062.96;51214.07]	AIC*

AIC* = best model; Values are the mean AIC values across 25 imputed datasets; All models adjusted for gender, ethnicity and EAL.

Table S5. ***AIC and ΔAIC values for a model containing all predictors simultaneously vs a composite factor.***

	Mean AIC [95% CIs]							
	Age 3 Language (AIC)	ΔAIC	Age 5 Language (AIC)	ΔAIC	Age 11 Language (AIC)	ΔAIC	Age 14 Language (AIC)	ΔAIC
Composite Factor	47034.01[46992.18;47075.84]	189.46	47347.51[47310.99;47384.04]	179.9	49891.61[49833.8;49949.43]	108.93	51138.52[51062.96;51214.07]	166.25
All predictors (simultaneous)	46844.55[46804.36;46884.75]	AIC*	47167.61[47132.05;47203.17]	AIC*	49782.68[49726.64;49838.71]	AIC*	50972.27[50896.97;51047.57]	AIC*

AIC* = best model
Values are the mean AIC values across 25 imputed datasets
All models adjusted for gender, ethnicity and EAL

Appendix I

Appendix I contains the coefficients for the associations between SEC indicators and vocabulary in the MCS2001 cohort.

Table S6: ***Associations between SEC indicators and vocabulary at ages 3, 5, 11 and 14 in the MCS2001 cohort.***

		β [95% CIs] *p value*			
	Indicator	Age 3	Age 5	Age 11	Age 14
Parent Education	NVQ2	.20[.13;.26] * * * p<.001	.24[.17;.30] * * * p<.001	.18[.10;.26] * * * p<.001	.15[.06;.24] * * * p<.001
	NVQ3	.34[.28;.41] * * * p<.001	.36[.29;.44] * * * p<.001	.32[.24;.41] * * * p<.001	.26[.17;.35] * * * p<.001
	NVQ4	.58[.52;.65] * * * p<.001	.66[.60;.73] * * * p<.001	.55[.48;.63] * * * p<.001	.55[.47;.64] * * * p<.001
	NVQ5	.74[.66;.82] * * * p<.001	.90[.82;.97] * * * p<.001	.80[.71;.89] * * * p<.001	.93[.82;1.03] * * * p<.001
	None of these/overseas qualifications	-.11[-.19;-.04] * * * p<.001	-.09[-.17;-.01] * p= .020	-.06[-.15;.03] p= .190	-.04[-.15;.06] p= .410
Income	Income Quintile 1	REFERENCE	REFERENCE	REFERENCE	REFERENCE
	Income Quintile	.17[.13;.21] * * * p<.001	.16[.11;.21] * * * p<.001	.12[.07;.18] * * * p<.001	.15[.09;.21] * * * p<.001
	Income Quintile 3	.43[.39;.48] * * * p<.001	.43[.38;.48] * * * p<.001	.31[.26;.36] * * * p<.001	.30[.24;.36] * * * p<.001
	Income Quintile 4	.55[.50;.59] * * * p<.001	.56[.51;.60] * * * p<.001	.44[.39;.49] * * * p<.001	.47[.40;.53] * * * p<.001

	Income Quintile 5	.64[.59;.70] * * * p<.001	.74[.68;.79] * * * p<.001	.66[.60;.72] * * * p<.001	.65[.57;.72] * * * p<.001
Wealth	Wealth Quintile 1	REFERENCE	REFERENCE	REFERENCE	REFERENCE
	Wealth Quintile 2	.04[-.03;.12] p= .220	.02[-.05;.09] p= .580	.05[-.02;.13] p= .170	-.01[-.08;.06] p= .730
	Wealth Quintile 3	.26[.21;.32] * * * p<.001	.24[.19;.29] * * * p<.001	.26[.21;.32] * * * p<.001	.16[.10;.22] * * * p<.001
	Wealth Quintile 4	.35[.31;.40] * * * p<.001	.35[.30;.40] * * * p<.001	.35[.29;.41] * * * p<.001	.25[.18;.32] * * * p<.001
	Wealth Quintile 5	.48[.43;.53] * * * p<.001	.54[.49;.58] * * * p<.001	.52[.47;.58] * * * p<.001	.48[.41;.55] * * * p<.001
Occupational Status	Occupational Status (routine)	REFERENCE	REFERENCE	REFERENCE	REFERENCE
	Occupational Status (unemployed)	-.24[-.29;-.19] * * * p<.001	-.26[-.30;-.21] * * * p<.001	-.18[-.24;-.12] * * * p<.001	-.14[-.20;-.08] * * * p<.001
	Occupational Status (intermediate)	.22[.17;.26] * * * p<.001	.20[.15;.24] * * * p<.001	.18[.13;.23] * * * p<.001	.12[.06;.17] * * * p<.001
	Occupational Status (higher managerial)	.39[.36;.43] * * * p<.001	.47[.44;.51] * * * p<.001	.42[.38;.46] * * * p<.001	.44[.39;.49] * * * p<.001
Neighbourhood Deprivation	Relative Neighbourhood Deprivation (most deprived decile)	REFERENCE	REFERENCE	REFERENCE	REFERENCE
	Relative Neighbourhood Deprivation (10 - <20%)	.11[.05;.16] * * * p<.001	.10[.04;.16] * * * p<.001	.19[.12;.26] * * * p<.001	.12[.04;.19] * * * p<.001
	Relative Neighbourhood Deprivation (20 - <30%)	.18[.12;.24] * * * p<.001	.23[.16;.29] * * * p<.001	.24[.16;.31] * * * p<.001	.17[.09;.24] * * * p<.001
	Relative Neighbourhood Deprivation (30 - <40%)	.27[.21;.34] * * * p<.001	.28[.22;.34] * * * p<.001	.31[.24;.38] * * * p<.001	.22[.14;.30] * * * p<.001

	Relative Neighbourhood Deprivation (40 - <50%)	.29[.23;.36] * * * p<.001	.33[.26;.39] * * * p<.001	.31[.24;.39] * * * p<.001	.25[.17;.32] * * * p<.001
	Relative Neighbourhood Deprivation (50 - <60%)	.36[.30;.42] * * * p<.001	.41[.34;.47] * * * p<.001	.36[.29;.44] * * * p<.001	.28[.20;.36] * * * p<.001
	Relative Neighbourhood Deprivation (60 - <70%)	.43[.37;.50] * * * p<.001	.47[.40;.53] * * * p<.001	.49[.41;.56] * * * p<.001	.36[.28;.44] * * * p<.001
	Relative Neighbourhood Deprivation (70 - <80%)	.49[.42;.55] * * * p<.001	.57[.50;.63] * * * p<.001	.49[.41;.56] * * * p<.001	.46[.38;.54] * * * p<.001
	Relative Neighbourhood Deprivation (80 - <90%)	.48[.42;.55] * * * p<.001	.56[.50;.63] * * * p<.001	.49[.41;.56] * * * p<.001	.45[.37;.53] * * * p<.001
	Relative Neighbourhood Deprivation (least deprived decile)	.60[.54;.66] * * * p<.001	.68[.62;.75] * * * p<.001	.62[.54;.69] * * * p<.001	.55[.48;.63] * * * p<.001
Composite	Composite SEC	.28[.26;.29] * * * p<.001	.32[.30;.33] * * * p<.001	.28[.26;.29] * * * p<.001	.28[.26;.30] * * * p<.001

All coefficients taken from models adjusted for gender, ethnicity and English as an additional language (EAL).
*p<.05 ; ** = p<.01; *** p<.001.

Appendix J

Appendix J contains the methods and results for the sensitivity analysis whereby age 14 SEC predictor variables were used to predict age 14 vocabulary.

Rationale

Our main analysis used SES indicators taken at age 3. We found that the strongest associations were with age 5 language ability. We conducted a sensitivity analysis with age 14 SES indicators, to check whether this result was due to the proximity of the SES exposure to the age 5 language outcome. We therefore predicted age 14 language with age 14 SES indicators, using the same methodology as the main analyses (see methods in main manuscript)

Method

Vocabulary measures.

Ages 3 & 5. Cohort members completed the Naming Vocabulary BAS II subscale, as a measure of expressive vocabulary. Cohort members were shown a series of images and were asked to name each item in the image(Moulton, 2020).

Age 11. Cohort members completed the Verbal Similarities BAS II subscale. This is a measure of verbal reasoning and verbal knowledge. Three words were read out to the cohort member, usually by the interviewer, and cohort members had to say how the words were related to each other(Moulton, 2020).

Age 14. Word Activity task. This test was a subset of items from the Applied Psychology Unit (APU) Vocabulary Test(Closs, 1986). Cohort members were given a list of 20 target words, each presented alongside 5 other words. Cohort members had to choose the word which meant the same, or nearly the same as the target word, from the 5 options(Fitzsimons, 2017; Moulton, 2020). Total scores out of 20 were converted into *z* scores for analyses.

Measures of Socioeconomic Circumstance

Five indicators of family SEC were used: parent education, family income, wealth, occupational status and relative neighbourhood deprivation. Operationalisation of these variables is discussed below. These were taken from the age 14 sweep of the MCS2001 cohort.

Parent education. As a measure of parent's education, highest household NVQ level was used (both academic and vocational qualifications derived into NVQ levels 1-5, with level 5 equating to higher qualifications).

Family income. UK OECD weighted income quintiles were used (an indication of household income 1=lowest, 5=highest, accounting for family size).

Wealth. A measure of total net wealth, taken from the age 14 sweep of the MCS2001 cohort. This measure was derived from 4 variables: amount outstanding on all mortgages, house value, amount of investments and assets, and amount of debts owed. Outstanding mortgages were subtracted from the house value, to give a measure of housing wealth. Debts owed were taken from the amount of investments and assets, to give a measure of financial wealth. Housing wealth and financial wealth were then summed to give an overall measure of total net wealth.

Occupational status. Highest household occupational status (National Statistics Socioeconomic Classification (NS-SEC) 3 categories: higher managerial; intermediate; routine, with a fourth category for those who were unemployed) at 14 years.

Relative neighbourhood deprivation. Indices of multiple deprivation (IMD) are the government official measure of relative deprivation (Mclennan, 2019). We used IMD deciles at age 14 (with 1= most deprived and 10=least deprived) as a measure of relative neighbourhood deprivation.

Analyses.

Language scores at age 14 were considered as the outcome variable. Separate models were conducted for SEC measure when the cohort members were aged 3, and when they were aged 14. Drop-one analyses were used to assess the unique contribution of each SEC predictor; a model with all 5 SEC predictors was compared to models with each predictor removed in turn (see main manuscript). A composite factor was included as the predictor variable, adjusting for the potential confounding variables. Results for models considering age 3 SEC predictors of age 14 language ability were compared to that of models considering age 14 SEC predictors of age 14 language ability.

Results

Partial R^2 values for age 3 SEC indicators predicting age 14 vocabulary, compared to age 14 SEC indicators predicting age 14 vocabulary, can be found in Table S7 and Figure S9. With the exception of parent education and occupational status, individual indicators measured at age 14 contributed more variance to age 14 vocabulary. Regression coefficients can be found in Table S8 and are plotted in Figures S10 and S11. Figure S10 displays the regression coefficients for age 3 SEC indicators predicting age 14 vocabulary, compared to age 14 SEC indicators predicting age 14 vocabulary, whilst Figure S11 shows the age 14 SEC coefficients plotted against the main analysis results for all ages. As can be seen from Figure S10, the slopes are similar in steepness, regardless of which age SEC indicators were measured, although the age 14 SEC measures indicate wider inequalities than the age 3 measures. However, when compared to vocabulary at other ages, the main pattern of results remains (see Figure S11): inequalities are widest at the age of 5 and remain persistently wide throughout childhood and into adolescence. Proximity of the SEC measure to age 14 vocabulary does not appear to affect the main pattern of results.

Model comparisons were conducted to determine whether each SES predictor contributed unique variance in vocabulary; a model with all indicators included simultaneously was compared to a model with each removed in turn. All age 14 SES indicators predict unique variance in age 14 vocabulary: Compared to a model without parent education, a model with all SEC predictors was a significantly better fit (Dm(5, 9215)= 26.86, *p*<.001). Compared to a model without income, a model with all SEC predictors was a significantly better fit (Dm(4, 2560.91)= 5.02, *p*<.001). Compared to a model without wealth, a model with all SEC predictors was a significantly better fit to the data (Dm(4, 2188.22)= 11.12, *p* <.001). Compared to a model without occupational status, a model with all SEC predictors was a significantly better fit to the data (Dm(3,8985.4)= 7.98, *p*<.001). Finally, compared to a model without relative neighbourhood deprivation, a model with all SEC predictors was a significantly better fit to the data (Dm(9, 10706.37)= 5.11, *p*<.001).

These findings are in line with that of the main analysis, with the exception of relative neighbourhood deprivation. When measured at the age of 3, relative neighbourhood deprivation did not contribute unique variance in age 14 vocabulary. This perhaps indicates that the proximity of neighbourhood deprivation is important regarding age 14 vocabulary.

Table S7. ***Model R^2 for age 3 SEC predictors and age 14 SEC predictors predicting age 14 language***

Indicator	Age 3 SEC measures	Age 14 SEC measures
Parent Education	7.1	5.5
Income	4.3	4.4
Wealth	3.4	4.6
Occupation	5.3	3.1
Relative Neighbourhood Deprivation	2.6	3.9
SEC composite	7.4	7.7
All predictors simultaneously	8.54	8.74

R^2 of models adjusted for gender, ethnicity and English as an additional language.

Figure S9. ***Partial R² Values for SEC indicators (Ages 3 & 14) for predicting Age 14 Vocabulary***

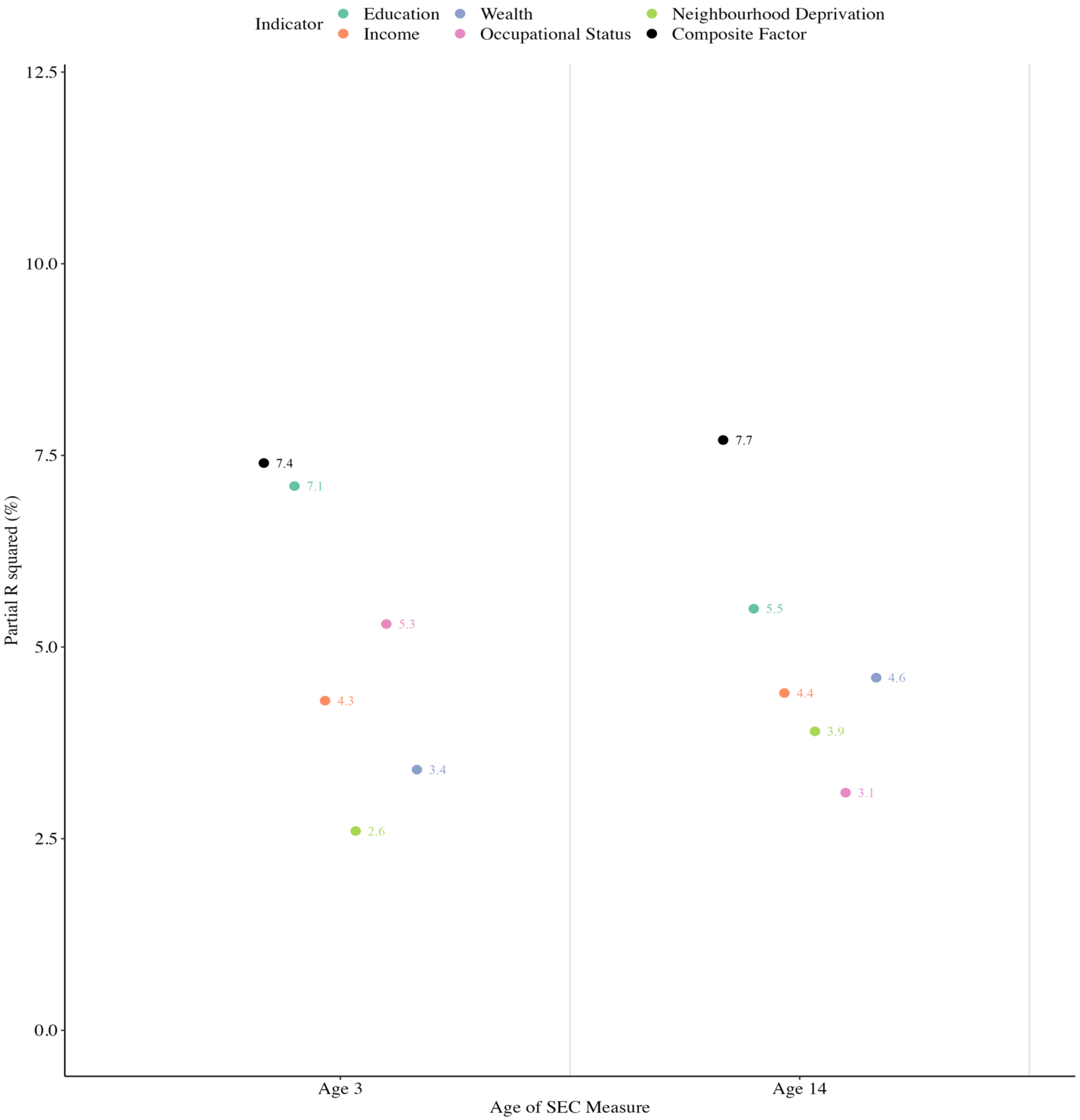


Table S8: Age 14 sensitivity check with Age 14 SES measures: [95% CIs]

		β [95% CIs] *p value*	
	Indicator	Age 14 Vocabulary	
Parent Education	NVQ1	REFERENCE	
	None of these/ oversees qualifications	-.04[-.14;.05] p= .360	
	NVQ2	.18[.10;.27] * * * p<.001	
	NVQ3	.29[.20;.38] * * * p<.001	
	NVQ4	.51[.43;.59] * * * p<.001	
	NVQ5	.68[.59;.77] * * * p<.001	
Income	Income Quintile 1		REFERENCE
	Income Quintile		.15[.09;.21] * * * p<.001
	Income Quintile 3		.24[.18;.30] * * * p<.001
	Income Quintile 4		.42[.36;.48] * * * p<.001

	Income Quintile 5	.63[.57;.70] * * * p<.001			
Wealth	Wealth Quintile 1		REFERENCE		
	Wealth Quintile 2		.15[.09;.21] * * * p<.001		
	Wealth Quintile 3		.26[.19;.32] * * * p<.001		
	Wealth Quintile 4		.39[.33;.45] * * * p<.001		
	Wealth Quintile 5		.60[.54;.66] * * * p<.001		
Occupational Status	Routine			REFERENCE	
	Unemployed			-.07[-.12;-.02] * * * p<.001	
	Intermediate			.18[.13;.24] * * * p<.001	
	Higher managerial			.41[.36;.47] * * * p<.001	
Relative neighbour-	Most deprived decile				REFERENCE
	10 - <20%				.10[.02;.17] * * p= .010

	20 - <30%	.14[.06;.22] * * * p<.001	
	30 - <40%	.23[.16;.31] * * * p<.001	
	40 - <50%	.29[.21;.36] * * * p<.001	
	50 - <60%	.36[.28;.44] * * * p<.001	
	60 - <70%	.45[.37;.53] * * * p<.001	
	70 - <80%	.43[.35;.51] * * * p<.001	
	80 - <90%	.50[.42;.58] * * * p<.001	
	Least deprived decile	.66[.58;.74] * * * p<.001	
Composite	Composite SEC		.28[.26;.30] * * * p<.001

Figure S10. ***Relationships between SEC indicators (Ages 3 & 14) and Vocabulary (Age 14)***

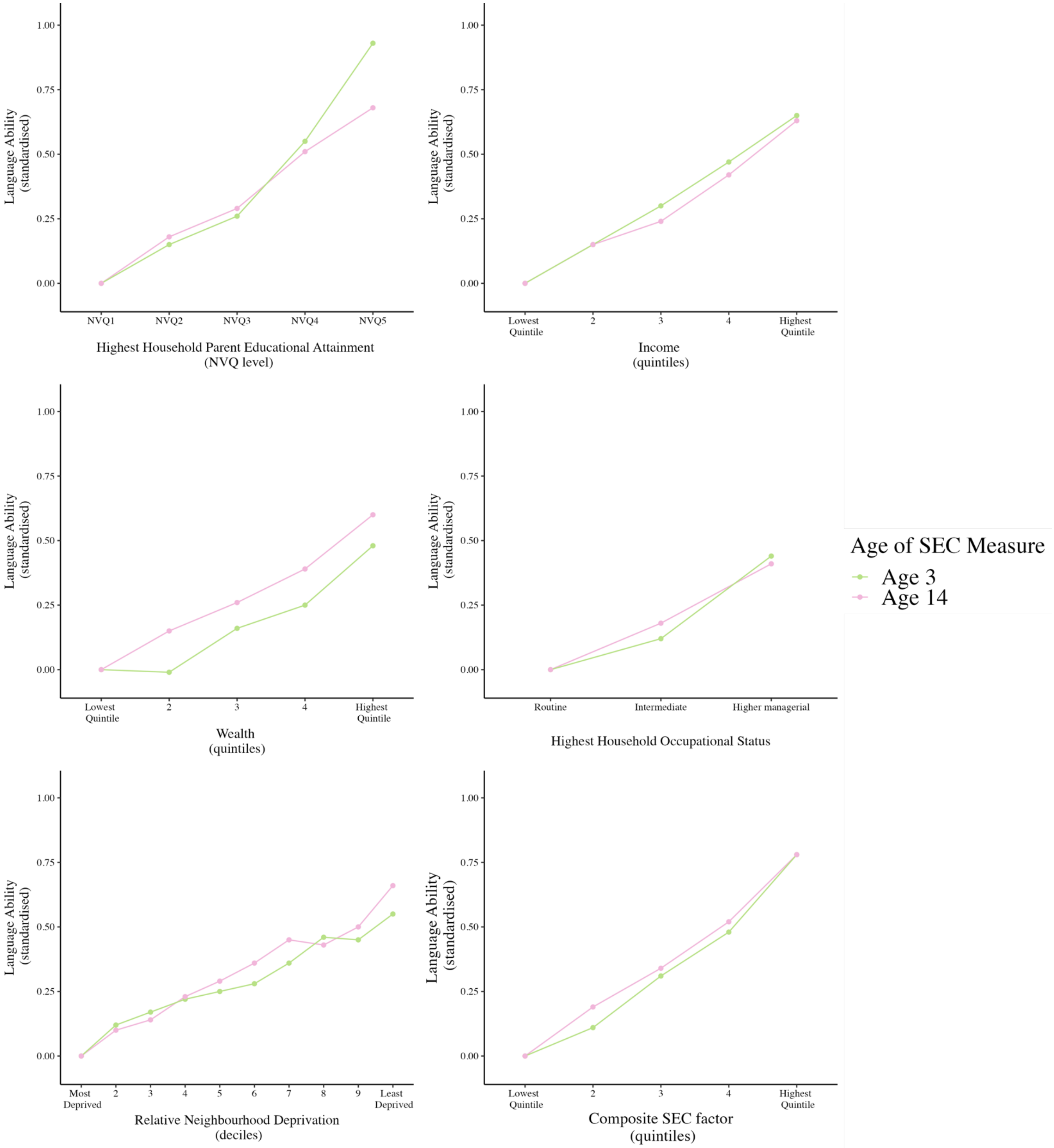
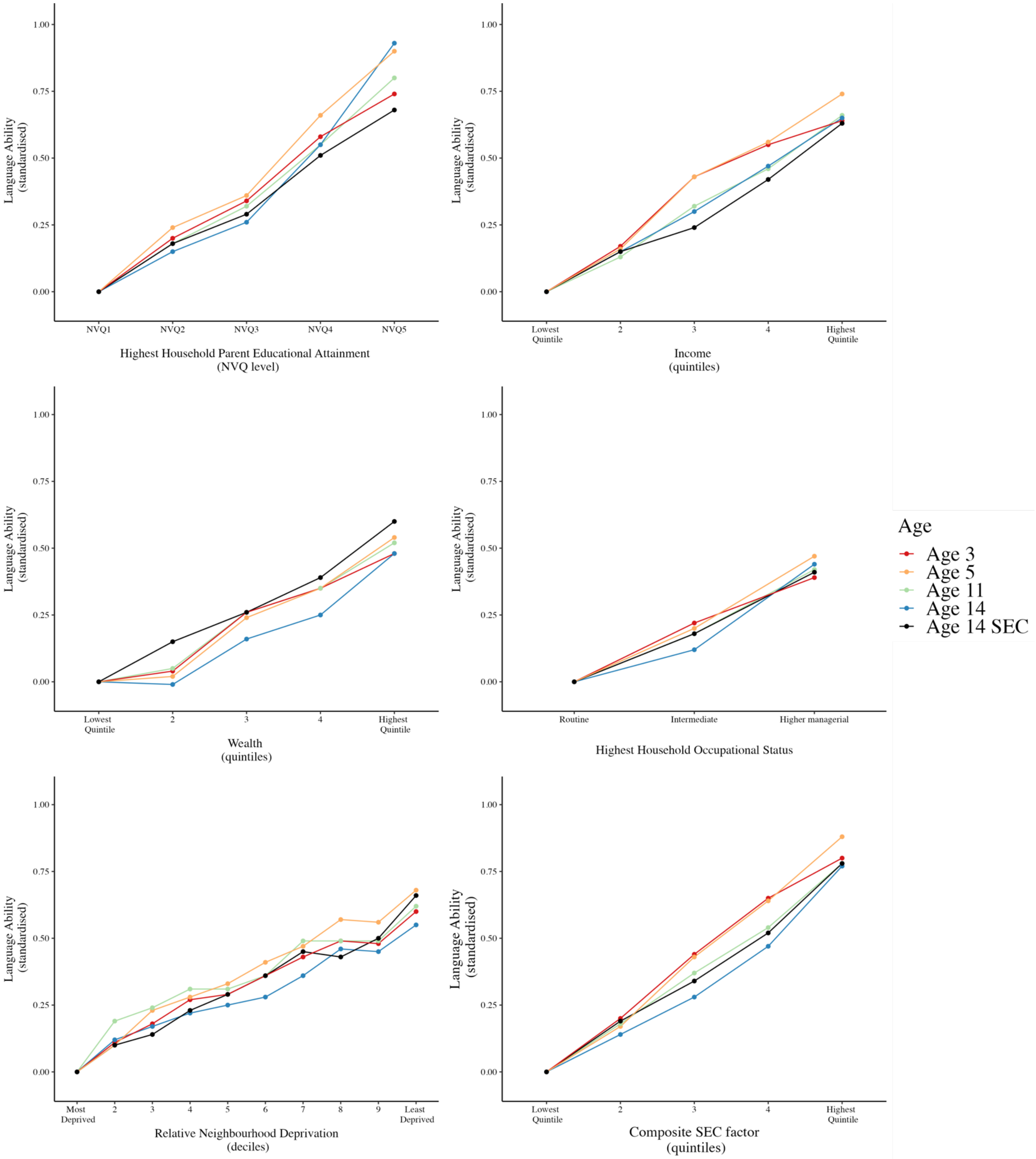


Figure S11. ***Relationships between SEC indicators and Vocabulary at ages 3, 5, 11 and 14. Age 14 vocab predicted by age 3 and age 14 SEC indicators***



Appendix K

Appendix K contains the descriptive statistics and regression coefficients for the cross-cohort comparison.

Table S9. ***Descriptive statistics for language measure by SEC group in each cohort.***

	BCS1970			MCS2001		
	Early childhood language *Range=0-56*	Late childhood language *Range=0-20*	Adolescent language *Range=0-74*	Early childhood language *Range=10-170*	Late childhood language *Range= 10-179*	Adolescent language *0-20*
Parent Education: No /low level qualifications	32.12(11.36) , [31.87;32.37]	11.37(2.6) ,[11.32;11.43]	38.66(13.24) , [38.38;38.95]	99.49(16.97) [98.93;100.04]	113.52(18.58), [112.91;114.13]	6.08(2.35), [6;6.16]
Parent Education:O levels/GCSEs grades A*-C	36.49(9.93), [36.14;36.84]	12.32(2.46) , [12.23;12.41]	42.31(12.55) , [41.86;42.75]	106.66(14.97), [106.24;107.07]	118.12(16.77), [117.66;118.59]	6.57(2.32) ,[6.51;6.63]
Parent Education: Post 16 education	37.9(10.22), [37.31;38.49]	12.8(2.42), [12.66;12.94]	44.63(12.06) , [43.93;45.32]	110.26(14.29), [109.78;110.73]	122.22(14.93), [121.72;122.71]	7.07(2.41) ,[6.99;7.15]
Parent Education: University level	39.45(10.1), [39.06;39.84]	13.41(2.38) , [13.31;13.5]	48.05(11.62) , [47.6;48.5]	114.88(14.56) [114.43;115.34]	126.7(14.04), [126.26;127.14]	8.28(2.81), [8.19;8.37]
Income (Quintile 1 - Lowest)	30.38(11.64) , [29.98;30.79]	10.99(2.68) ,[10.9;11.09]	37.65(13.43) , [37.18;38.12]	101.32(17.37), [100.75;101.89]	115.78(18.38), [115.18;116.38]	6.4(2.4), [6.32;6.48]
Income (Quintile 2)	33.57(11.28) , [33.16;33.98]	11.78(2.6) ,[11.69;11.87]	40.12(13.32) , [39.63;40.6]	104.47(15.76) [103.93;105.01]	116.73(16.95), [116.15;117.31]	6.53(2.37), [6.45;6.61]
Income (Quintile 3)	35.07(10.73) , [34.69;35.45]	11.99(2.54) , [11.9;12.07]	41.18(12.83) , [40.73;41.63]	107.99(15.45), [107.46;108.52]	120.11(16.5), [119.55;120.68]	6.84(2.53), [6.75;6.92]
Income (Quintile 4)	36.52(10.26) , [36.16;36.88]	12.46(2.49) , [12.37;12.55]	43.17(12.59) , [42.73;43.62]	111.81(14.32), [111.3;112.32]	122.4(15.61), [121.84;122.95]	7.28(2.65), [7.19;7.38]
Income (Quintile 5 - Highest)	38.88(10.03) , [38.49;39.27]	13.14(2.4), [13.05;13.23]	46.26(12.32) , [45.78;46.74]	114.03(13.99), [113.52;114.54]	125.68(14.54), [125.14;126.21]	7.93(2.75), [7.83;8.03]

Occupational Status (Unemployed)	30.95(11.85), [28.7;33.19]	11.56(3.07), [10.98;12.14]	39.31(13.24), [36.81;41.81]	100.2(16.8), [99.63;100.78]	114.16(18.55), [113.53;114.79]	6.19(2.4), [6.11;6.27]
Occupational Status (Routine)	30.59(11.7), [30.1;31.08]	11.01(2.63), [10.9;11.12]	37.43(13.19), [36.88;37.98]	104.92(15.73), [104.42;105.42]	117.12(17.5), [116.57;117.67]	6.56(2.36), [6.48;6.63]
Occupational Status (Intermediate)	34.15(10.78), [33.91;34.39]	11.78(2.55), [11.73;11.84]	40.34(13.11), [40.05;40.63]	108.3(15.63), [107.74;108.86]	120.56(15.85) ,[119.99;121.13]	6.86(2.45), [6.77;6.95]
Occupational Status (higher managerial)	37.52(10.84), [37.21;37.83]	12.86(2.53), [12.78;12.93]	45.1(12.58), [44.74;45.47]	113.56(13.92), [113.2;113.92]	124.85(14.46), [124.48;125.22]	7.74(2.72), [7.68;7.81]

Table S10. ***Associations between SEC and language ability in the MCS2001 and BCS1970 cohorts in early childhood, late childhood and adolescence***

	Indictor	Early Childhood Vocabulary (BCS)	Late Childhood Vocabulary (BCS)	Adolescent Vocabulary (BCS)	Early Childhood Vocabulary (MCS)	Late Childhood Vocabulary (MCS)	Adolescent Vocabulary (MCS)
Parent Education	No/low level qualifications	REFERENCE	REFERENCE	REFERENCE	REFERENCE	REFERENCE	REFERENCE
	O levels/ GCSEs grades A*-C	.35[.30;.40] * * * p<.001	.35[.30;.39] * * * p<.001	.27[.21;.33] * * * p<.001	.30[.26;.35] * * * p<.001	.25[.20;.30] * * * p<.001	.18[.13;.23] * * * p<.001
	Post 16 education	.48[.41;.54] * * * p<.001	.53[.45;.60] * * * p<.001	.43[.35;.52] * * * p<.001	.54[.50;.59] * * * p<.001	.50[.45;.54] * * * p<.001	.37[.32;.43] * * * p<.001
	University level qualifications	.65[.60;.69] * * * p<.001	.76[.71;.82] * * * p<.001	.70[.63;.77] * * * p<.001	.85[.80;.89] * * * p<.001	.76[.71;.81] * * * p<.001	.84[.79;.89] * * * p<.001
Occupa-	Routine	REFERENCE	REFERENCE	REFERENCE	REFERENCE	REFERENCE	REFERENCE
	Intermediate	.28[.23;.34] * *	.28[.23;.33] * * *	.21[.15;.27] * * *	.31[.27;.36] * * *	.25[.20;.30] * * *	.13[.08;.19] * * *

		* p<.001	p<.001	p<.001	p<.001	p<.001	p<.001
	Higher managerial	.62[.56;.68] * * * p<.001	.70[.64;.76] * * * p<.001	.58[.51;.65] * * * p<.001	.44[.39;.49] * * * p<.001	.42[.37;.48] * * * p<.001	.31[.25;.36] * * * p<.001
	Unemployed	.09[-.16;.35] p= .480	.25[.02;.48] * p= .030	.19[-.05;.43] p= .120	.06[.01;.11] * p= .030	.03[-.03;.09] p= .370	-.02[-.08;.03] p= .410
Income Quintiles	Quintile 1 (Most Deprived)	REFERENCE	REFERENCE	REFERENCE	REFERENCE	REFERENCE	REFERENCE
	Quintile 2	.25[.19;.32] * * * p<.001	.29[.23;.34] * * * p<.001	.18[.10;.25] * * * p<.001	.63[.58;.68] * * * p<.001	.56[.51;.61] * * * p<.001	.51[.46;.57] * * * p<.001
	Quintile 3	.37[.31;.43] * * * p<.001	.36[.30;.41] * * * p<.001	.26[.19;.33] * * * p<.001	-.26[-.31;-.20] * * * p<.001	-.17[-.22;-.12] * * * p<.001	-.14[-.19;-.08] * * * p<.001
	Quintile 4	.48[.42;.54] * * * p<.001	.53[.47;.59] * * * p<.001	.40[.33;.46] * * * p<.001	.20[.15;.25] * * * p<.001	.20[.15;.25] * * * p<.001	.12[.06;.18] * * * p<.001
	Quintile 5 (Least Deprived)	.70[.64;.76] * * * p<.001	.79[.74;.85] * * * p<.001	.63[.55;.72] * * * p<.001	.48[.44;.52] * * * p<.001	.45[.40;.49] * * * p<.001	.46[.41;.51] * * * p<.001

All coefficients taken from models adjusted for gender, ethnicity, English as an additional language (EAL) and age of cohort member at the time of the language test.
*p<.05
** = p<.01
*** p<.001

Appendix L

Appendix L contains the methods and results for the sensitivity analysis that used Ridit scores in the cross cohort comparison.

Rationale

The education system and occupational structure of the UK has changed over the period that separates the BCS1970 and MCS2001 cohorts, leading to changes in the composition of these two SEC indicators. We therefore conducted a supplementary analysis to our cross-cohort comparison, whereby highest household occupational status and highest household educational attainment were converted to Ridit scores to aid comparability across cohorts (Donaldson, 1998). Ridit scores put ordered categories onto a scale of 0-1, based on the distribution of the categories within any dataset. The resulting coefficients of regression models with SEC Ridit scores as the predictor provide the slope index of inequality (SII). The SII represents the estimated absolute inequalities in an outcome (here, vocabulary) between the highest and lowest SEC groups (Bann, Johnson, Li, Kuh, & Hardy, 2018; Renard, Devleesschauwer, Speybroeck, & Deboosere, 2019; WHO, 2013) and accounts for the changes in the composition of the SEC indicator (Regidor, 2004; WHO, 2013). Therefore, this method allows us to compare inequalities in vocabulary in two cohorts, despite the underlying distributions of SEC variables differing across cohorts. However, as this is an absolute measure of inequalities, this method is not able to discern gradients within the distribution and so hence this method forms our supplementary analysis.

Method

Highest household educational attainment and highest household occupational status were converted to Ridit scores separately in each cohort. The *toridit()* function from the ridittools package in R was used (Bohlman, 2018). Ridit scores were calculated for each imputed dataset and regression models, where the Ridit score was the main predictor, were ran for each imputed dataset. The results of the regression models were then pooled. This procedure was conducted on separate regression models, where early childhood vocabulary, late childhood vocabulary and adolescent vocabulary in each cohort were the outcome variables. This results in 9 separate regression models in each cohort (see Table S11). All models controlled for gender, ethnicity and English as an additional language (EAL).

Results

Regression coefficients can be found in Table S11. Because our Ridit scores rank occupation and education from the lowest SEC to the highest SEC, positive coefficients are indicative of higher vocabulary abilities among the highest SEC group (WHO, 2013). Coefficients indicate better vocabulary scores in the most advantaged group. This is the case for all ages and in both cohorts.

The results from this supplementary analysis confirm the results of the main cross cohort comparison analysis. As can be seen from Table S11, inequalities based on highest household education are largest in the MCS2001 cohort for early childhood and adolescent vocabulary, but in the BCS1970 cohort for late childhood language ability. Turning to highest household occupational status, inequalities are largest for vocabulary at all ages in the BCS1970 cohort, indicated by the bigger coefficients for this cohort. A comparison of the partial R^2 values for the main analysis and Ridit score analysis can be found in Table S12 and Figure S12. These are similar across both analyses.

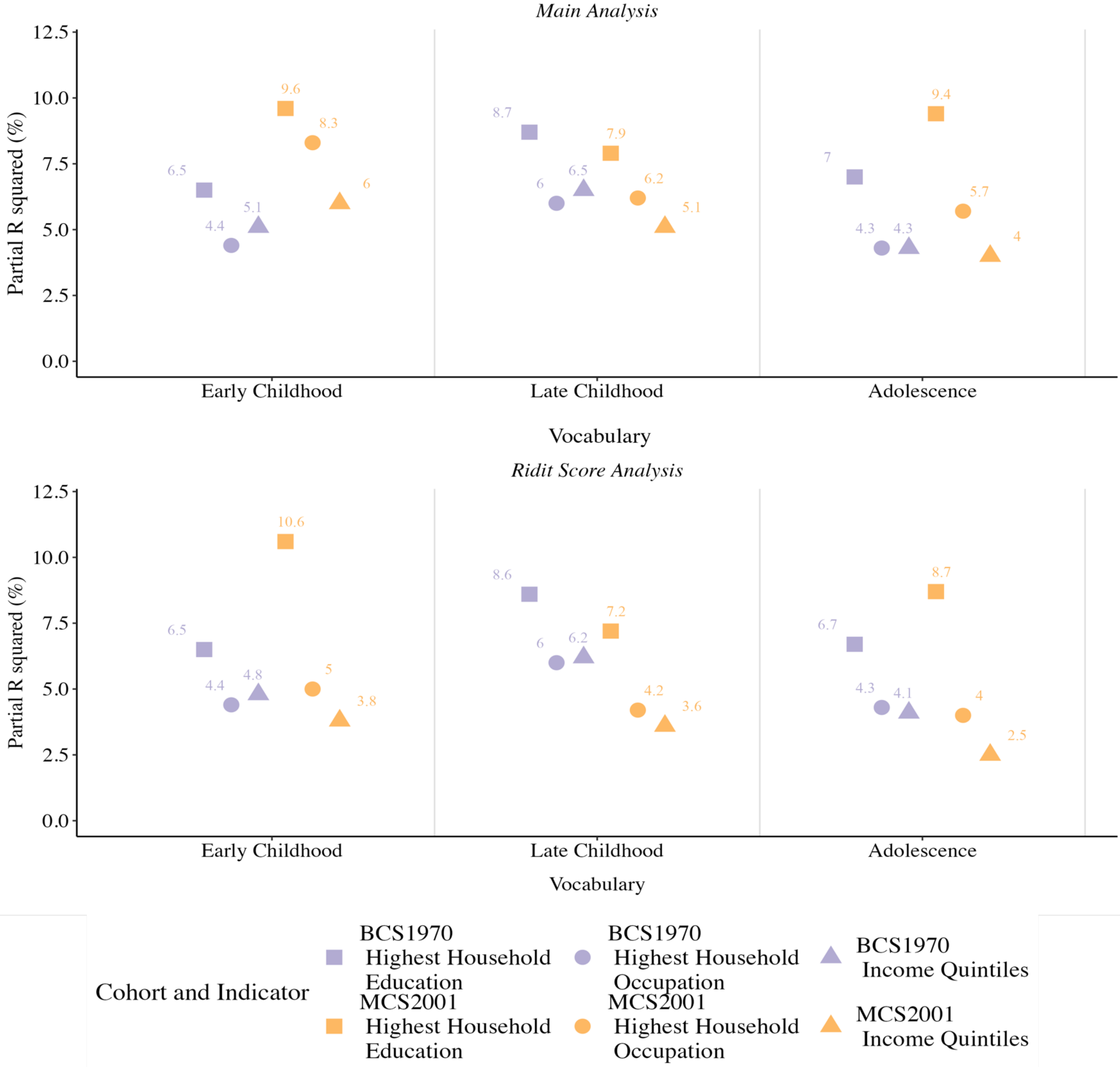
Table S11. ***Regression coefficients for models predicting vocabulary using SEC Ridit scores***

	β [95% CIs] *p value*					
	Highest Household Education (Ridit score)		Highest Household Occupation (Ridit score)		Income (Ridit score)	
Vocabulary	1970	2001	1970	2001	1970	2001
Early Childhood	.98[.92;1.05]*** *p<.001*	1.08[1.03;1.14]*** *p<.001*	.81[.73;.88]*** *p<.001*	.78[.71;.84]*** *p<.001*	.76[.47;1.06]*** *p<.001*	.64[.59;.70]*** *p*<.001
Late Childhood	1.12[1.05;1.19]*** *p<.001*	.98[.93;1.04]*** *p<.001*	.94[.86;1.01]*** *p<.001*	.72[.65;.78]*** *p<.001*	.86[.53;1.19]*** *p<.001*	.62[.56.69]*** *p<.001*
Adolescent	.99[.89;1.08]*** *p<.001*	1.07[1.01; 1.13]*** *p<.001*	.79[.70;.88]*** *p<.001*	.76[.69;.84]*** *p<.001*	.70[.42;.97]*** *p<.001*	.57[.49; .65]*** *p<.001*

Table S12. ***Partial R^2 values for Ridit scores predicting vocabulary throughout childhood in two cohorts***

	Partial R^2 (%)					
	Highest Household Education (Ridit score)		Highest Household Occupation (Ridit score)		Income (Ridit score)	
	1970	2001	1970	2001	1970	2001
Early Childhood Vocabulary	6.5	10.6	4.4	5	4.8	3.8
Late Childhood Vocabulary	8.6	7.2	6	4.2	6.2	3.6
Adolescent Vocabulary	6.7	8.7	4.3	4	4.1	2.5

Figure S12. ***Partial R² Values for SEC indicators in each cohort: Comparison of Main and Ridit Score Analyses***



Appendix M

Appendix M contains the sensitivity analysis for the cross-cohort comparison which included only those of a White ethnicity.

Method

Vocabulary measures (cross-cohort comparison).

Early language ability. For the BCS1970 cohort, receptive vocabulary was measured at age 5 using the English Picture Vocabulary Test (EPVT), a UK version of the Peabody Picture Vocabulary Test (Brimer, 1962; Dunn, 1965). Cohort members were shown 56 sets of four diverse images and heard a specific word associated with each set of four images. They were asked to select one picture that matched the presented word and were awarded one point for every correct response(Moulton, 2020; Parsons, 2014). For the MCS2001cohort, expressive vocabulary was measured using the naming vocabulary sub-test of the BAS II (Colin D. Elliott, 1996). We adjusted for age in months at the time of the test in both cohorts. All scores and ages were converted to *z* scores for analyses.

Late childhood language ability. When the BCS1970 cohort members were aged 10, they completed the BAS word similarities subscale (Elliott, 1979). The test was made up of 21 items, each of which consisted of three words. The teacher read these sets of items out loud and cohort members had to a) name another word that was consistent with the three words in the item and b) state how the words were related. In order to receive a point, cohort members had to correctly answer both parts of the question (Moulton, 2020; Parsons, 2014). Details on the scoring of this vocabulary measure and the SPSS syntax used can be found in appendix 3 of "Childhood Cognition in the 1970 British Cohort Study" (Parsons, 2014). When MCS2001 cohort members were aged 11, they completed the BAS II verbal similarities subscale (detailed above). As already mentioned, test scores for the MCS2001 cohort were adjusted for item difficulty. In both cohorts, we controlled for age at the time of the test and converted all scores to *z* scores.

Adolescent language ability. When aged 16, BCS1970 cohort members completed the APU Vocabulary Test (Closs, 1986). This consisted of 75 items: an item consisted of a target word, presented with a multiple-choice list, from which cohort members had to select a word that meant the same as the target word(Moulton, 2020; Parsons, 2014). These items got progressively harder throughout the test. Details on the scoring of this vocabulary test can be found in appendix 3(Parsons, 2014). When MCS2001cohort members were aged 14, they completed the Word Activity Task (detailed above). Words used in the Word Activity Task were a subset of the words used in the BCS1970 cohort Vocabulary Test, which cohort members completed aged 16(Fitzsimons, 2017). Scores were adjusted for age and converted to *z* scores for analyses.

Indicators of socioeconomic circumstance.

Harmonised measures of the following two indicators were used as measures of SEC:

Parent education. The highest academic qualification achieved in the household when the cohort member was aged 5. Where this information is missing, information from previous sweeps was used.

Occupational status. Highest household occupational status at age 5. For the BCS1970 cohort, this was ascertained with the Registrar General's classification. For the MCS2001cohort, the NS-SEC classification system was used. Where this information is missing, information from previous sweeps was used.

Analysis plan.

We had 3 separate outcome variables in each cohort (early childhood language ability, late childhood language ability and adolescent language ability). We built two regression models per outcome, one with occupational status as the predictor variable and the other with parent education as the predictor variable. Because our measures of language ability were standardised within each cohort, we were able to directly compare coefficients between cohorts and establish the rate of inequality in language ability at each age in the two cohorts.

Results

Partial R^2 values can be found in Table S13, and regression coefficients can be found in Table S14. The results from this sensitivity analysis confirm the results of the main cross cohort comparison analysis. As can be seen from Table S14, inequalities based on highest household education are largest in the MCS2001 cohort for early childhood and adolescent vocabulary, but in the BCS1970 cohort for late childhood language ability. Turning to highest household occupational status, inequalities are largest for vocabulary at all ages in the BCS1970 cohort, indicated by the bigger coefficients for this cohort. Thus, the ethnic composition of the two cohorts do not appear to be driving the results of our cross-cohort comparison.

Table S13. ***Partial R2 Values for cross-cohort comparison (%)***

	Partial R^2 (%)					
	Highest Household Education		Highest Household Occupation		Income	
	1970	2001	1970	2001	1970	2001
Early Childhood Vocabulary	7.6	10.2	5.4	9.1	6	6.4
Late Childhood Vocabulary	9.3	7.9	6.5	6.5	6.5	5.1
Adolescent Vocabulary	7.1	9.6	4.7	5.9	4.3	4

Table S14. ***β[95% CIs] for SEC predicting vocabulary in MCS2001 and BCS1970 Cohorts***

			BCS1970 Cohort			MCS2001 Cohort	
		Indicator	Early Childhood Vocabulary	Late Childhood Vocabulary	Adolescent Vocabulary	Early Childhood Vocabulary	Late Childhood Vocabulary
Parent Education	No/low level qualifications	REFERENCE	REFERENCE	REFERENCE	REFERENCE	REFERENCE	REFERENCE
	O levels/ GCSEs grades A*-C	.37[.32;.41] * * * p<.001	.35[.30;.40] * * * p<.001	.27[.21;.33] * * * p<.001	.30[.26;.35] * * * p<.001	.26[.21;.32] * * * p<.001	.19[.13;.25] * * * p<.001
	Post 16 education	.49[.43;.56] * * * p<.001	.54[.47;.61] * * * p<.001	.44[.36;.53] * * * p<.001	.53[.48;.58] * * * p<.001	.50[.45;.56] * * * p<.001	.39[.32;.45] * * * p<.001
	University level qualifications	.66[.61;.70] * * * p<.001	.78[.73;.83] * * * p<.001	.70[.62;.77] * * * p<.001	.82[.77;.87] * * * p<.001	.76[.70;.81] * * * p<.001	.86[.80;.92] * * * p<.001
Occupational Status	Routine	REFERENCE	REFERENCE	REFERENCE	REFERENCE	REFERENCE	REFERENCE
	Unemployed	.06[-.19;.32] p= .630	.19[-.09;.47] p= .190	.15[-.23;.54] p= .440	.03[-.02;.08] p= .280	.01[-.06;.08] p= .770	-.03[-.09;.04] p= .400
	Intermediate	.30[.25;.35] * * * p<.001	.28[.23;.34] * * * p<.001	.23[.16;.30] * * * p<.001	.29[.24;.34] * * * p<.001	.24[.19;.30] * * * p<.001	.13[.07;.19] * * * p<.001

	Higher managerial	.66[.60;.71] * * * p<.001	.72[.66;.79] * * * p<.001	.61[.53;.69] * * * p<.001	.42[.37;.47] * * * p<.001	.41[.35;.47] * * * p<.001	.30[.24;.36] * * * p<.001
Income Quintiles	Quintile 1 (Most Deprived)	REFERENCE	REFERENCE	REFERENCE	REFERENCE	REFERENCE	REFERENCE
	Quintile 2	.27[.22;.33] * * * p<.001	.28[.22;.33] * * * p<.001	.17[.10;.24] * * * p<.001	.58[.53;.63] * * * p<.001	.53[.48;.59] * * * p<.001	.50[.44;.56] * * * p<.001
	Quintile 3	.39[.33;.45] * * * p<.001	.34[.28;.40] * * * p<.001	.26[.19;.33] * * * p<.001	-.28[-.34;-.22] * * * p<.001	-.20[-.26;-.15] * * * p<.001	-.16[-.22;-.10] * * * p<.001
	Quintile 4	.51[.45;.57] * * * p<.001	.52[.46;.59] * * * p<.001	.39[.32;.46] * * * p<.001	.19[.14;.23] * * * p<.001	.20[.15;.26] * * * p<.001	.12[.05;.19] * * * p<.001
	Quintile 5 (Least Deprived)	.72[.66;.78] * * * p<.001	.78[.72;.85] * * * p<.001	.63[.54;.72] * * * p<.001	.44[.40;.48] * * * p<.001	.42[.38;.47] * * * p<.001	.46[.40;.51] * * * p<.001

License

Early vocabulary and grammar development in Albanian-speaking children: An MB-CDI adaptation study

Enkeleida Kapia
University Ludwig Maximilian, Munich, Germany
Albanian Academy of Sciences, Tirana, Albania

Shanley E. M. Allen
University of Kaiserslautern-Landau, Germany

Doruntinë Zogaj
University of Saarbrucken, Germany

Abstract: This is the first exploratory study of a large sample of Albanian-speaking children and their communicative development, based on the Albanian adaptation of three sections of the MacArthur–Bates Communicative Development Inventory II Words and Sentences: The vocabulary checklist (Part I Section A), How children use words (Part I Section B), and Word Endings (Part II Section A). Parental report data were collected from 112 children between the ages of 13 and 36 months. Correlation analyses for early vocabulary were conducted. Developmental trajectories of children were compared based on the demographic characteristics of sex and parental education. Results show that lexical growth for Albanian is comparable to that reported for other languages, and neither sex nor maternal education level correlate with vocabulary size. However, fathers' educational level correlates with production of early morphological markers. We conclude that the Albanian CDI is a useful tool for the assessment of and research on the language development of Albanian-speaking children. Future directions include testing larger numbers of children from different dialect and socioeconomic backgrounds.

Keywords: MB-CDI, early lexical and grammar development, Albanian

Corresponding author: Enkeleida Kapia, Institute for Phonetics and Speech Processing, LMU, Schellingstrasse 3, 80799, Munich, Germany. Email: enkeleida.kapia@phonetik.uni-muenchen.de.

Orcid ID: https://orcid.org/0000-0002-9676-3509

Citation: Kapia, E., Allen, S. E. M., & Zogaj, D. (2024). Early vocabulary and grammar development in Albanian-speaking children: An MB-CDI adaptation study. *Language Development Research, 4*(1), 175–206.http://doi.org/10.34842/wve8-yn80

Introduction

To date very little research exists on the early language development of Albanian-speaking infants and toddlers. This limits our understanding of how Albanian is learned and what the relationship is between various aspects of early language development. Information about Albanian-speaking children's language development from an early age is valuable not only for the purpose of discovering milestones and developmental steps for children with typical development, but also for comparing the process of language development in Albanian with that of other languages with the same or different typologies. Furthermore, such information is essential for identifying children who have developmental delays or other atypical language profiles such as autism and aphasia, as well as for creating and evaluating effective intervention strategies for language disorders, and for distinguishing linguistic and cognitive delays.

The present study is the first to provide a detailed exploratory examination of the early acquisition of vocabulary and grammar in Albanian. It is based on data from 112 Albanian-speaking infants and toddlers aged 13 to 36 months using the Albanian adaptation of three sections of the MacArthur-Bates Communicative Development Inventories II Words and Sentences: The vocabulary checklist (Part I Section A), How children use words (Part I Section B), and Word Endings (Part II Section A). To this end, this study describes the characteristics of early vocabulary in child Albanian including its size, its composition and some language-internal and language-external factors that may be associated with it.

Early vocabulary and grammar development in child language

Results from studies of numerous languages using their own adaptations of the MB-CDI indicate very similar routes and speeds in early vocabulary and grammar development across languages among infants and toddlers (Bates et al., 1994; Bleses et al., 2008; Devescovi et al., 2005; Frank et al., 2021; Gendler-Shalev & Dromi, 2022; Stolt et al., 2009). Despite large individual differences (Bates et al., 1994; Bleses et al., 2008; Devescovi et al., 2005; Fernald et al., 2001), most children around the world are found to produce their first words between 1;0 and 1;8 years of age, experiencing a vocabulary spurt soon after 1;6 (Bates & Goodman, 2001; Fernald et al., 2001).

Early vocabulary consists of items that denote concrete things in children's immediate environment, such as body parts, kitchen tools, food and drinks, toys, routine activities, etc. (Caselli et al., 1999; Karmiloff & Karmiloff-Smith, 2001). These open-class or content words, which show up in high numbers during the first years of life, have been argued to aid the later emergence of closed-class words such as prepositions and articles (Bates & Goodman, 2001). The later acquisition of closed-class words has also been linked to them being phonologically shorter and less emphasized in speech, which makes them harder to be noticed by children from early on (Morgan et al., 1996).

In addition, extensive data from children crosslinguistically has shown that the size of early vocabulary is closely linked to the development of grammar. For example, for two-word utterances to begin appearing in children's speech, children need to have already acquired around 50-100 words (Bates & Goodman, 2001; Marchman & Bates, 1994). A recent study looking at data from nearly 8000 English-speaking infants and toddlers has suggested that lexical and syntactic ability begins to grow roughly in the 20- to 24-month age range; however, variability is high and this ability does not stabilize until children reach 30 months or so (Day & Elison, 2022). This study reveals a distinct lag effect, where children's ability to use syntax is almost always lower than their vocabulary skills. However, although a relationship between early vocabulary size and later morphological and syntactic development is evident across languages, this relationship is not identical across languages, reflecting differences in language structure as well as cultural differences in language use. For example, Thordardottir et al. (2002) show that Icelandic-speaking children need to have acquired a larger vocabulary than English-speaking children before they begin to use plural inflections on nouns and past tense inflections on verbs, linking this observation with the more complex inflectional system in Icelandic as compared to English. Thus, it is important that more studies investigate the timing of the relationship of early grammar to early lexicon, especially from languages with a complex grammatical system such as Albanian.

And lastly, long standing work has also shown that several demographic factors including sex and parental education play a role in early language development. On average, it has been reported that boys produce fewer words at a given age than girls (Bates et al., 1994; Bouchard et al., 2009; Day & Elison, 2022; Hulle et al., 2004), and children of more highly educated mothers (Bates et al., 1994) – with mixed findings with respect to fathers (Bates et al., 1994; Pancsofar & Vernon-Feagans, 2006) – produce more words.

The main objective of the present study is to build on this literature in order to establish the characteristics of early vocabulary in Albanian-speaking infants and toddlers from 13 – 36 months of age, as reported by the Albanian MB-CDI. In Section 2 we present general information about the Albanian language and its early development in infants and toddlers, as well as detail the development of the Albanian MB-CDI form.

Albanian language and cultural context

Albanian language

Albanian is a language of the Indo-European family with 6-7 million speakers (J. S.

Klein et al., 2017; Rusakov, 2017) who live mostly in the Republic of Albania, the Republic of Kosovo, Montenegro and the Republic of North Macedonia, as well as in Albanian-speaking minority communities in Italy, Greece, Croatia and Ukraine. In this article, however, we focus on Albanian as spoken in Albania. Albanian is widely accepted to form a branch of its own within the Indo-European language family (Bopp, 1855; Çabej, 1976; Pedersen, 1897); no evidence to date relates Albanian to any other language within this family (Demiraj, 2018). Albanian has traditionally been described as comprising two main dialects: Gheg, spoken in northern and central Albania, and Tosk, spoken in the south of the country (e.g. Desnickaja, 1976; Gjinari, 1988; Hahn, 2013). The Shkumbin River located in the center of the country forms the approximate boundary between the two dialect regions. A third variety, the Standard variety, is also present alongside the two dialects; it was institutionalized in 1972 via a National Congress of Orthography and was based mostly on the Tosk dialect (Ismajli, 1998). Standard Albanian, which will be the focus of this article, is characterized by a relatively free word order (Rushi, 1983) and has a fairly complex fusional morphology (Agalliu et al., 2002). For example, it has noun declensions in 5 different cases, which are also marked for gender, number, and definiteness, as well as verbs which are marked for mood, tense, person, and number (Agalliu et al., 2002).

Albanian early language development

Although children's language acquisition has been studied extensively in many languages, research on the acquisition of Albanian has been almost completely absent from this literature. Exceptions are just a few studies published in recent years (Cenko, 2017; Cenko & Budwig, 2007; Kapia, 2010, 2014; Shashaj, 1996), and a couple of recent Master's theses (Dule, 2023; Jia, 2023; Sehitaj, 2015; Zogaj, 2021). These studies provide a preliminary view of children's speech at an early age, with quantitative and qualitative descriptions focusing on the lexicon, syntax and, more broadly, grammatical constructions.

The earliest study that focused on children's language development was motivated by the educational policy demands during the communist regime (Shashaj, 1996). This policy sought ways to understand how children acquired language in relation to general education, with the purpose of influencing their language habits more towards the Standard norm. Shashaj (1996) investigated the speech of three children, of which two produced their first words before 8 months old, and the third child after 12 months. According to this study, which does not report any details about its methodology, Albanian-speaking children at age 1 typically produce around 20 words while children at age 3 produce around 1000 words. These words are not the same for every child, but depend on the environment in which they grow up including their home environs, playmates, kindergarten, and caregivers (Shashaj, 1996).

A small number of studies have also investigated the acquisition of syntax and pragmatics, and provide preliminary reports about children's ability to combine words together into phrases or initial short sentences. Shashaj (1996) reports that children begin to be able to form more complex phrases between the second and third year of life. Studies by Cenko and Budwig (2007) as well as Kapia and colleagues (Kapia, 2010, 2014) have focused on different language structures and different stages of acquisition across age groups in monolingual and bilingual children. Kapia (2010, 2014), for instance, found that Albanian-speaking children perform at an adult-like level with clitic doubling of dative object nouns at around age 2-3 years, but show difficulty with clitic doubling of accusative object nouns since these require differentiation between old and new information – a pattern which seems to suggest a delay in their development of pragmatics, but not of syntax. In another study, Cenko & Budwig (2007) found that most 2-year-old Albanian-speaking children are able to use at least one verb with the correct morphological marking in both the transitive and the unaccusative form. This demonstrates their flexibility in verb construction use, in contrast to the findings for English-speaking 2-year-old children. This flexibility was suggested to be linked to the rich morphological markings on the verb that emphasize the differences between transitive and unaccusative constructions, consistent with similar findings for other morphologically rich languages (Cenko & Budwig, 2007). Apart from these studies, our knowledge about lexical and morphosyntactic development in Albanian is very limited, which is why more detailed and larger scale studies are needed at this point. In the present study, we aim to fill some of this gap by focusing on early vocabulary and grammar development, their relation to each other, and some of the language-external factors that may influence both of these components of early language. We achieve this by investigating early language development using the Albanian MB-CDI tool.

In terms of vocabulary learning, we do not have any reason to believe that Albanian-speaking children generally learn words any differently than children from other language contexts, apart from perhaps learning certain words earlier or later depending on the cultural context. Nonetheless, determining the trajectory of vocabulary developmental for Albanian-speaking children specifically is crucial for clinical and educational reasons given that assessing their vocabulary level using norms taken from other languages is likely to lead to inaccurate results. In relation to how children use words (operationalized by 'displaced events' in the CDI), to the extent that these are conceptual, we also do not think that Albanian-speaking children will show different developmental tracks than children learning other languages. Where we do expect a difference, however, is the rate of acquisition of morphology based on the fact that Albanian has quite a complex morphological system. For instance, verb forms are often realized not just by adding a suffix or one or two particles to them, but also sometimes by metathesis or stem change altogether, as is the case for *ha* 'to eat' vs *hëngra* 'ate' vs *pata ngrënë* 'had eaten' or *shoh* 'to see' vs *shihja* 'saw' vs *pata parë* 'had seen'. We predict that learning of grammatical morphemes will take longer and occur later in life than for children with simpler morphological systems such as English.

Social context

Considering that parental gender and education has been reported to be linked to early lexical and grammatical development (see Section 1.1), a secondary aim of this study is to investigate whether mother's and father's educational levels play a role in early child language (vocabulary and grammar) development. Thus, we outline here some relevant details about the role of women and men in Albanian society. Although Albania is now on the rise from one of the poorest countries in Europe to a middle-income country, there are still gaps in both education levels and labor market opportunities between women and men. The World Bank reports that, unlike in most patriarchal societies, more women than men in Albania receive a postsecondary education, with the gap between genders being around 25% (World Bank, 2018). However, when it comes to the labor market, women's participation in the labor force has dropped drastically from 78% in 1989 to 46% in 2005, finally reaching 50% participation in 2013 (INSTAT, 2017). This is likely due to the Albanian society holding onto strong patriarchal values that place women of reproductive age outside the labor market, with few opportunities for retraining and qualification (Young, 2018).

Development of the Albanian MB-CDI form

Version 1

The present adaptation of the CDI for Albanian was developed by the first author in collaboration with Enila Cenko (University of New York Tirana), a researcher of early child development, in close consultation with the second author and Nancy Budwig (Clark University). They received authorization for the Albanian adaptation project from the CDI Advisory Board in 2008, and had regular cooperation with the European Network on Communicative Development Inventories including researchers from all over Europe developing CDIs for various European languages. The adaptation was completed in 2010.

The starting point in constructing the Albanian CDI was the American English MB-CDI *Words and Sentences* which consists of two parts (Bates et al. 1994; Fenson et al., 2007). Part I, *Words Children Use,* focuses on the child's use of words and is split into sections A and B. Section A, *Vocabulary Checklist,* is a checklist of 680 words divided into 22 semantic categories. Section B, *How Children Use Words,* asks whether or not the child has started using words to talk about displaced events such as events in the future or past, or objects not present in the context.

Part II, *Sentences and Grammar,* deals with early grammar and is divided into five sections. Section A, *Word Endings I,* tests the child's use of word endings such as -ed, -ing etc. Section B, *Words Forms,* deals with word forms, nouns and verbs. Section C, *Word Endings II,* checks the proper use and errors when using word endings. Section D, *Examples,* asks for the three longest sentences the child has said recently. Finally, Section E, *Complexity,* deals with the complexity of the child's morphosyntactic skills, for example, does the child say "two foot" or "two feet"?

Our adaptation of the Albanian CDI only includes Part I Sections A & B and Part II Section A. Adaptations of the remaining sections are ongoing.

Part I Section A: Lista e fjalëve (Vocabulary Checklist)

For the adaptation of Part I Section A, we began by translating all the words of the American English version (all 22 semantic categories) from English to Albanian. We then had a person unfamiliar with the project translate those items back from Albanian to English, in order to ensure that the back-translation resulted in the same items as originally intended. All words at this step of the process were back-translated as intended.

Some sections, however, required adaptation due to differences between English and Albanian morphosyntax. In the action word category, for example, listing verbs in their root form as in English would not be appropriate because Standard Albanian (the dialect used here) lacks infinitives. Instead, verbs were listed in typical citation form used in Albanian dictionaries: first person singular form using indicative mood, active voice and present tense, and preceded by a first person pronoun (e.g., *(unë) punoj* '(I) work', *(unë) shkruaj* '(I) write'). Differently from English, in Section II Part A (Word Endings I), we inquired whether children have knowledge of the subjective form of verbs which are formed with particles *për të* + verb. These forms denote an action to be completed and are the quasi-analogue of the infinitive which only exists in the Gheg dialect, but not in the Standard variety tested here (Cipo, 1949). Furthermore, adjectives in Albanian are obligatorily inflected for gender. Thus, the words in the descriptive word category included the relevant gender variations (e.g., *i bukur / e bukur* 'beautiful.M / beautiful.F', *jeshil / jeshile* 'green.M / green.F). Additionally, words in the pronoun category were changed to fit the Albanian pronoun system which agrees in gender with the noun it refers to, so gender variations were included when necessary (e.g., *i imi / e imja* 'mine.M / mine.F', *këta / këto* 'these.M / these.F').

Once the list was finalized, we tested it with two focus groups comprised of caregivers of children aged 1-3 years. Focus Group I, consisting of 15 caregivers (7 had 8 or fewer years of education) received the full vocabulary checklist. Their task was to mark the words that their children produced and comprehended. Focus Group II, with another set of 15 caregivers (6 had 8 or fewer years of education), received a blank form with just the semantic categories of the American English MB-CDI. Their task was to write down as many words as they could remember that their children produced and comprehended within these categories and add any other words that were not captured by the categories on the form. After the two focus groups, we combined the responses from the 30 participants into one large list of words. For each word, we determined the number of children out of 30 that produced that word, and rank-ordered the words in terms of the frequency of occurrence. Several words on the original list of the American English MB-CDI were excluded such as *hamburger, peanut butter, babysitter, backyard, soda,* and *pancake*. Other words relevant to Albanian culture and culinary practices were added in, such as *çiçi* 'peepee', *dum dum* 'small van', *kola* 'coca cola', *petulla* 'a type of fried dough', *byrek* 'a type of savory pie', and *gjizë* 'a type of cottage cheese'. The final result was the first complete draft of Part I Section A, *Lista e fjalëve* (Vocabulary Checklist), of the Albanian CDI.

Part I Section B: Si i përdorin fjalët fëmijët (How Children Use Words)

Part I Section B comprised five questions that check whether children refer to displaced events. Two of these questions focus on concepts of time, asking whether the child talks about events in the past (e.g., the child visited the beach last week, later he/she mentions sea, sun, sand) or in the future (e.g., if they are going to visit grandmother, the child says *nëna* 'grandmother'). Three further questions focus on displaced objects and people, asking whether the child talks about objects or people that are not present (e.g., asking for the father while he is at work), whether the child understands requests for objects or people that are not present (e.g., when asked "where is the ball?", s/he can get the ball from another room), and if the child can point to an object that belongs to a person that is not present (e.g., the child can point to mother's shoes and says "mother"). For each question, caregivers selected either 'never', 'sometimes', or 'often' as their answer. We discussed each of these questions with the caregivers in Focus Groups I and II. These consultations resulted in no changes in Part I Section B of the Albanian CDI from the American English CDI, as all the categories of time and displaced events were deemed appropriate and necessary.

Part II Section A: Mbaresat e fjalëve I (Word Endings I)

Finally, Part II Section A is constructed in the same way as in the American CDI, and also benefitted from discussions with caregivers from both Focus Groups mentioned earlier. Our goal in adapting this section was to include elements of grammar that are relevant to the rich morphology of noun and verbal systems in Albanian. Consultations with caregivers revealed similar elements of grammar as relevant for this part of the Albanian CDI as for the American English CDI: plurality in nouns and several different verb tenses. This section contains five questions, one each about whether the child uses the forms in (1-5), whose morpheme-by-morpheme glosses follow the Leipzig glossing conventions (Croft, 2002; Lehmann, 1982) .

(1) Regular plural forms
vajzë *vajza*
girl.F.NOM.SG girl.F.NOM.PL

(2) Verbs in simple past tense
ha *hëngra*
eat.IND.PRS.1SG eat.IND.PST.1SG

(3) Verbs in present continuous tense
jam duke ngrënë
am -ing eat.PRS.PTCP

(4) Verbs in indicative form
për të ngrënë
for to eat.PRS.PTCP

(5) Verbs in future tense
do të ha
will to eat.PRS.IND

Questions are structured as in (6).

(6) *A e keni dëgjuar fëmijën tuaj të përdori emra në shumës, si për shembull: libër-libra, vajzë-vajza, mollë-mollë?*
'Have you ever heard your child use nouns in plural, as in the examples book-books, girl-girls, apple-apples?'

For each question, caregivers selected either 'never', 'sometimes', or 'often' as their answer.

In sum, Version 1 of the Albanian CDI form resulted in having two parts: Part I Section A *Lista e fjalëve* (Vocabulary Checklist) and Section B *Fjalë të tjera që përdorin fëmijët* (Words Children Use), as well as Part II Section A *Mbaresat e fjalëve* (Word Endings).

Version 2

Version 1 of the Albanian CDI form was tested in a pilot study with a group of 40 parents and grandparents (it is quite common in Albania that children grow up in extended households where they receive input from grandparents and parents at the same time). The results of this pilot were used to develop Version 2.

Words from Part I Section A were rank-ordered in terms of frequency of occurrence.

Words that never occurred were removed from the CDI, for example *dëshiroj* 'wish'. Table 1 compares the number of words per category of Part I Section A of Version 2 of the Albanian CDI with the American English, Danish and Norwegian forms (because these were the CDI forms for which we could readily find word lists). As seen from this table, the difference between these forms in the number of words for each semantic category is relatively small. One aspect to highlight here is that the biggest differences are in action words, descriptive words and prepositions and locations. The reasons for these differences are many, but in the case of action words, for example, some words in the American CDI such as *skate* or *ride* are irrelevant in the Albanian context, as there are not many opportunities for skating in the warm Albanian Mediterranean climate or for riding since not many families own cars and neither do children ride ponies or bikes when they are little, as is pretty standard in some Western cultures. Another example in this section that seemed inappropriate to focus groups at the time of the adaptation was the action word *hate;* no one imagined children this young using the word *hate* 'urrej' in their daily speech. Other examples of the American CDI action word list that ended up being deleted from the Albanian counterpart were examples that were not expressed via one word but a group of words, such as *cuddle* 'rri gushe-gushe/përqafuar dhe duke u përkëdhelur'. These expressions were generally avoided unless they denoted some really basic activity in children's lives such as *pee* 'bëj çiçin'.

Table 1: ***Comparison of the categories and number of items in the vocabulary lists of Albanian, Danish, Norwegian and American MB-CDIs***

	Albanian	American English	Danish	Norwegian
1. Sound effects & animal sounds	11	12	12	12
2. Animals (real or toy)	42	43	43	44
3. Vehicles (real or toy)	15	14	14	14
4. Toys	12	18	18	18
5. Food and drink	70	68	68	68
6. Clothing	34	28	30	30
7. Body parts	24	27	28	27
8. Small household items	70	50	50	50
9. Furniture and rooms	-[1]	33	33	34
10. Outside things	31	31	31	31
11. Places to go	18	22	22	22
12. People	31	29	40	36

[1] *Small household items* and *Furniture and rooms* are combined in the Albanian CDI in one semantic category labelled *Things and rooms around the house*.

13. Games and routines	32	25	27	27
14. Action words	83	103	103	108
15. Descriptive words	54	63	63	62
16. Words about time	15	12	15	16
17. Pronouns	26	25	31	31
18. Question words	7	7	7	7
19. Prepositions and locations	29	26	41	41
20. Quantifiers and articles	14	17	21	22
21. Auxiliary verbs	26	21	21	22
22. Connecting words	6	7	6	9
Total Vocabulary	**650**	**680**	**725**	**731**

Part I Section B and Part II Section A remained the same as in the previous version since they functioned as expected and no changes were necessary. This version - Version 2 - was used in three unpublished MA theses (Jia, 2023; Sehitaj, 2015; Zogaj, 2021), the latter two of which comprise the data for the present study.

Methods

Participants and materials

A total of 112 Albanian-speaking infants and toddlers aged 13 to 36 months old were recruited for this study. Not much is known about Albanian vocabulary acquisition. Thus, in line with the protocol followed during the earlier phases of the development of the American English CDI inventories (Fenson et al., 2007), we wanted to ensure that the form we chose to adopt first showed steady developmental regularity in all of its components not just for the suggested age window (16-30 months), but also for a few months below (3 months) and above (6 months). As a result, we expected to see numerous floor and ceiling effects. This way we could firmly say that the decision to cut off the age range for the norming study to 16-30 months was dictated by the results of our preliminary study, not just by following verbatim the American CDI. All participants lived in Albania. Data from two participants were excluded from the analysis because their caregivers did not complete some part of the CDI. Additionally, data from one participant was excluded from the analysis due to a very low vocabulary size – more than 1.5 SD below the mean of his/her age group. This yielded a final participant pool of 109 children with an age range from 13 to 36 months (*M* = 26.3, *SD* = 6.2, 55 females and 54 males). All children had normal birth weight, no serious illnesses, and no developmental delays. Children were recruited from five different cities in Albania to ensure a certain degree of generalizability, as this was the first study of this kind: Tirana, Fier, Vlorë, Elbasan, and Pogradec. Note that almost all these cities come from the Tosk-speaking areas, the dialect which also forms the base for the Standard variety. We purposefully did not recruit from the Gheg-speaking areas other than Tirana (which is the capital and where almost half of the population lives). We believe that a separate CDI form is needed for the deep Gheg-speaking areas of northern Albania as well as Kosovo due to the lexical, and sometimes grammatical, differences between the two varieties. Children were also recruited from homes with different parental education levels (see Table 3).

In checking the distribution of our data, we noticed two instances of skewness. First, the majority of our participants fall within the upper age ranges of 25-36 months, as shown in Table 2 and in the density plot in Figure 1. Second, the majority of the parents of our participants have a relatively high level of education - either postsecondary (for mothers) or postsecondary or secondary (for fathers) – as shown in Table 3 and Figure 2. We had very few participants whose parents have completed only primary school. The skewness of education level may have occurred for two reasons. One reason is that, as a former communist state, Albania has a tradition of granting access to higher education to all, making university degrees common among the population. In recent years, the multiplication of private universities has further increased the graduation rate at the Master's level; in fact, Albania is reported to be the country with the highest number of private universities in Southeast Europe. A second reason is that parents with higher education may be more willing to take part in research studies like this – a general trend also noticed in other CDI studies (deMayo et al., 2021).

Thus, given that our sample does not have a normal distribution, and also because of our inability to make evidence-based hypotheses about children's performance on the CDI due to the absence of knowledge about Albanian language acquisition, our study will be exploratory and descriptive in nature. We perform statistical testing solely as an auxiliary tool to explore the description of the trends we observe.

Table 2. ***Number and sex of participants, divided in 3-month intervals***

Age	Female	Male
13-15	4	4
16-18	1	6
19-21	5	4
22-24	6	8
25-27	13	6
28-30	12	13
31-33	6	7
34-36	8	6

Table 3. ***Education level of participants' parents***

Completed Education	Mother	Father
Primary School	6 (5.5%)	6 (5.5%)
Secondary School	29 (26.6%)	48 (44.0%)
Post-Secondary Education (Bachelor's degree or higher)	74 (67.9%)	55 (50.5%)

Data for the study were provided by a primary caregiver of the child, either a parent or grandparent (as noted in Section 2.3, grandparents often live with the family and/or babysit children during the day). Prior to data collection, permission to conduct the study was obtained from the school boards of the corresponding municipalities. Children were recruited in two ways. First, preschool teachers distributed the CDI test to interested parents of the children that were enrolled in their preschools. Second, researchers involved in the study handed out the CDI test to interested relatives, neighbours, friends, etc., who had children corresponding to the ages relevant for the study. In both cases, parents provided written consent for their child's participation in the study. Data were collected over the period 2013-2015 using Version 2 of the Albanian CDI, as detailed above.

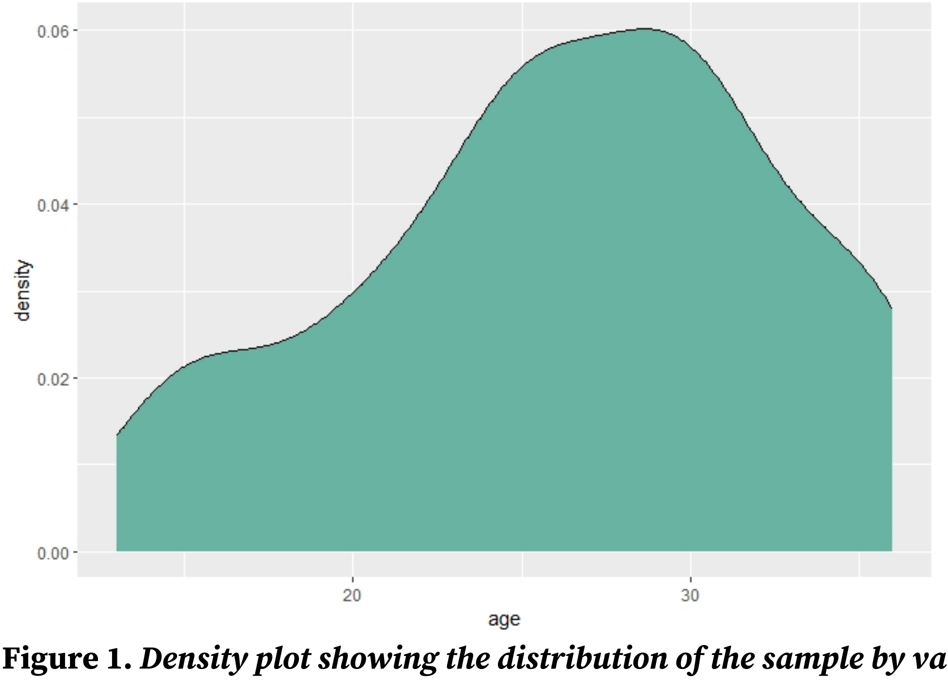


Figure 1. ***Density plot showing the distribution of the sample by variable age in months***

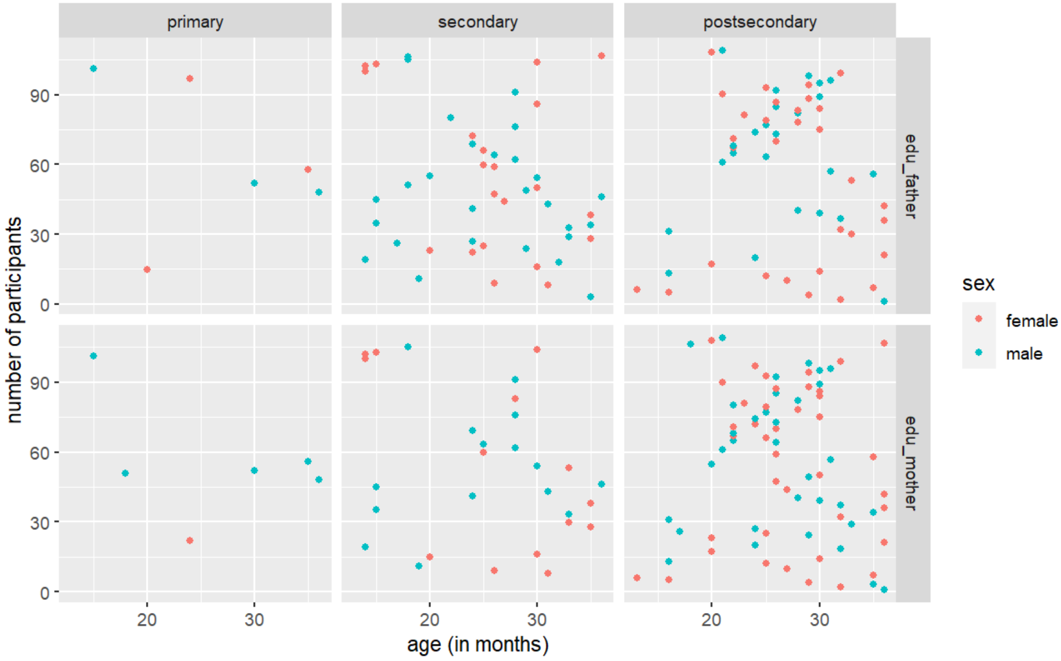


Figure 2. ***Distribution of our participants per demographic variables of sex and level of parental education, i.e. maternal education (edu_mother) and paternal education (edu_father)***

Analysis

For ease of reference in the analysis, we labelled Part I Section A as 'vocabulary size', Part I Section B as 'displaced events', and Part II Section A as 'morphology'. We also divided the words in the vocabulary list from Part I Section A into two categories: open class words (nouns, verbs, adjectives, adverbs) and closed class words (prepositions, exclamations, articles, pronouns). The vocabulary size was calculated as one point per word reported as used by the child, with a maximum score of 650. The scoring for displaced events and morphology was calculated as one point per item that the caregiver reported used either 'sometimes' or 'often'; items reported as used 'never' were scored as 0 points. Thus, the maximum score was 5 for each of these two parts, since each part contained 5 questions.

Given that some of the variables are bound within certain ranges, as was clear from the distribution plots presented above, we used Spearman correlation analyses to explore two types of relationships: a) the relationship between age and vocabulary size, and between age and grammatical development (i.e., score for displaced events + morphology); and b) the relationship between vocabulary size and other aspects of language development, such as number of open class words, number of closed class words, score for displaced events, and score for morphology. Lastly, we explored the possible effects of demographic factors (children's sex, maternal education and paternal education) om language development (vocabulary size, displaced events, morphology) using a beta regression model (Ferrari & Cribari-Neto, 2004). This is a form of regression used when the response variable – for example, total vocabulary size – takes values within (0,1), and is assumed to take a beta distribution. The values of the response variables were beta-transformed as suggested in Smithson & Verkuilen (2006)[2]. Importantly, however, due to the unusual shape of the sample distribution and its bias towards older children and parents with higher educational levels, these analyses are used only as an auxiliary tool in the descriptions of the trends that we observe in the data.

Results

Age and acquisition of vocabulary, displaced events and morphology

Most frequent words

As a first step, a descriptive analysis showed that several semantic categories are included in the most frequently produced words in the Albanian CDI. These include the following (percentage of all children who produced the word is shown in parentheses).

a) kinship terms: *babi* 'father' (97.2%), *mami* 'mother' (90.8%), *teta* 'aunt' (83.5%), *gjyshi* 'grandfather' (79.8%), and the child's own name (69.7%)
b) terms used in the contexts of greeting and parting: *alo* 'answering phone call' (88.9%), *jo* 'no' (84.4%), *po* 'yes' (74.3%)
c) animals and their sounds: *macja* 'cat' (77.9%), *qeni* 'dog' (76.1%), *lopa* 'cow' (73.4%), *pula* 'chicken' (72.5%), *bee* 'baa' (82.6%), *ciu ciu* 'tweet-tweet' (81.7%), *ham ham* 'woof-woof' (84.5%)
d) food items: *buka* 'bread' (80.7%), *banane* 'banana' (77.1 %), *biskota* 'cookies' (74.3%)
e) toys: *topi* 'ball' (77.9%), *lapsi* 'pen' (73.4%), *lodra* 'toys' (73 %)

On the other hand, words that were used less frequently and only by older children (from 21 months old) were as follows: *i/e ngathët* 'clumsy' (26.4%), *i/e përgjumur* 'sleepy' (31,8%), *i/e varfër* 'poor' (33.6%), *pasnesër* 'day after tomorrow' (29.1%), *i/e tyre* 'theirs' (27.3%), *i yni* 'ours' (30.1%), and also two traditional desserts *kadaif* 'type of pastry dessert' (32.7%), *mualebia* 'type of baby food' (31.8%). As we can see, words that are learned latest are mostly adjectives, adverbs, and pronouns. The low frequency counts of words such as *kadaif* and *mualebia* in this study show the fast evolving nature of post-communist Albanian society; these words were deemed as important for this list by two different focus groups in the early phases of the Albanian CDI, but appear

[2] We thank one of the anonymous reviewers for this helpful suggestion which allows a more appropriate analysis for a sample with a distribution such as ours.

to have gone "out of fashion" now. It is very likely that we will exclude them from the norming phase of the CDI test.

Relationship between age and vocabulary

After a first descriptive analysis, we then assessed the relationship between children's age and their vocabulary size. A Spearman's correlation analysis for monotonic relationships between two variables revealed a moderately strong and statistically significant correlation (rho 0.40, $p < 0.01$). As can be seen in Figure 3, the total number of words produced by children increases as a function of age. However, two trends deserve mention here: a) data points are much sparser for children younger than 25 months of age, and b) a large number of children appear to reach ceiling, i.e. produce all the words in the CDI list, around 25 months old.

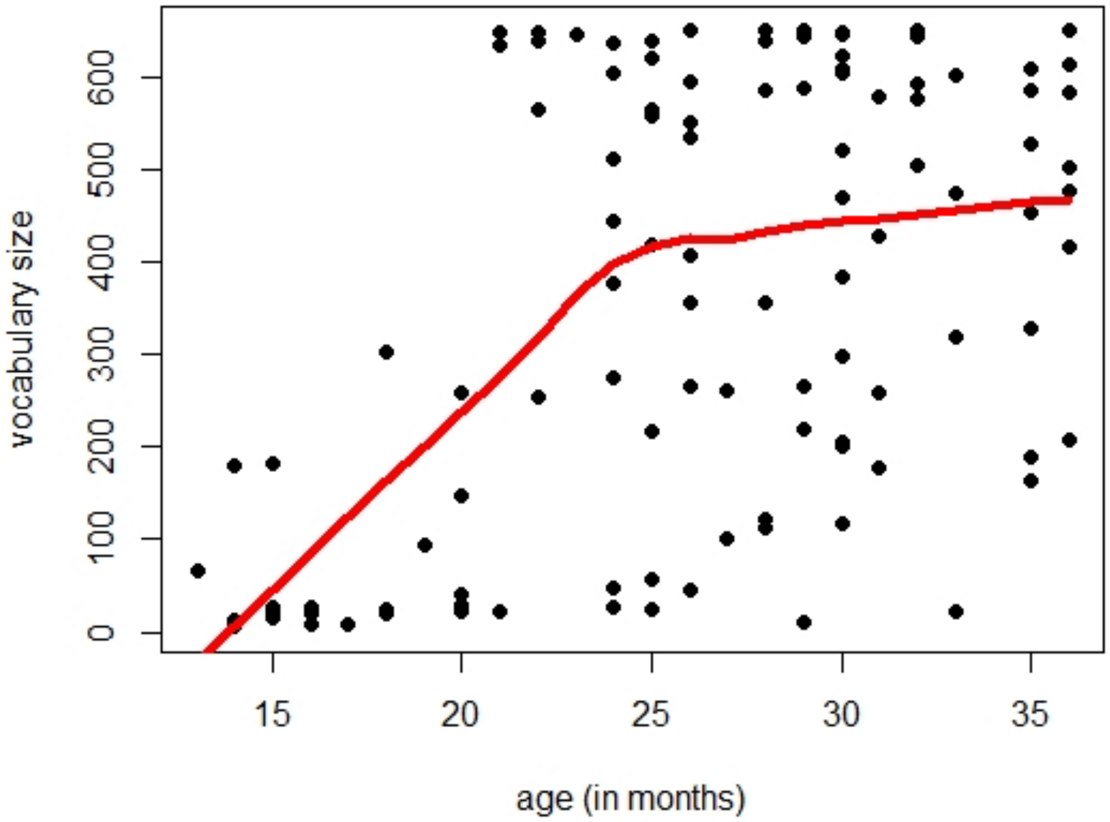


Figure 3. ***Number of words produced as a function of age (in months)***

Relationship between age and displaced events

We also looked at 'displaced events' – at how children performed with regards to whether they a) talk about events that happened in the past, b) talk about events that will happen in the future, c) talk about objects or people that are not present, d) understand requests for objects or people that are not present, and e) point/talk to an object that belongs to a person that is not present. Each participants' score varied depending on the number of displaced events they showed evidence of having conceptualized; for example, a score of 5 means that the participant showed evidence of having conceptualized all five of them. A Spearman's correlation analysis revealed a moderately strong and statistically significant correlation (rho 0.41, $p < 0.01$). Again, we note the high number of participants that are either at ceiling or at floor.

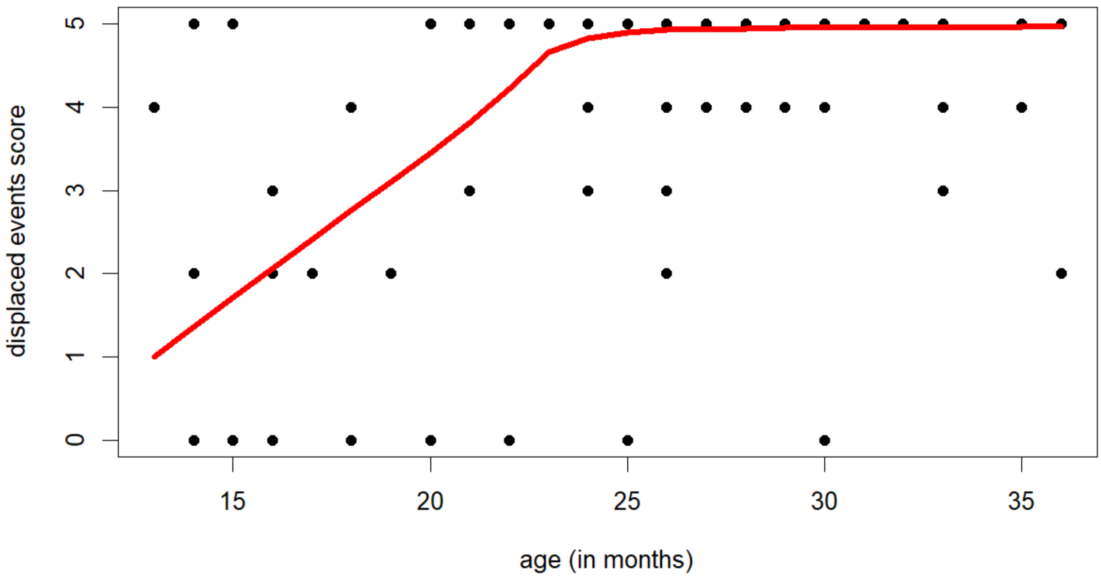


Figure 4. ***Displaced events score produced as a function of age (in months)***

Relationship between age and morphological knowledge

In the same way, we also analysed the development of morphology – whether children produce 1) regular plural forms (*vajzë-vajza* 'girl-girls'); 2) verbs in simple past tense (*hëngra* 'ate'); 3) verbs in present continuous tense (*jam duke ngrënë* '(I) am eating'); 4) verbs in the subjunctive form (*për të ngrënë* 'to eat'); and 5) verbs in future tense (*do të ha* '(I) will eat'). Again here, each participants' score varied depending on the number of morphological forms they showed evidence of having produced; for example, a score of 5 means that the participant showed evidence of having produced all five of them. As Figure 5 shows, children's abilities with these morphological forms increases as they age. A correlation analysis revealed a moderate and statistically significant correlation (rho 0.53, $p < 0.01$). As with displaced events, a high number of participants are either at ceiling or floor in terms of their morphology as assessed in this section of the CDI.

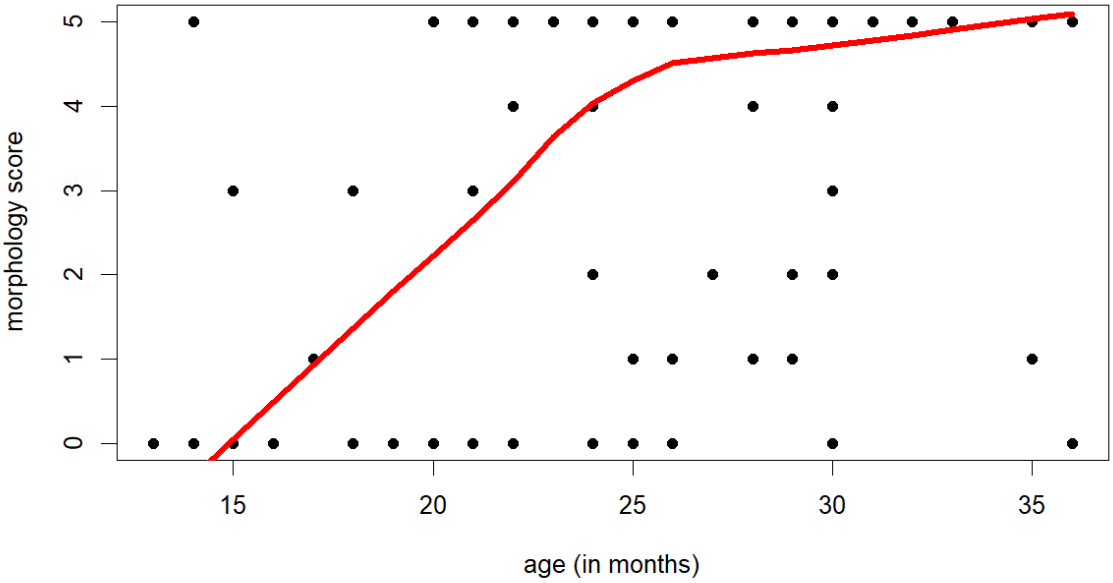


Figure 5. ***Morphology score as a function of age (in months)***

Relationship between vocabulary size and other aspects of language development

Next, we examined the relationship between the overall vocabulary size and production of other language-related phenomena, such as open class words and closed class words, displaced events and morphology. Following Bates et al (1994), Figure 6 portrays developmental trends in vocabulary composition for both open class and closed class words as a function of vocabulary size. Both open class and closed class words increase with age and there is a clear linear relationship between both of them and vocabulary size. Correlation analyses reveal strong correlations for both open class words (rho 0.99, $p < 0.01$) and closed class words (rho 0.97, $p < 0.01$).

We also found a moderately strong significant correlation between vocabulary size and the expression of displaced events (rho 0.57, $p < 0.01$), as well as between vocabulary size and the use of morphology (rho 0.64, $p < 0.01$), using Spearman's correlation analysis for monotonic relationships. As can be seen in Figure 7, the total number of grammatical concepts realized through morphology increases as a function of vocabulary, but there is noticeable variability as shown by many data points in the upper end of the scale, revealing a ceiling effect.

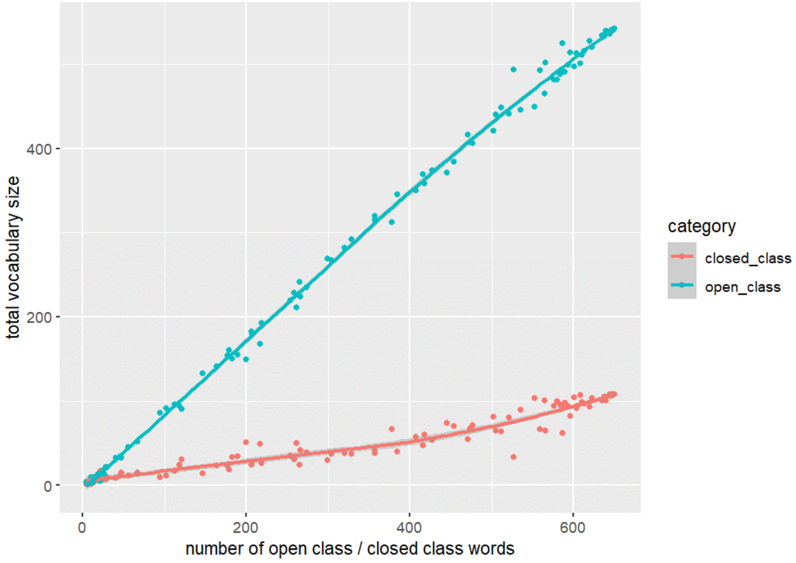


Figure 6. ***Number of open class and closed class words as a function of vocabulary size***

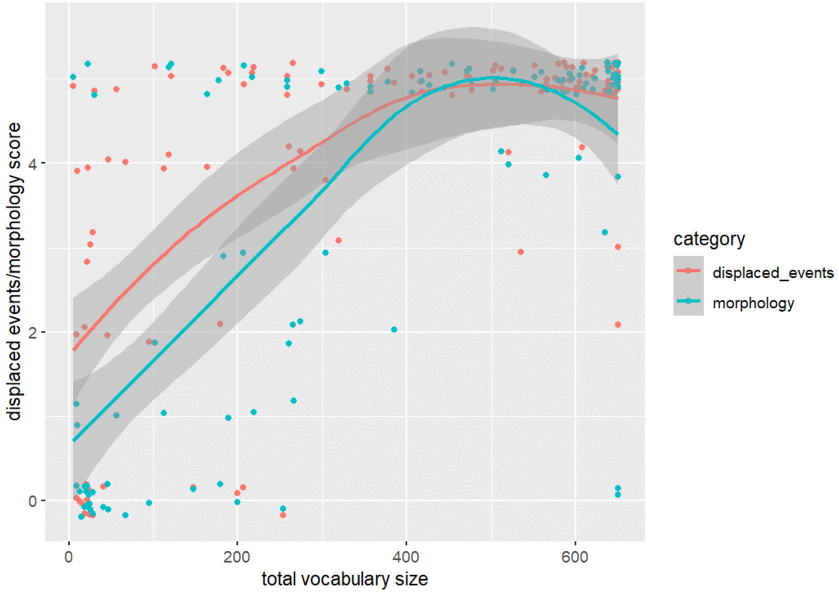


Figure 7. ***Relationship between vocabulary size and frequency of mention of displaced events (blue) and morphological inflections (red)***

The role of demographic variables

Finally, we investigated the effect of three demographic variables – children's sex, maternal education, and paternal education – on vocabulary size, displaced events and morphology. We fitted three beta regression models via the *betareg* function (Ferrari & Cribari-Neto, 2004) in R (R Core Team, 2023) to predict the total number of words, the displaced events score, and the morphology score based on each of the three variables, employing the standard logit link in *betareg*. The data is visualized in Figure 8 for sex, in Figure 9 for maternal education and in Figure 10 for paternal education.

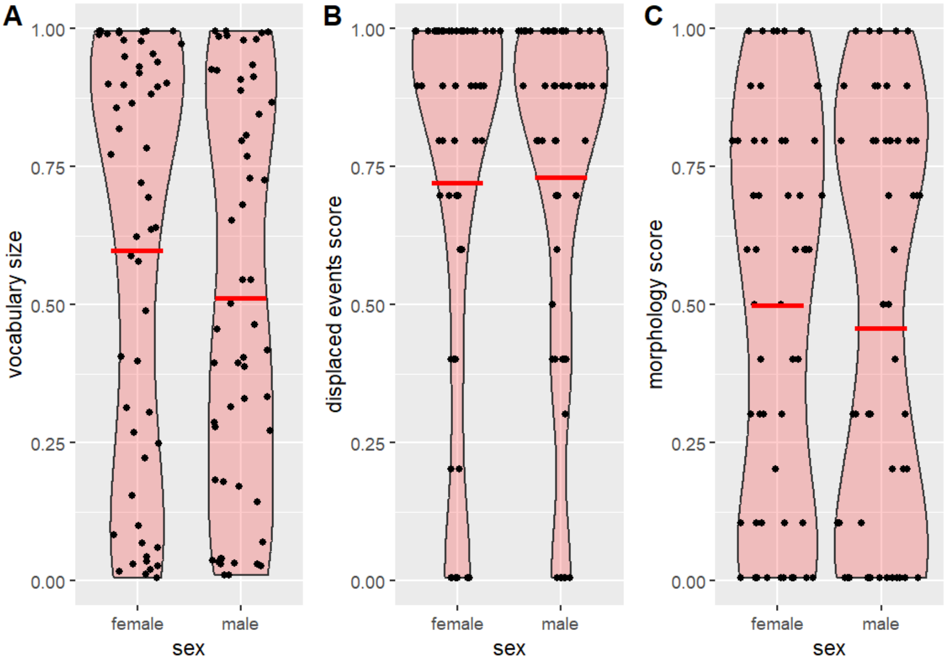


Figure 8. ***The relationship between children's sex and vocabulary size (plot A), displaced events score (plot B), and morphology score (plot C)***

In general, our observations, aided by the beta analysis, showed that none of the three demographic variables tested, i.e. sex, maternal education and paternal education had an effect on the total number of words that children learned over their first three years of life, or on the displaced event score or morphology score. However, an effect of paternal education was found on children's morphology score, as shown in the output of this analysis in Appendix 1. Children of fathers that had postsecondary education degrees were more likely to have higher morphology scores than were children of other fathers. However, these results should be considered with caution given that the distribution of our data is skewed towards older children that have parents with postsecondary degrees.

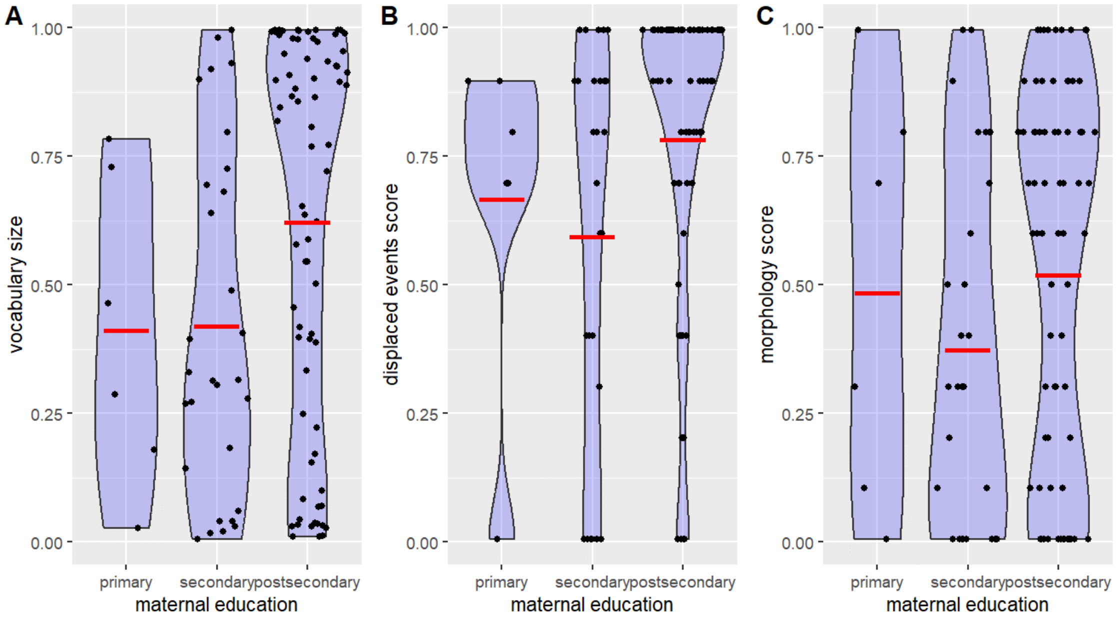


Figure 9. ***The relationship between maternal education and vocabulary size (plot A), displaced event score (plot B), and morphology score (plot C)***

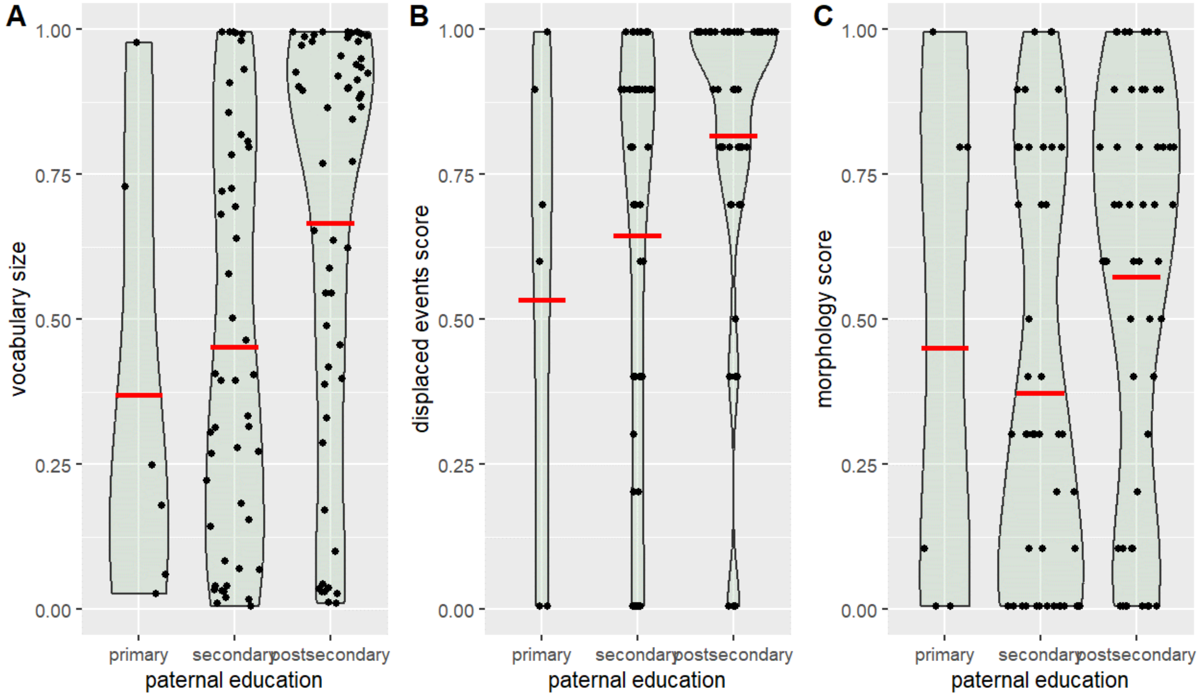


Figure 10. ***The relationship between paternal education and early expressive vocabulary (plot A), displacement score (plot B), and morphology score (plot C)***

Discussion and Conclusion

This study is the first study investigating the characteristics of early vocabulary and grammar acquisition (i.e. displaced events and morphology) in Albanian-speaking infants and toddlers using data from the first adaptation of the Albanian CDI. It serves as a first step in the development of this instrument for Albanian, a largely understudied language of the Indo-European family. In addition, it provides the first description of lexical and grammatical development of a large group of Albanian-speaking children, at least compared to what has been previously reported in the literature. Generally speaking, we observed language development trends similar to other reported languages in that vocabulary size, conceptualization of displaced events, and use of morphology to indicate plurality and tense all increase as children grow older. However, given that our sample distribution was skewed towards older children and children whose parents had postsecondary degrees, as well as to children living in urban areas, our results should be seen as mainly explorative in nature to guide specific hypotheses that will be tested during our forthcoming norming study of the Albanian CDI with a more representative distribution of children.

In measuring early vocabulary growth, we found that the correlation between age and vocabulary size reported here for Albanian appears similar to that reported for other languages at a general level (Frank et al., 2017). However, the clear levelling out in growth around 25 months as illustrated in Figure 3 seems to differ somewhat from other languages, and overall, the Albanian-speaking children seem to master a higher number of words than children in many other languages (see Bleses et al., 2008, p. 641). We believe that two reasons may explain this result. First, the distribution of our sample is largely skewed towards older children, as shown in the density plot in Figure 1, and it is likely that these older children know more words in general than the younger children. Note that 27 of the 109 children are older than 30 months of age, which is the recommended upper age limit for CDI II (the version adapted and tested here). The ceiling effect in these children's performance validates the CDI recommendations for 30 months being the upper age limit for CDI II. Second, our sample is also biased towards children whose parents have postsecondary degrees. This high level of education may indeed lead to higher vocabulary levels, or some other factor might be at play (e.g., these parents may have a tendency to claim that their children are doing well by checking off all the words in the CDI checklist). We plan to explore the relationship between children's vocabulary development and parental education level further by having a wider distribution of parental education in future studies.

Another trend noticeable in our data is that during the 2nd year of life, children's vocabularies increase dramatically. Although the number of words for 24-month-old Albanian-speaking children seems quite high, previous work has suggested that toddlers of the same age across different languages show substantial differences when it comes to the number of words in the expressive vocabulary (e.g Bates et al., 1988;

Fernald et al., 2001). The observed difference between the 19- to 21-month old group and the 22- to 24-month old group might also indicate the occurrence of two stages in the development of vocabulary, thus matching previous findings reported in the literature (e.g Bates & Goodman, 2001; Gendler-Shalev & Dromi, 2022). The stages that we discern here roughly match those reported for English, for instance, for which a second phase of vocabulary development begins around 2;0 to 2;6 (Day & Elison, 2022). For Albanian-speaking children, it seems that it is around age 2;0 that a new phase of vocabulary learning begins. These phases have been linked to neuro-maturational changes that impact the route and the speed with which children can acquire new vocabulary (Gendler-Shalev & Dromi, 2021).

Children's vocabulary patterns revealed that the overall number of open-class words (nouns, verbs, and adjectives) was higher than that of closed-class words (prepositions, pronouns, adverbs, interjections, etc.) from the very start (see Figure 6). However, a closer look at children's first 100 words below the age of 20 months showed only 5 verbs alongside 77 nouns and 13 interjections, 2 adverbs, 2 pronouns, and 1 conjunction. These results generally match the findings reported for other languages (Conboy & Thal, 2006; Day & Elison, 2022; Marjanovič-Umek et al., 2011), which establish that toddlers tend to produce more open-class or content words such as nouns, verbs and adjectives earlier than closed-class words such as prepositions, determiners and pronouns.

A second research question focuses on the relationship between children's vocabulary development and whether they refer to displaced events (things that are absent in space or time) or morphology knowledge use. We found that the use of both of these correlated strongly with vocabulary size. On a first look, it appears that the more words children produced, the more knowledge of morphology they used in their speech. These results resonate with those found in other languages in which infants' and toddlers' grammatical development is closely linked to vocabulary size (e.g Bates et al., 1988, 1994; Bleses et al., 2008; Caselli et al., 1999; Day & Elison, 2022; Devescovi et al., 2005; Jackson-Maldonado et al., 1993; Stolt et al., 2009, 2009). But it is not clear if this relationship between early vocabulary and displaced events/morphology is linear or not. It may also be the case that both the lexicon and displaced events/morphology might actually be developing synchronously in the early years of life, along the lines of the proposal put forth in Dixon and Marchman (2007). Indeed, the scores for displaced events, for example, associate positively with vocabulary size, converging with the idea that these close-timing synchronies can be interpreted as evidence that lexicon is not necessarily learned earlier, but is "part and parcel of the child's transition to grammatical language" (Anisfeld et al., 1998).

Our study also investigated the role of language-external factors in early vocabulary development, specifically sex and parental education. The results revealed no effect of sex in children's early vocabulary, displaced events and morphology, which is not coherent with previous work showing girls outperforming boys at this age (Eriksson et al., 2012; Simonsen et al., 2014). However, another aspect of previous CDI work that has found partial support in our study is the idea that children's early language differs based on their parents' education level (Day & Elison, 2022; Fernald & Marchman, 2012; Hoff & Ribot, 2015). More specifically, we found that paternal educational levels correlated with children's morphology score. The more educated fathers were, the more knowledge of morphology was reported in children's early language. Interestingly, this effect is also reflected in the fact that the word *babi* 'father' is also at the top of Albanian-speaking children's first 100 words, followed by *mami* 'mom' in third place. This role of fathers could be linked to more modern trends in recent years with changes in family structure across the globe and the changing role of men in these structures, wherein more fathers play an active role in children's development (Pancsofar & Vernon-Feagans, 2006). It is not clear, however, that this is the case for Albanian society to date. Studies have also previously shown that in collectivist cultures like those in Asia or South America, fathers indirectly influence children through their effects on the mother-child relationship, providing resources that promote learning and language, and directly through their interactions with children (Tamis-LeMonda et al., 2008). In addition, fathers with more education are typically able to provide more resources and learning opportunities to their children compared to fathers with less education (Cabrera & Peters, 2000). Whether this is true of the Albanian context is still unclear and deserves further investigation, especially given the fact that our sample distribution was somewhat biased with respect to the education levels of caregivers. The patriarchal nature of Albanian society should also be considered in understanding the patterns that we have uncovered here.

A limitation of this study is that it did not fully capture the diverse nature of the Albanian society, with highly educated and urban parents being over-represented in our sample. Future steps should aim to test children that were under-represented in this study, such as those from less urban areas, those brought up by parents that have low educational attainment, as well as those from different dialectal backgrounds in order to better understand lexical and grammatical development of early language in Albanian-speaking children. Procedures for norming of the CDI in the future will be better informed if we have tackled some of these populations beforehand and compared their development to the developmental trajectories reported here.

In sum, this study examined empirical parental report data from the Albanian version of the MacArthur-Bates Communicative Development Inventory (Albanian CDI) concerning the vocabulary and grammatical development of Albanian-speaking children aged 13-36 months. The adaptation reported here covered three sections of the original CDI II Words and Sentences: The vocabulary checklist (Part I Section A), How children use words (Part I Section B), and Word Endings (Part II Section A). The findings outlined here provide insight into patterns of vocabulary growth and the development of communicative competence in Albanian-speaking children from 1 to 3 years old. We have shown that an adaptation of the MB-CDI for Albanian - a yet unstudied language/culture – is able to produce data comparable to those found for other languages using the same instrument, and thus conclude that this adaptation is successful and promising so far. A full adaptation of the entire test is necessary to develop the tool for the assessment of and research on the language development of Albanian-speaking children. Further data collection, instrument improvement, and the full development of Albanian MB-CDI will contribute to shedding more light on the patterns of communicative development in a language and a cultural context that has been largely understudied.

Appendix 1

Table 1. Summary of the beta regression analysis with *morphology* score as dependent variable and *gender, maternal education* and *paternal education* as independent

```
Call:
betareg(formula = morf_beta ~ gender + edma + edfa, data = df2)

Standardized weighted residuals 2:
    Min      1Q  Median      3Q     Max 
-1.7720 -0.8010  0.1090  0.5689  1.9643 

Coefficients (mean model with logit link):
                Estimate Std. Error z value Pr(>|z|)  
(Intercept)      0.19306    0.20576   0.938   0.3481  
gendermale      -0.14089    0.25023  -0.563   0.5734  
edmaelementary   0.38062    0.63029   0.604   0.5459  
edmahigh school -0.08982    0.31714  -0.283   0.7770  
edfaelementary  -0.59109    0.62979  -0.939   0.3480  
edfahigh school -0.61218    0.29280  -2.091   0.0365 *

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)    
(phi)   0.9365     0.1007   9.302   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Type of estimator: ML (maximum likelihood)
Log-likelihood: 47.42 on 7 Df
Pseudo R-squared: 0.07335
Number of iterations: 13 (BFGS) + 2 (Fisher scoring)
```

References

Agalliu, F., Demiraj, S., Domi, M., & Instituti i Gjuhësisë dhe i Letërsisë (Akademia e Shkencave e RSH) (Eds.). (2002). *Gramatika e gjuhës shqipe*. Botimi i Akademisë së Shkencave.

Anisfeld, M., Rosenberg, E. S., Hoberman, M. J., & Gasparini, D. (1998). Lexical acceleration coincides with the onset of combinatorial speech. *First Language, 18*(53), 165–184. https://doi.org/10.1177/014272379801805303

Bates, E., Bretherton, I., & Snyder, L. S. (1988). *From first words to grammar: Individual differences and dissociable mechanisms.* Cambridge University Press.

Bates, E., & Goodman, J. C. (2001). On the inseparability of grammar and the lexicon: Evidence from acquisition. In *Language development: The essential readings.* (pp. 134–162). Blackwell Publishing.

Bates, E., Marchman, V., Thal, D., Fenson, L., Dale, P., Reznick, J. S., Reilly, J., & Hartung, J. (1994). Developmental and stylistic variation in the composition of early vocabulary. *Journal of Child Language, 21*(1), 85–123. https://doi.org/10.1017/S0305000900008680

Bleses, D., Vach, W., Slott, M., Wehberg, S., Thomsen, P., Madsen, T. O., & Basbøll, H. (2008). Early vocabulary development in Danish and other languages: A CDI-based comparison. *Journal of Child Language, 35*(3), 619–650. https://doi.org/10.1017/S0305000908008714

Bopp, F. (1855). *Über das Albanesische in seinen verwandtschaftlichen Beziehungen.* Dümmler.

Bouchard, C., Trudeau, N., Sutton, A., Boudreault, M.-C., & Deneault, J. (2009). Gender differences in language development in French Canadian children between 8 and 30 months of age. *Applied Psycholinguistics, 30*(4), 685–707. https://doi.org/10.1017/S0142716409990075

Çabej, E. (1976). *Studime gjuhësore, III.* Academy of Sciences.Tirana.

Cabrera, N., & Peters, H. E. (2000). Public Policies and Father Involvement. *Marriage & Family Review, 29*(4), 295–314. https://doi.org/10.1300/J002v29n04_04

Caselli, C., Casadio, P., & Bates, E. (1999). A comparison of the transition from first words to grammar in English and Italian. *Journal of Child Language, 26*(1), 69–111. https://doi.org/10.1017/S0305000998003687

Cenko, E. (2017). The Early Acquisition of Verb Constructions in Albanian: Evidence from Children's Verb Use in Experimental Contexts. *Academic Journal of Interdisciplinary Studies, 6*(1), 87–96. https://doi.org/10.5901/ajis.2017.v6n1p87

Cenko, E., & Budwig, N. (2007). The acquisition of early verb constructions in Albanian: A first look at transitives and intransitives. *A Supplement to the Proceedings of the 31st Boston University Conference on Language, 31.*

Cipo, K. (1949). *Gramatika e gjuhës shqipe*. Instituti i Shkencave. Tirana.

Conboy, B. T., & Thal, D. J. (2006). Ties Between the Lexicon and Grammar: Cross-Sectional and Longitudinal Studies of Bilingual Toddlers. *Child Development, 77*(3), 712–735. https://doi.org/10.1111/j.1467-8624.2006.00899.x

Croft, W. (2002). *Typology and Universals* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511840579

Day, T. K. M., & Elison, J. T. (2022). A broadened estimate of syntactic and lexical ability from the MB-CDI. *Journal of Child Language, 49*(3), 615–632. https://doi.org/10.1017/S0305000921000283

deMayo, B. E., Kellier, D., Braginsky, M., Bergmann, C., Hendriks, C., Rowland, C., Frank, M., & Marchman, V. (2021). *Web-CDI: A system for online administration of the MacArthur-Bates Communicative Development Inventories.* https://doi.org/10.34842/KR8E-W591

Demiraj, B. (2018). The evolution of Albanian. In J. Klein, B. Joseph, & M. Fritz (Eds.), *Handbook of Comparative and Historical Indo-European Linguistics* (pp. 1812–1815). De Gruyter. https://doi.org/10.1515/9783110542431-021

Desnickaja, A. V. (1976). Çështje të dialektologjisë historike të gjuhës shqipe. *Jehona, 4,* 305–310.

Devescovi, A., Caselli, M. C., Marchione, D., Pasqualetti, P., Reilly, J., & Bates, E. (2005). A crosslinguistic study of the relationship between grammar and lexical development. *Journal of Child Language, 32*(4), 759–786. https://doi.org/10.1017/S0305000905007105

Dixon, J. A., & Marchman, V. A. (2007). Grammar and the Lexicon: Developmental Ordering in Language Acquisition. *Child Development, 78*(1), 190–212. https://doi.org/10.1111/j.1467-8624.2007.00992.x

Dule, X. S. (2023). MA Thesis. University of Kaiserslautern.Germany.

Eriksson, M., Marschik, P. B., Tulviste, T., Almgren, M., Pérez Pereira, M., Wehberg, S., Marjanovič-Umek, L., Gayraud, F., Kovacevic, M., & Gallego, C. (2012). Differences between girls and boys in emerging language skills: Evidence from 10 language communities. *British Journal of Developmental Psychology, 30*(2), 326–343. https://doi.org/10.1111/j.2044-835X.2011.02042.x

Fenson, L., Marchman, V., Thal, D., Dale, P., Reznick, J. S., & Bates, E. (2007). *Mac-Arthur-Bates Communicative Development Inventories (CDI) Words and Sentences.* Brookes Publishing.

Fernald, A., & Marchman, V. A. (2012). Individual Differences in Lexical Processing at 18 Months Predict Vocabulary Growth in Typically Developing and Late-Talking Toddlers: Lexical Processing and Vocabulary Growth. *Child Development, 83*(1), 203–222. https://doi.org/10.1111/j.1467-8624.2011.01692.x

Fernald, A., Swingley, D., & Pinto, J. P. (2001). When Half a Word Is Enough: Infants Can Recognize Spoken Words Using Partial Phonetic Information. *Child Development, 72*(4), 1003–1015. https://doi.org/10.1111/1467-8624.00331

Ferrari, S., & Cribari-Neto, F. (2004). Beta Regression for Modelling Rates and Proportions. *Journal of Applied Statistics, 31*(7), 799–815. https://doi.org/10.1080/0266476042000214501

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language, 44*(3), 677–694. https://doi.org/10.1017/S0305000916000209

Gendler-Shalev, H., & Dromi, E. (2022). The Hebrew Web Communicative Development Inventory (MB-CDI): Lexical Development Growth Curves. *Journal of Child Language, 49*(3), 486–502. https://doi.org/10.1017/S0305000921000179

Gjinari, J. (1988). *Dialektologjia shqiptare* (revised and expanded). Akademia e Shkencave e Shqipërisë.Tirana.

Hahn, J. G. (2013). *Studime shqiptare. Origjinali gjermanisht: Albanesiche Studien. Wien: Friedrich Mauke, 1854.* Akademia e Shkencave e Shqipërisë.Tirana.

Hoff, E., & Ribot, K. M. (2015). Language Development: Influence of Socio-Economic Status. In *International Encyclopedia of the Social & Behavioral Sciences* (pp. 324–328). Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.23132-2

Hulle, C. A. V., Goldsmith, H. H., & Lemery, K. S. (2004). Genetic, Environmental, and Gender Effects on Individual Differences in Toddler Expressive Language. *Journal of Speech, Language, and Hearing Research, 47*(4), 904–912. https://doi.org/10.1044/1092-4388(2004/067)

INSTAT. (2017). *Burrat dhe gratë në Shqipëri—Men and Women in Albania*. Albanian National Institute of Statistics. http://www.instat.gov.al/media/2316/burrat_dhe_grat__ne_shqiperi_2017_libri.pdf

Ismajli, R. (1998). *"Në gjuhë" dhe "për gjuhën" (rrjedha të planifikimit të shqipes në Kosovë (1945-1968)*. Dukagjini.

Jackson-Maldonado, D., Thal, D., Marchman, V., Bates, E., & Gutierrez-Clellen, V. (1993). Early lexical development in Spanish-speaking infants and toddlers. *Journal of Child Language, 20*(3), 523–549. https://doi.org/10.1017/S0305000900008461

Jia, X. (2023). *Early lexical acquisition of Albanian: An online: CDI-based study*. MA Thesis. University Ludwig-Maximilian, Germany.

Kapia, E. (2010). *The Role of Syntax and Pragmatics in the Structure and Acquisition ofClitic Doubling in Albanian.*PhD Dissertation. Boston University. Boston, MA, USA.

Kapia, E. (2014). Acquisition of Dative and Accusative Clitic Doubling in Albanian: A Syntactic-Pragmatic Approach. In *Developments in the Acquisition of Clitics*. Cambridge Scholars Publishing.

Karmiloff, K., & Karmiloff-Smith, A. (2001). *Pathways to language: From fetus to adolescent.* (pp. ix, 256). Harvard University Press.

Klein, J. S., Joseph, B. D., & Fritz, M. (Eds.). (2017). *Handbook of comparative and historical Indo-European linguistics*. De Gruyter Mouton.

Lehmann, C. (1982). Directions For Interlinear Morphemic Translations. *Folia Linguistica, 16*(1–4). https://doi.org/10.1515/flin.1982.16.1-4.199

Marchman, V. A., & Bates, E. (1994). Continuity in lexical and morphological development: A test of the critical mass hypothesis. *Journal of Child Language, 21*(2), 339–366. https://doi.org/10.1017/S0305000900009302

Marjanovič-Umek, L., Fekonja, U., Podlesek, A., & Kranjc, S. (2011). Assessing toddler language competence: Agreement of parents' and preschool teachers' assessments. *European Early Childhood Education Research Journal, 19*(1), 21–43. https://doi.org/10.1080/1350293X.2011.548957

Morgan, J. L., Shi, R., & Allopenna, P. (1996). Perceptual bases of rudimentary grammatical categories: Toward a broader conceptualization of bootstrapping. In *Signal to syntax: Bootstrapping from speech to grammar in early acquisition.* (pp. 263–283). Lawrence Erlbaum Associates, Inc.

Pancsofar, N., & Vernon-Feagans, L. (2006). Mother and father language input to young children: Contributions to later language development. *Journal of Applied Developmental Psychology, 27*(6), 571–587. https://doi.org/10.1016/j.appdev.2006.08.003
Pedersen, H. (1897). Die albanesichen l-Laute. *KZ, 33,* 535–551.

R Core Team. (2023). *R: A language and environment for statistical computing* [Computer software]. https://www.R-project.org.

Rusakov, A. (2017). Albanian. In M. Kapović (Ed.), *The Indo-European Languages* (2nd ed., pp. 552–608). Routledge.

Rushi, T. (1983). Rreth rendit të gjymtyrëve në fjalinë dëftore. [De l'ordre des termes dans la phrase énonciative]. *Studime Filologjike, 1,* 71–85.

Sehitaj, G. (2015). *Zhvillimi i fjalorit dhe i gramatikës tek fëmijet shqipfolës nga mosha 1-3 vjeç.* MA Thesis. Akademia e Studimeve Albanologjike, Insituti i Gjuhësisë dhe Letërsisë.

Shashaj, A. (1996). *Të folurit e fëmijëve të moshës parashkollore dhe puna për zhvillimin e tij.* Shtëpia botuese Onufri.

Simonsen, H. G., Kristoffersen, K. E., Bleses, D., Wehberg, S., & Jørgensen, R. N. (2014). The Norwegian Communicative Development Inventories: Reliability, main developmental trends and gender differences. *First Language, 34*(1), 3–23. https://doi.org/10.1177/0142723713510997

Smithson, M., & Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods, 11*(1), 54–71. https://doi.org/10.1037/1082-989X.11.1.54

Stolt, S., Haataja, L., Lapinleimu, H., & Lehtonen, L. (2009). Associations between lexicon and grammar at the end of the second year in Finnish children. *Journal of Child Language, 36*(4), 779–806. https://doi.org/10.1017/S0305000908009161

Tamis-LeMonda, C. S., Adolph, K. E., Lobo, S. A., Karasik, L. B., Ishak, S., & Dimitropoulou, K. A. (2008). When infants take mothers' advice: 18-month-olds integrate perceptual and social information to guide motor action. *Developmental Psychology,* *44*(3), 734–746. https://doi.org/10.1037/0012-1649.44.3.734

Thordardottir, E. T., Weismer, S. E., & Evans, J. L. (2002). Continuity in lexical and morphological development in Icelandic and English-speaking 2-year-olds. *First Language, 22*(1), 3–28. https://doi.org/10.1177/014272370202206401

World Bank. (2018). *Education Statistics: Country at a Glance Albania.* https://datatopics.worldbank.org/education/country/albania

Young, E. (2018). The Borgen Project. *10 Facts about Girls' Education in Albania.* https://borgenproject.org/tag/girls-education-in-albania/

Zogaj, D. (2021). *Design and analysis of the MacArthur Bates Communicative Development Inventory for Albanian.* MA Thesis. University of Kaiserslautern, Germany.

Data, Code and Materials Availability Statement

Data, code, and materials are available at https://osf.io/9fzyt/

Ethics statement

Ethics approval was obtained from the ethics committee of the Academy of Albanological Sciences in Tirana, Albania. Caregivers of the children who participated in this study gave informed written consent for their participation.

Authorship and Contributorship Statement

All authors approved the final version of the manuscript and agree to be accountable for all aspects of the work in ensuring the accuracy or integrity of any part of the work presented here. EK conceived of the study. EK designed the experiment. EK, SA, DZ performed the analyses. EK and SA wrote the manuscript.

Acknowledgements

The first author was supported by a DFG grant (project number 499336853) for this study. We would like to thank Gjyste Sehitaj who did prior work in her Master's thesis leading up to the present paper. This study has also benefitted tremendously from the comments and feedback of two anonymous reviewers and the LDR statistics reviewer, as well as statistics discussions with Stefano Coretta at University of Edinburgh and Michele Gubian at the Institute of Phonetics and Speech Processing at LMU Munich. In addition, we are grateful to all the parents that filled out the questionnaires and participated in this study without any financial compensation.

License

Is the effect of gross motor development on vocabulary size mediated by language-promoting interactions?

Sivan Bar Or
Naomi Havron
University of Haifa, Israel

Abstract: Previous research suggests there is a positive correlation between infants' motor and language development. Several reasons for this effect have been suggested, but little empirical research directly addressed them. Here we tested the hypothesis that motor development is related to an increase in language-promoting interactions with parents (such as naming objects that the infant is interested in), and that these activities are related to language development. 93 Israeli parents filled in questionnaires about their 8- to-18-month-old infants' language and motor development, as well as about their engagement in language-promoting interactions. Contrary to previous research, we found no evidence that motor development was related to language development, and only partial evidence that motor development was related to language-promoting interactions, and language-promoting interactions to vocabulary size. Possible reasons are discussed.

Keywords: motor development; language development; social interactions

Corresponding author(s): Naomi Havron, School of Psychological Science & The Center for the Study of Child Development, University of Haifa, Abba Khoushy Ave 199, Haifa, Israel. Email: nhavron@psy.haifa.ac.il.

ORCID ID(s): https://orcid.org/0000-0001-6429-1546

Citation: Bar Or, S. & Havron, N. (2024). Is the effect of gross motor development on vocabulary size mediated by language-promoting interactions? *Language Development Research, 4*(1), 207–232. https://doi.org/10.34842/swwf-e586

Introduction

During their first years of life, infants acquire motor skills that significantly alter the way their body moves, affects their environment, and is affected by it. New motor skills change the way infants interact with objects and people around them. Honing these motor skills includes different physical aspects such as manual dexterity and changes in posture and mobility, and allows infants opportunities to act on the world around them actively and proactively (Iverson, 2010, 2021).

Do infants who develop motor skills faster than others also develop faster linguistically? Some empirical evidence suggests that they do. Two recent systematic reviews found an overall positive correlation between motor and language skills (Gonzalez et al., 2019; Leonard & Hill, 2014). These findings have three general explanations, which are not mutually exclusive. We survey each before focusing on the third of these explanations - that new motor abilities change infants' relationship with their environment in a cascading manner (Iverson, 2010, 2021) - the explanation we explore in depth in the current study.

The shared-resources explanation

The first explanation is that motor and language skills share the same set of resources. When infants are engaged with acquiring a novel motor skill, such as learning to stand upright, they produce less vocalizations, being wholly absorbed in acquiring this new skill (Berger et al., 2017). Boudreau and Bushnell (2000) found an interference effect between motor and cognitive activities (and between cognitive and motor activities) in 12 months old infants, not only during transitional periods of mastering a new skill. They called this the attention-driven cognition/action trade-off. Overall, it seems that interference between motor and cognitive performance occurs when the demand for the second task is greater than the system's resource capacity (Abou Khalil et al., 2020). It stands to reason that infants with larger system resource capacity (e.g., larger attentional or cognitive capacities) would be better in both language learning and motor performance - which could explain why most previous studies do find a relation between motor and language skills. However, while there is a relationship between the onset of walking and vocabulary comprehension, this relationship becomes weaker after two weeks of walking experience (Walle, 2016; Walle & Campos, 2014). This seems at odds with a joint attentional resources view, as children with larger cognitive/attentional capacities should excel in both sets of skills and there would be no reason for this relationship to weaken.

The direct-physical-link explanation

The second general explanation is that motor development affects articulation through a direct, physical, route. The physical route is manifested in several motor millstones. Infants' ability to hold objects and bring them to their mouth is an effective way to explore vocal production. It was found that when 11- to-13-month-olds play, they produce sounds whose quality changes in relation to the size of objects they explore. When infants play with larger objects, they open their fingers wider, and also open their mouth wider, changing their vocal productions (Bernardis et al., 2008). Posture is another factor which affects the physical aspect of speech production. A vertical position of the head during upright sitting, for example, changes the way the vertebra and vocal cords are aligned, and the tongue is pushed towards the front of the mouth, making it easier to produce syllables (Yingling, 1981).

Another line of evidence that physical factors affect language development directly is evidence that children with articulation delay are prone to also show motor delays (Gaines & Missiuna, 2007), and infants with atypical motor development, such as cerebral palsy and preterm infants, tend to also show language delays (Ross et al., 2018). There is also a great amount of evidence for both language and motor delays in autistic children (Iverson et al., 2019; West et al., 2019), however, Autistic Spectrum Condition (ASC) is a pervasive condition affecting many areas, and could also be related to the cascading relation with the infant's environment (Iverson et al., 2022), so it will be discussed in detail below.

In surveying the physical route, we assume that its effect on language production will be stronger than its effect on language comprehension, since articulation will be most directly affected. However, strong ties and correlations exist between language production and comprehension. Even simple articulation processes may affect language comprehension. For example, research has shown that blocking the ability to temporarily produce certain phonemes during learning in infancy is related to the ability to discriminate between these phonemes (Bruderer et al., 2015; Choi et al., 2019). Nonetheless, since the effect of motor development on language production is more direct than its effect on comprehension, we suggest that larger gains in production than in comprehension following motor milestones would support the physical route, while larger or equal gains in comprehension will support the interactional route (detailed below) as well. This is because production involves a physical activity (articulation) which is either affected by physical abilities or even defined in itself as a fine motor skill.

Walking infants do show larger vocabularies in both production and comprehension than crawling infants of the same age (He et al., 2015; Walle & Campos, 2014; but see Moore et al., 2019 who do not find a relationship between the onset of walking and vocabulary size). Even when infants who did not yet walk independently were placed in walkers, they still produced less sounds and gestures, including pointing and capturing the mother's attention, than infants who could walk independently (Clearfield, 2011). This could mean that it was more than posture which affected speech and communication. Rather, infants' experience with the world as independent walkers might have affected their language and communication skills.

The cascading-effects explanation: changes in the relationship with parents and with the environment

The third general explanation is that by changing infants' relationship with the world around them, and especially with their caregivers, motor development causes a cascading effect of increasing language promoting activities and interactions (Iverson, 2010, 2021). We refer to this route as the interactional route, a route that is at the heart of the current study. We first describe motor skills' effect on infants' interaction with objects before describing social interactions.

Exploring objects by themselves allows infants, for example, to connect an object with its use and meaning (for example, when they put beads in a canister). Toddlers' manipulation of objects affords them a better view of the objects they are exploring than when parents manipulate the same object. Infants' view is more diverse and captures higher-quality object views than parents' view, and when neural networks were trained on child-generated data, they achieved better performance than when trained on adult-generated data (Bambach et al., 2018). In another study, Slone et al., (2019) found that infants who generated such object views through object manipulation at 15 months of age experienced greater vocabulary growth over the next six months. Moreover, infants attribute meaning to objects when they engage in recognition gestures (such as holding a phone to their ear). It has been suggested that naming an object using a word or a gesture begins with the motor act of using an object, such as a phone in the example above (Bates et. al.,1979). Such gestures are easier to perform when one can sit upright or stand, or move in space by crawling or walking to reach toys of interest.

On the social side, motor development also affords infants more opportunities to engage in language-promoting interactions with their caretakers (Iverson, 2010, 2021; West et al., 2019). For example, sitting without support allows a wider and more flexible field of vision (Iverson, 2010), which might increase the chances of creating eye contact and shared attention with parents. In addition, walking infants can pick up an object and bring it to their caretaker, creating joint interest and attention around an object that is especially attractive to the infant at that moment. Indeed, caregiver utterances contain more labels during infant object manipulation, and these labels frequently corresponded to the infants' held object and their gaze (West & Iverson, 2017).

Walle (2016) found that infant initiation of joint engagement such as bringing objects to the parent, as well as following of the parent's joint engagement cues such as their gaze, increased as a function of infant walking experience. Parents might also talk more to walking than crawling infants (Karasik et al., 2011; Schneider & Iverson, 2022), and tend to use verbs that correspond with the action the infant is engaged in (e.g., describing the action; West et al., 2022). West et al. (2023) found that while they were walking, 13- and 18-month-old infants received triple the rate of locomotor verbs compared to when they were stationary.

While the overall amount of speech directed to children (and possibly also overheard by them, see Akhtar, 2005, but cf Shneidman & Goldin-Meadow, 2012) is seen as extremely important for language development (Hoff & Naigles, 2002; Weisleder & Fernald, 2013), quality child-directed speech is seen as even more useful to their language development. Parents' congruent and thoughtful engagement with infants is thus a major contributor to their cognitive and linguistic development. Previous prospective research found that children whose mothers were more responsive during the first few years of life achieved language-development milestones earlier than those with less responsive mothers (Tamis-LeMonda et al., 2001; Paavola et al., 2006). Incidentally, Tamis-LeMonda et al. (2001) found stronger relations between responsiveness at 13 months and language milestones, then between responsiveness at 9 months and the same milestones - coinciding with the age at which most children begin to walk. Thus, in typically developing children, if motor development promoted such quality, responsive, and adapted child-directed speech - then it stands to reason that motor development supports language development indirectly through increasing adapted and useful linguistic interactions.

Such a cascading effect for motor development on language development has also been shown in children with ASC (West et al., 2019). ASC manifests itself in (among other things) qualitative impairments in communication including a delay in or total lack of development of spoken language. It also includes social atypicalities such as a lack of spontaneous seeking to share enjoyment, interests, or achievements with other people, or a lack of social or emotional reciprocity (Hodges et al., 2020) - the same processes thought to link motor and language development. Motor challenges in ASC are also very common, with up to 87% of the autistic population affected (Zampella et al., 2021). Given that all three of these fields (the social, the motor, and the language fields) are atypical in the autistic population, it is important to also examine whether a cascading effect can be directly viewed, rather than only through correlations between these three fields. Calabretta et al., (2022) tested links between infants' walking and parental responsiveness in typically developing children and siblings to autistic children - that were later diagnosed with, or not diagnosed with, ASC themselves. They found that out of all the infants' in the sample, infants' moving bids (infants' sharing with their caregivers of objects they carry from a distance, by approaching them and using gestures to show or offer their discoveries) were related to highly elevated parental responding with language. However, parents of siblings later diagnosed as autistic were more likely to respond when their infants simply approached them (with or without an object in hand). This particular finding demonstrates that cascading effects between motor abilities, proactive eliciting of language-promoting interactions by the infant, parental responsiveness and language outcomes in autistic infants are nuanced and merit further investigation. As motor-language cascades in ASC are not the focus of the current study, we refer interested readers to Iverson (2018) for a review of additional studies on the subject.

The current study

While we find the explanation of these cascading effects linking motor to language skills compelling, there are very few studies examining the mediating effect of language-promoting interactions in the relation between motor and language development in typically developing children (though see Walle, 2016; as well as West et al., 2019 who test an ASC and a typically-developing comparison group). Generally speaking, there is strong evidence that motor development promotes language-promoting interactions (e.g., Schneider & Iverson, 2022; Walle, 2016), and that language-promoting interactions are related to language development (e.g., Hirotani, et al., 2009), but less evidence that these interactions with infants mediate the relationship between motor and language development. In the current study we wanted to test this hypothesis using parental reports of motor development, language-promoting interactions, and language development in 8- to 18-month-olds. We chose this age range because it is a time of rapid development in both motor and language areas. In addition, the widely used vocabulary parental-report questionnaire, the Macarthur-Bates Communicative Development Inventory (Fenson et al., 1994), which was also used in the current study, only starts at 8 months of age. In terms of motor development, according to the Alberta Infant Motor Scale-AIMS (Darrah et al., 1998) norms, 90% of infants will have achieved unsupported sitting by 8 months, independent standing by 13 months, and independent walking by 14 months. Thus, we expected to find large variability in our sample in both domains.

There was some challenge with operationalizing the concept of language-promoting interactions. Interactions can either be initiated or led by the infant, or they can be initiated, led or controlled by the parent, as can be seen from the different examples above. Thus, parents could be compelled by the infant's motor abilities to behave in a certain, language-promoting way towards them, but motor development might also drive the infant's own behavior regardless of the parent. We generally hypothesized that motor development will be positively related to language development, as was previously found. We expected motor development to also be related to language-promoting interactions (which include both child-initiated and parent-initiated interactions). We expected, like previously found in the literature, these same language-promoting interactions to be related to language development. Last, we expected language-promoting interactions to mediate the effect of motor development on language development, as was notably suggested by Iverson (2010, 2021), as well as others (e.g., Walle, 2016).

Method

Participants

143 parents filled in at least some of the online questionnaires. Of these,102 filled in all questionnaires. Of these we excluded 9 infants: 5 were bilingual (over 25% exposure to a second language reported in the CDI demographic questions, based on the criterion in Frank et al. (2020), 3 were born preterm (more than 4 weeks early according to the CDI questionnaire demographic section - one of these infant was not reported as being premature in the CDI questionnaire, but was reported in our demographic questionnaire as being born on the 30th week of gestation and was therefore excluded), and 1 was below 8 months of age. Another exclusion criterion was parental reports of developmental concerns (these were screened for content, such that reports of non-serious issues - early treated torticollis, for example - could still be included). No parent of the included sample reported serious concerns. The mean age of these infants was 12.42 months (SD 3.24 months, 42% girls). Since we relied on norms for the the Hebrew Web Communicative Development Inventory, and norms were not available for 8-month olds, we removed these children from the analyses of their production, but not comprehension (see below in the Measures section for justification). The comprehension analyses thus included 93 infants, while the production analyses only included 81 infants, who were, on average, older (13.025, SD = 2.868, 38% girls). Out of these 93 infants, at the time of the study, parents reported that 84 were already crawling, 79 were standing unsupported, 75 were sitting unsupported, and 36 were already walking (see Table 1).

Design and Procedure

Parents were recruited online through social media. They filled in five online questionnaires: the gross-motor development subsection of the Ages and Stages Questionnaire (ASQ, to measure motor development), the Hebrew adaptation of the Communicated Development Inventory (CDI, to measure language development), a language-promoting interactions questionnaire developed in our lab, the StimQ home cognitive environment questionnaire (Availability of Learning Materials and Reading subscales), and a demographic details questionnaire. Parents signed online consent forms and the study was approved by the University of Haifa's IRB. The study was not preregistered but the data and analyses scripts, as well as the measure we developed are available on the OSF https://osf.io/hrmp6/?view_only=1248633dd1e943babdf316e4ed205191.

Measures

Ages and Stages (ASQ, Squires et al., 1997). This tool includes 21 separate questionnaires for 2- to-66-month-olds. Each questionnaire contains 30 items querying about five different areas of development: communication, gross motor, fine motor, problem solving, everyday activities and personal-social development. For each item the parent marks whether the infant performs this activity (10 points for "yes", 5 for "inconsistently" and 0 for "not yet"). We used only the gross-motor-skills subset of the instrument, and the forms for 8- to-18-month-olds. Overall, the ASQ has a re-test reliability of .94, and a high correlation (r = .88) with the Bayley Scales of Infant Development (Squires et al,1997). We translated the ASQ relevant forms to Hebrew, and back to English to ascertain the quality of translation before administering them.

The Hebrew Web Communicative Development Inventory - MB-CDI (Maital et al., 2002; Gendler-Shalev & Dromi, 2021). This is a Hebrew adaptation of the English CDI parental questionnaire (Fenson et al., 1994). It was recently adapted for Hebrew, validated, and normed by Gendler-Shalev and Dromi (2021). For each of 428 words, the parent is asked to indicate whether the infant understands the word, understands and says the word, or not mark anything if the infant does not say and does not understand the word. The original CDI has high internal reliability (.95-.96, Fenson et al., 1994), as does the adapted Hebrew version (.98, Maital et al., 2002). The original CDI has a high correlation with infants' performance on the One Word Picture Vocabulary Test (.79) and their mean length of utterance (Fenson et al., 1994). For Hebrew, the test was not validated against an existing measure (since such a measure was not available) but rather, age-related growth curves were shown to be similar to those in the original English version, and expected effects such as an advantage for girls, and an effect of birth order were also demonstrated (see Gendler-Shalev & Dromi, 2022).

Norms exist from 12 months of age, but for the sake of this study, Gendler-Shalev provided us with unpublished norms from 9 months of age. For 8-month-olds (12 infants), we used the 9-months quantiles for comprehension, but removed these children from the analysis for the production models, since, even at 9 months, infants in the 50th quantile only produce 1 word, and it is thus unclear whether an infant who does not yet produce a single word is in the 10th or 40th quantile. Since 8-months-olds would reasonably produce even fewer words, we reasoned that it was uninformative whether an 8-month-old produced 1 word or none. Thus, the analysis of expressive vocabulary only included 81 infants. The online forms of the Hebrew CDI include some demographic questions, which we used to describe our sample's demographics in addition to our own questionnaire which we describe below.

Language-promoting Interactions. We developed this questionnaire based on a previous longitudinal study which included observations of 9- to-18-months-old infants (Alison & Clarke, 1973), as well as items borrowed from the StimQ questionnaire Parental Involvement in Developmental Advance subscale (Dreyer et al.,1996) such as: “Do you have opportunities, daily, to point at objects in the environment of the house and name them (such as point at a tree and say “tree”)?”. For each question the parent indicated whether the infant or themselves often act in this way (2 points), sometimes act in this way (1 point) or does not yet act in this way (0 points).

As mentioned above, there was some challenge with operationalizing the concept of language-promoting interactions. Interactions can either be initiated or led by the infant, or they can be initiated, led, or controlled by the parent. Given this complexity, we opted to develop a questionnaire which captures both types of behaviors. 11 items ask about the infant’s proactive behavior, such as “Does your child bring you books to read to her/him[1]?” and “Does your child attempt to draw your attention by throwing an object out of reach?”. 10 items ask about parental behaviors, such as “Do you have opportunities, daily, to point at objects in the environment of the house and name them (such as point at a tree and say “tree”)?” and “Do you teach your child the names of body parts while touching her/him and naming the body part (e.g., “here is your nose”)?”. Most of the items pertaining to parental behavior are *related* to items asking about infant behavior. For example, the infant-behavior item “When your child needs help, or wants an object that is out of reach, do they try to draw your attention to it in some way (e.g., by looking at it, vocalizing or pointing to the object)?” is followed by the parental-behavior item: “Do you tell the child in words what they asked for (for example, “did you want me to give you the pacifier?”)?”. Three items directly relate to an interaction (e.g., “Does your child come over to you when you call him/her?”), and the remaining two are “Does your child make sounds?” and “Does your child play with an object and explore it in different ways (e.g., banging on it or throwing it)?”.

See supplementary materials for the full questionnaire in the OSF link:
https://osf.io/hrmp6/?view_only=1248633dd1e943babdf316e4ed205191.

Demographic details. We asked about children’s date of birth, sex, maternal years of education, parents’ native languages, the percentage of time infants hear each language, the number of children in the family, birth order, week of gestation at birth, and birth weight.

Additional measures. We also asked parents about the age at which their child began sitting, standing, crawling and walking. This data was not analyzed but the number of sitting, standing, crawling, and walking infants, as well as the mean age at which they reached theses millstones are summarized in Table 1[2].

[1] The questionnaire was in Hebrew, which does not have a gender-neutral pronoun. Parents received a version of the questionnaire which fit their and their infant’s gender.

[2] We additionally collected the Availability of Learning Materials and Reading subscales of the StimQ (Dreyer, et al., 1996). The Availability of Learning Materials subscale of the StimQ produced a ceiling

Table 1. ***Participants' characteristics on all collected measures (N = 93)***

Measure	Range	Mean (SD) / median or mode were appropriate
Infant's sex	42% girls	
CDI filler's gender	94% mothers, 4% fathers, 2% both parents filled in the CDI	
Infant's age	8-18	12.4 (3.24)

Birth order	1-5	1.82 (0.97)
		Mode = 1
		Median = 2

Maternal years of education	13-23	16.8 (1.39)
ASQ gross-motor scale	0-6	4.34 (1.68)
Language-promoting interactions questionnaire	7.5-33.5	24.75 (5.03)
CDI comprehension (number of words)	0-401	111.96 (104.33)
CDI comprehension (quantile)	10-90	42.45 (28.27) Median = 50
CDI production (number of words)	0-220	21.63 (37.21)
CDI production (quantile)	10-90	34.81 (29.06) Median = 25
StimQ (RD) reading scale	0-16	9.59 (3.97)
Age at which began sitting unsupported (N = 75)	5.5-11.5	7.6 (1.56)
Age at which began standing unsupported (N = 79)	5.5-14	8.89 (1.85)
Age at which began crawling (N = 84)	4.5-10.5	6.59 (1.63)
Age at which began walking (N = 36)	8.5-18	12.78 (1.94)
Duration sitting unsupported (N = 75)	0.11-13.68	5.85 (3.26)
Duration standing unsupported (N = 79)	0.1-13.68	4.47 (3.13)
Duration crawling (N = 84)	0.12-12.29	6.45 (3.45)
Duration walking (N = 36)	0.1-8.39	3.09 (1.76)

effect where all participants achieved the highest score, and was not used. The READ scale was finally not used as a control in the analyses, as there is no a priori reason for it to be related to motor development, only language development.

Note: The reason Ns differ from the total sample size for motor milestones' age and duration is that some children did not yet achieve these milestones, for example 8 out of 93 infants did not yet independently sit unsupported, resulting in N = 75.

Analysis Plan

For each model (CDI comprehension and production quantile separately), we first examined the relationship between motor development and language-promoting interactions with a linear model (the lm function). We then examined the relationship between language-promoting interactions and CDI quantile with the glmmTMB function (glmmTMB package, Brooks et al., 2017), and a beta family parameter (since the dependent variable is measured in quantiles), and then examined the relationship between motor development and CDI quantile with the glmmTMB function. For these models, we also calculated a Baysian approximation using Bayesian Information Criterion (BIC) values, relying on the R package bayestestR and the function bic_to_bf (Makowski et al., 2019). A BIC provides an approximation to a Bayesian hypothesis test, but does not require the specification of priors (see Wagenmakers, 2007). Finally, we ran a mediation analysis using the robmed package (Alfons et al., 2022) and the test_mediation function via bootstrapping (5000 interaction).

In all models, we statistically controlled for factors hypothesized to be related to both motor and language development (Wysocki et al., 2022): child's age and sex, birth order, and maternal education.

Results

Before examining our hypotheses, we descriptively present Pearson correlations between our variables in Figures 1 and 2. Infants' age was correlated with infants' ASQ gross-motor scale scores, their language-promoting interaction scores and StimQ parental reading scores. Their CDI comprehension quantiles were correlated with their ASQ gross-motor scale scores and StimQ parental reading scores. Their ASQ gross-motor scale scores were correlated with their CDI language production and comprehension quantiles, StimQ parental reading scores, language-promoting interactions score and their age. Their language-promoting interaction scores were correlated with CDI language production quantile, their ASQ gross-motor scale scores, age, maternal education and StimQ parental reading scores.

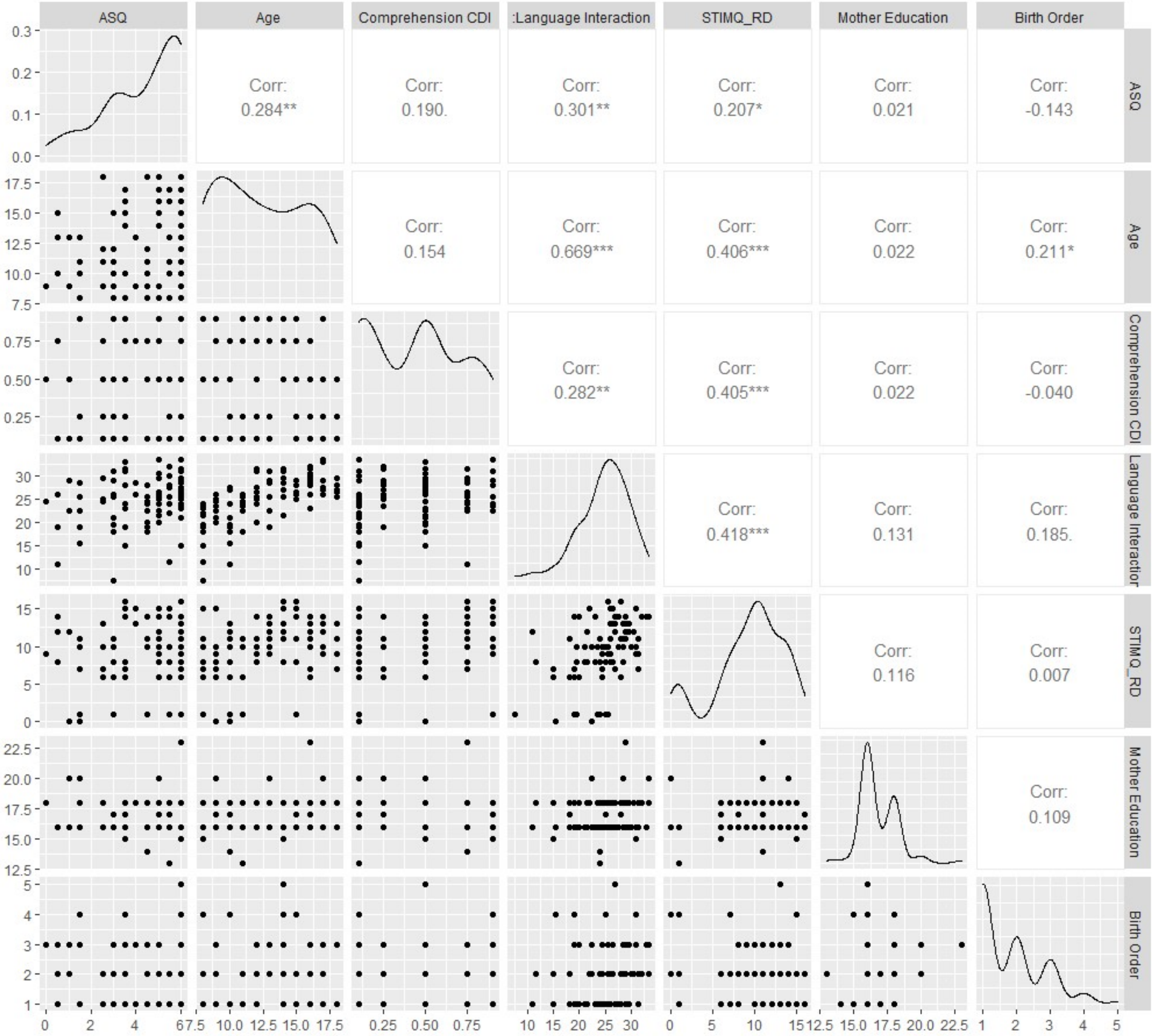


Figure 1: ***Relationship between all measured variables for the language comprehension sample. The diagonal shows density of the distribution of each of the variables. Panels below the diagonal show the scatter plot for the two variables involved (e.g., age and ASQ gross-motor subsection, first column). Those above the diagonal show the Pearson correlation for the two variables involved.***

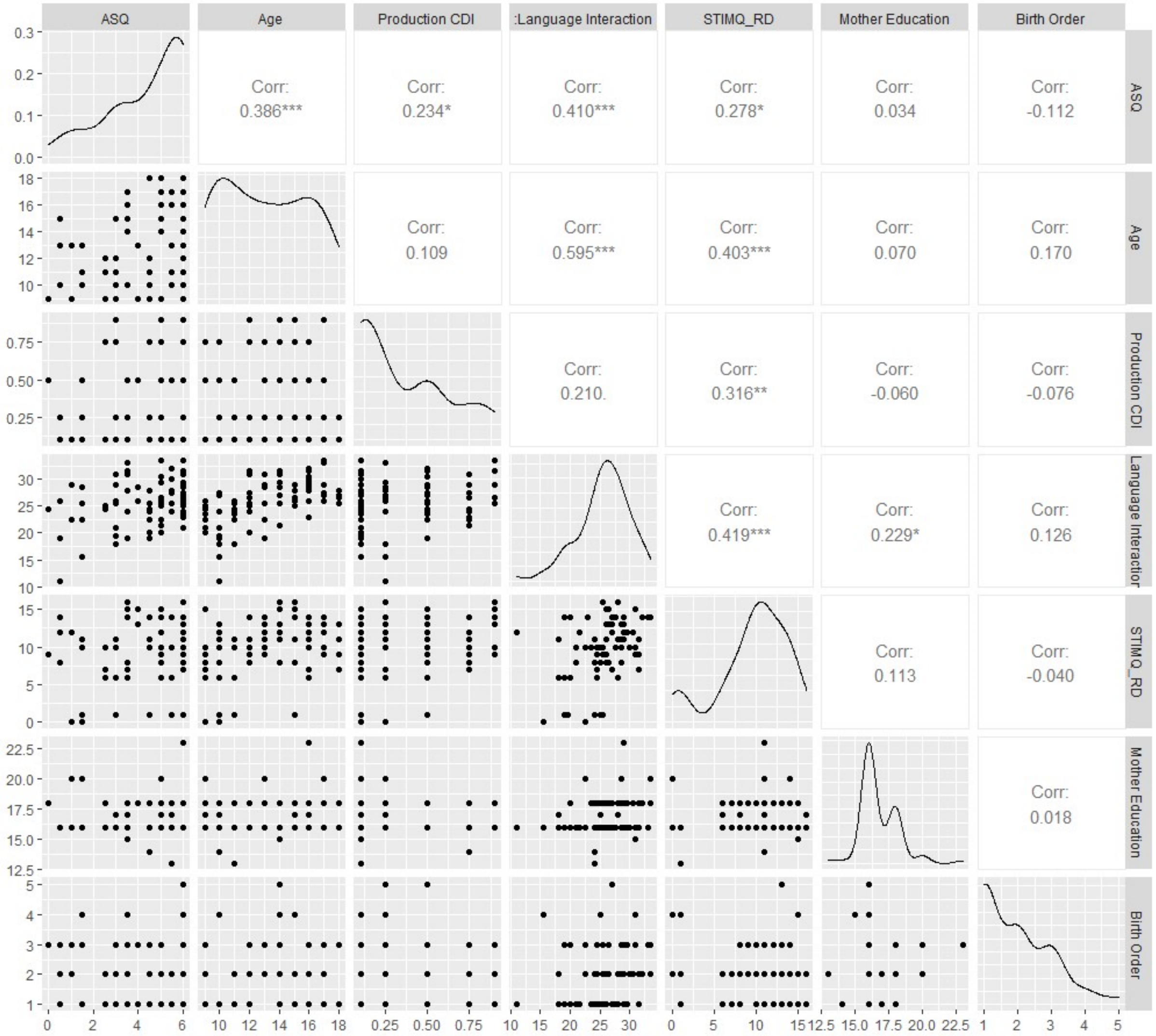


Figure 2: *Relationship between all measured variables for the language production sample. The diagonal shows density of the distribution of each of the variables. Panels below the diagonal show the scatter plot for the two variables involved (e.g., age and ASQ gross-motor subsection, first column). Those above the diagonal show the Pearson correlation for the two variables involved.*

We next tested our hypotheses about the relationship between motor development, language-promoting interactions, and language development (CDI comprehension and production separately).

Motor development, language-promoting interactions and language comprehension (N = 93)

There was a significant relationship between language-promoting interactions scores and CDI comprehension quantiles, with anecdotal evidence for a relationship between the two in the Bayesian analysis (beta = 0.071, SE = 0.033, *p* = .029, BF = 1.212, see table 2). There was no significant relationship between ASQ gross-motor scale and language-promoting interactions, with anecdotal evidence against a relationship between the two in the Bayesian analysis (beta = 0.382, SE = 0.233, p = .104, BF = 0.442, see table 2). The relationship between ASQ gross-motor scale and CDI comprehension quantiles was not significant, and the Bayesian analysis showed moderate evidence against a relationship between the two (beta = 0.73, SE = 0.069, *p* = .289, BF = 0.187, see table 2).

Table 2. ***Summary statistics for the models testing the relationship between CDI comprehension quantiles, ASQ gross-motor scale and language-promoting interactions.***

Model 1: CDI by Interactions questionnaire	Estimate	Std. Error	z value	p
(Intercept)	-1.525	1.410	-1.081	.280
Language-promoting interactions questionnaire	0.071	0.033	2.189	.0286*
Age	-0.025	0.047	-0.531	.595
sex - Male	0.055	0.226	0.243	.808
Maternal education	-0.001	0.079	-0.018	.985
Birth order	-0.093	0.116	-0.803	.422

Model 2: Interactions questionnaire by ASQ	Estimate	Std. Error	z value	p
(Intercept)	4.470	4.774	0.936	.352
ASQ	0.382	0.233	1.642	.104
Age	0.942	0.126	7.503	< .0001***
sex - Male	0.614	0.778	0.789	.432
Maternal education	0.383	0.271	1.417	.160

Birth order	0.226	0.415	0.546	.587

Model 3: CDI by ASQ	Estimate	Std. Error	z value	p
(Intercept)	-1.231	1.393	-0.884	.377
ASQ	0.073	0.069	1.061	.289
Age	0.036	0.036	0.981	.327
sex - Male	0.077	0.227	0.338	.736
Maternal education	0.020	0.079	0.252	.801
Birth order	-0.064	0.121	-0.528	.597

Bootstrapped mediation analysis found that the point estimate of the total effect of the ASQ gross-motor scale on the CDI comprehension quantile was .032 (*p* = .21), the direct effect was .029 (*p* = .272), and the indirect effect was .004, thus, no mediated effect was attested (see Table 3).

Table 3. ***Summary statistics for the mediation analysis between the CDI comprehension quantiles and ASQ gross-motor scale, mediated by language-promoting interactions.***

Total effect of ASQ on CDI:	Data	Boot	Std. Error	Z value	*p*
	0.033	0.032	0.026	1.257	.209
Direct effect of ASQ on CDI:					
	0.029	0.027	0.025	1.093	.275
Indirect effect of ASQ on CDI though Language-promoting interactions questionnaire:					
	Data	Boot	95% CI Lower	95% CI Upper	
	0.004	0.005	-0.006	0.025	

Motor development, language-promoting interactions and language production (N = 81)

There was no significant relationship between language-promoting interactions scores and CDI production quantiles, with anecdotal evidence against a relationship between the two in the Bayesian analysis (beta = 0.056, SE = 0.035, p = .106, BF = 0.43, see Table 4). There was a significant relationship between ASQ gross-motor scale scores and language-promoting interactions scores, with anecdotal evidence for a relationship between the two in the Bayesian analysis (beta = 0.589, SE = 0.24, p = .017, BF = 2.62, see Table 4). The relationship between ASQ gross-motor scale and CDI production quantiles was not significant, with moderate evidence against a relationship between the two in the Bayesian analysis (beta = 0.098, SE = 0.072, p = .176, BF = 0.285, see Table 4).

Table 4. ***Summary statistics for the models testing the relationship between CDI production quantiles, ASQ gross-motor scale and language-promoting interactions.***

Model 1: CDI by Interactions questionnaire	Estimate	Std. Error	z value	*p*
(Intercept)	-0.684	1.387	0.493	.622
Language-promoting interactions questionnaire	0.056	0.035	1.614	.106
Age	0.007	0.051	0.140	.889
Sex - Male	-0.044	0.248	0.179	.858
Maternal education	-0.073	0.079	0.921	.357
Birth order	-0.090	0.118	0.762	.446

Model 2: Interactions questionnaire by ASQ	Estimate	Std. Error	z value	*p*
(Intercept)	3.241	4.816	0.673	.503
ASQ	0.589	0.240	2.454	.017 *
Age	-0.509	0.841	0.605	.547
Sex - Male	0.774	0.152	5.086	< 0.001 ***
Maternal education	0.570	0.271	2.102	.039 *
Birth order	0.347	0.399	0.871	.387

Model 3: CDI by ASQ	Estimate	Std. Error	z value	*p*
(Intercept)	-0.511	1.374	0.372	.710
ASQ	0.098	0.072	1.352	.176
Age	0.036	0.044	0.824	.410
Sex - Male	-0.093	0.248	0.377	.706
Maternal education	-0.048	0.078	0.619	.536

Birth order	-0.049	0.118	0.412	.681

Bootstrapped mediation analysis found that the point estimate of the total effect of the ASQ gross-motor scale on the CDI production quantile was .02 (p = .365), the direct effect was .016 (p = .43), and the indirect effect was .002 - thus, no mediated effect was attested (see Table 5).

Table 5. ***Summary statistics for the mediation analysis between the CDI production quantiles and ASQ gross-motor scale, mediated by language-promoting interactions.***

Total effect of ASQ on CDI:	Data	Boot	Std. Error	Z value	*p*
	0.018	0.02	0.022	0.892	.372
Direct effect of ASQ on CDI:					
	0.016	0.017	0.022	0.772	.44
Indirect effect of ASQ on CDI though Language-promoting interactions questionnaire:					
	Data	Boot	95% CI Lower	95% CI Upper	
	0.002	0.002	-0.004	0.025	

Discussion

The present study examined the theory that the relationship between motor development and language development is mediated by language-promoting interactions. According to Iverson (2010, 2021), motor development helps children have more complex and self-initiated interactions with their environment, in ways that encourage parents to produce quality language-promoting input. For example, a walking infant may carry their favorite toy to their caretaker, encouraging joint attention around an object of their interest. This makes the infant a proactive, and not just an active, partner in these interactions. Such interactions, in turn, have been shown to support language development (Brooks & Meltzoff, 2005; Morales et al., 2000).

Here, we tested this suggestion by asking parents about their infant's motor development, language development, and their interactions with their infants. We found that, in comprehension, there was no significant direct or mediated relationship between motor and language development. However, there was a significant relationship between language-promoting interactions and language comprehension as previously found (e.g., Ramírez-Esparza, et al., 2014). We found no relationship between motor development and language-promoting interactions.

The picture was slightly different for children's production scores. Here, motor development was related to language-promoting interactions, but there was no relationship between language-promoting interactions and language production, nor a direct or mediated relation between motor development and language development.

Note that, unlike their vocabulary scores, which differ in the comprehension and production analysis, infants' motor and language-promoting interaction scores are the same in both models. However, the production sample is a subsample of the comprehension sample and it therefore smaller and biased towards older ages in the language-production analysis - given that 8-month-olds did not have production quantiles, but did have comprehension quantiles. We therefore do not want to give too much weight to the fact that we found a significant effect of motor development on language-promoting interactions in the production but not comprehension analysis. The fact we only find this effect in a subsample of our study might be because the effect only exists in older children, but we believe it is likely just due to chance. It is, however, also possible that this effect only exists in older infants, who are more likely to have made the transition to walking, as will be discussed below.

Since most of our results, and especially the direct relationship between motor and language development, were not significant, we should consider the possibility that such a relationship indeed does not exist. The first possible reason could be that one crucial tipping point in the co-development of language and motor ability is walking. This milestone has been found to be particularly related to gesture growth, gesture and vocalization coordination and contingent talk by parents (Schneider & Iverson, 2021; West & Iverson, 2020). Schneider & Iverson (2021) found infants were more likely to hear caregiver language and gestures that either requested or described movement or provided information about objects - after they made the transition to walking. Moreover, they found an effect of infants' real-time behavior, such that infants were more likely to hear language from their caregivers when they moved while upright than when they crawled. These parental behaviors are most likely captured by our language-promoting interactions measure. Other researchers also focused on walking onset and experience and their relation to productive and receptive vocabulary. For example, Walle (2016) found that infant initiation of joint engagement and following of the parent's joint engagement cues increased as infants gained walking experience (again, these behaviors should be captured by the measure we developed). He also found that walking experience predicted infants' receptive and productive language. Our sample only included 36 infants who could already walk, as opposed to 57 who could not yet walk. Future studies should focus more on the transition to walking.

We should also consider the possibility of a meaningful null result, one that signifies that a relationship between language and motor development is not statistically reliable. However, this goes against much of the literature today, on both typically developing (see Gonzalez et al., 2019; and Leonard & Hill, 2014 for systematic reviews), and disabled children (e.g.,Gaines & Missiuna, 2007; Ross et al., 2018). While these previous findings might also represent a biased literature base, we would be very wary to suggest so, given their consistency across different populations. We find it more likely that the limitations of the current study (a small sample size, exclusive use of parental reports from the same parent, cross-sectional design, little focus on the transition to walking described above) prevented us from finding an effect of motor skills on language skills.

As for a lack of significant mediation by language-promoting interactions, it might be due to the same limitations that prevented us from finding a significant relation between motor and language development. It could also be that the measure of language-promoting interactions we developed was not valid or not sensitive enough to individual differences in behavior. Indeed, we described in the introduction the challenge of developing a measure that would capture both parent-initiated and child-initiated interactions, as well as parental responses to child-initiated interactions. It could also be that parental reports are not a good way to assess interactions, as parents might be driven by social desirability to report behaviors which sound supportive or beneficial for child development. Indeed, this might be the reason studies of parental interactions with their children tend to use observational methods (e.g., Alison & Clarke, 1973, on which we based a portion of our questionnaire). It might be worthwhile to test the same hypothesis with observational methods (for both motor development and interactions), however, this would most probably serve to lower even further the sample size.

Our findings do not lend support to the hypothesis that language-promoting interactions mediate the relationship between motor and language development. However, it would be wrong to rely on them to claim the opposite – most of our analyses did not find significant results, but Bayesian analysis shows them to be inconclusive rather than supporting the null hypothesis. We therefore suggest the main conclusion from this study should be that there is need for further research, especially research tackling the main limitations of the current study described above. In addition, the questionnaire we developed has not been validated. However, we hope others will use it, validate, and improve it for use in the study of motor development, language development, and the relationship between the two.

References

Agler, R., & De Boeck, P. (2017). On the interpretation and use of mediation: multiple
perspectives on mediation analysis. *Frontiers in Psychology, 8*, 1984. https://doi.org/10.3389/fpsyg.2017.01984

Alfons A., Ates, N.Y., & Groenen P.J.F. (2022). Robust Mediation Analysis: The R Package robmed. *Journal of Statistical Software, 103*(13), 1-45. doi: 10.18637/jss.v103.i13

Bambach, S., Crandall, D., Smith, L., & Yu, C. (2018). Toddler-inspired visual object learning. *Advances in Neural Information Processing Systems, 31*.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*(6), 1173.

Bates, E., Benigni, L., Bretherton, I., Camaioni, L., Volterra, V. (1979). The emergence of symbols: Cognition and communication in infancy. New York: Academic Press. https://doi.org/10.1016/0378-2166(83)90154-6

Bedford, R., Pickles, A., Lord, C. (2016). Early gross motor skills predict the

subsequent development of language in children with autism spectrum disorder. *Autism Research, 9*(9), 993-1001. https://doi.org/10.1002/aur.1587

Bernardis, P., Bello, A., Pettenati, P., Stefanini, S., Gentilucci, M. (2008). Manual actions affect the vocalizations of infants. *Experimental Brain Research 184*, 599–603. https://doi.org/10.1007/s00221-007-1256-x

Berger, S. E., Cunsolo, M., Ali, M., Iverson, J. M. (2017). The trajectory of concurrent motor and vocal behaviors over the transition to crawling in infancy. *Infancy, 22*(5), 681-694. DOI: 10.1111/infa.12179

Brooks ME, Kristensen K, van Benthem KJ, Magnusson A, Berg CW, Nielsen A, Skaug HJ,
Maechler M, Bolker BM (2017). "glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling." *The R Journal*, 9(2), 378–400. https://journal.r-project.org/archive/2017/RJ-2017-066/index.html.

Clearfield, M.W. (2011). Learning to walk changes infants' social interactions. *Infant*

Behavior and Development. 34(1), 15-25. https://doi.org/10.1016/j.infbeh.2010.04.008

Dreyer, B. P., Mendelsohn, A. L., Tamis-LeMonda, C. S. (1996). Assessing the

child's cognitive home environment through parental report, reliability and validity. *Early Development and Parenting: An International Journal of Research and Practice, 5*(4), 271-287. DOI: 10.1002/(SICI)1099-0917(199612)5:4<271::AID-EDP138>3.0.CO;2-D

Gendler-Shalev, H., Dromi, E. (2021). The Hebrew Web Communicative Development Inventory (MB-CDI): Lexical Development Growth Curves. Journal of Child Language, advanced online publication. https://doi.org/10.1017/S0305000921000179

He, M., Walle, E. A., Campos, J. J. (2015). A cross-national investigation of the relationship between infant walking and language development. *Infancy, 20*(3), 283-305. https://doi.org/10.1111/infa.12071

Hirotani, M., Stets, M., Striano, T., & Friederici, A. D. (2009). Joint attention helps infants
learn new words: event-related potential evidence. *Neuroreport, 20*(6), 600-605. doi: 10.1097/WNR.0b013e32832a0a7c

Iverson, J. M. (2021). Developmental variability and developmental cascades: Lessons from motor and language development in infancy. *Current Directions in Psychological Science, 30*(3), 228-235.

Iverson, J. (2010). Developing language in a developing body: the relationship between motor development and language development. *Journal of Child Language, 37*(2), 229-261. https://doi.org/10.1017/S0305000909990432

Iverson, J. M., West, K. L., Schneider, J. L., Plate, S. N., Northrup, J. B., & Britsch, E. R. (2022). Early development in autism: How developmental cascades help us understand the emergence of developmental differences.

Karasik, L. B., Tamis-LeMonda, C. S., Adolph, K. E. (2011). Transition from crawling to walking and infants' actions with objects and people. *Child Development, 82*(4), 1199-1209. https://doi.org/10.1111

Lifter, K., Bloom, L. (1989). Object knowledge and the emergence of language. *Infant Behavior and Development, 12*(4), 395-423. https://doi.org/10.1016/0163-6383(89)90023-4

Loeys, T., Moerkerke, B., & Vansteelandt, S. (2015). A cautionary note on the power of the test for the indirect effect in mediation analysis. *Frontiers in Psychology, 5,* 1549. https://doi.org/10.3389/fpsyg.2014.01549

Maital, S., Dromi, E., Sagi, A., Bornstein, M. (2000) The Hebrew Communicative Development Inventory: Language specific properties and cross-linguistic generalizations. *Journal of Child Language, 27*(1), 43-67. https://doi.org/10.1017/S0305000999004006

Makowski D, Ben-Shachar M, Lüdecke D (2019). bayestestR: describing effects and their uncertainty, existence and significance within the Bayesian framework. *Journal of Open Source Software, 4*(40), 1541. doi: 10.21105/joss.

Maxwell, S. E., & Cole, D. A. (2007). Bias in cross-sectional analyses of longitudinal mediation. *Psychological methods, 12*(1), 23. https://doi.org/10.1037/1082-989X.12.1.23

Moore, C., Dailey, S., Garrison, H., Amatuni, A., & Bergelson, E. (2019). Point, walk, talk:
Links between three early milestones, from observation and parental report. *Developmental psychology, 55*(8), 1579-1593. https://doi.org/10.1037/dev0000738

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

Ross, G., Demaria, R., Yap, V. (2018). The relationship between motor delays and language development in very low birthweight premature children at 18 months corrected age. *Journal of Speech, Language, and Hearing Research, 61*(1), 114-119. https://doi.org/10.1044/2017_JSLHR-L-17-0056

Schneider, J. L., & Iverson, J. M. (2022). Cascades in action: How the transition to walking
shapes caregiver communication during everyday interactions. *Developmental Psychology, 58*(1), 1-16. https://doi.org/10.1037/dev0001280

Squires, J., Bricker, D., Potter, L. (1997). Revision of a parent-completed developmental screening tool: Ages and Stages Questionnaires. *Journal of pediatric psychology, 22*(3), 313-328. https://doi.org/10.1093/jpepsy/22.3.313

Tingley D, Yamamoto T, Hirose K, Keele L, Imai K (2014). "mediation: R Package for Causal Mediation Analysis." *Journal of Statistical Software*, *59*(5), 1–38. http://www.jstatsoft.org/v59/i05/.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values.
Psychonomic bulletin & review, 14(5), 779-804.

Walle, E. A. (2016). Infant social development across the transition from crawling to

walking. *Frontiers in psychology, 7*, 960. https://doi.org/10.3389/fpsyg.2016.00960

Walle, E. A., Campos, J. J. (2014). Infant language development is related to the acquisition of walking. *Developmental Psychology, 50*(2), 336-348. https://psycnet.apa.org/doi/10.1037/a0033238

West, K. L., Fletcher, K. K., Adolph, K. E., & Tamis-LeMonda, C. S. (2022). Mothers talk
about infants' actions: How verbs correspond to infants' real-time behavior. *Developmental Psychology, 58*(3), 405–416. https://doi.org/10.1037/dev0001285

West, K. L., & Iverson, J. M. (2021). Communication changes when infants begin to walk.
Developmental Science, 24(5), e13102. https://doi.org/10.1111/desc.13102

West, K. L., & Iverson, J. M. (2017). Language learning is hands-on: Exploring links between infants' object manipulation and verbal input. *Cognitive Development, 43,* 190-200. https://doi.org/10.1016/j.cogdev.2017.05.004

West, K. L., Saleh, A. N., Adolph, K. E., & Tamis-LeMonda, C. S. (2023). "Go, go, go!" Mothers' verbs align with infants' locomotion. *Developmental science,* e13397. https://doi.org/10.1111/desc.13397

Wysocki, A. C., Lawson, K. M., & Rhemtulla, M. (2022). Statistical control requires causal
justification. *Advances in Methods and Practices in Psychological Science, 5*(2), 25152459221095823. https://doi.org/10.1177/25152459221095823

Yingling, J. M. (1982). *Temporal Features of Infant Speech: A Description of Babbling Patterns Circumscribed by Postural Achievement* (Unpublished doctoral dissertation). University of Denver.

Data, code and materials availability statement

The data and code are available on the OSF: https://osf.io/hrmp6/?view_only=1248633dd1e943babdf316e4ed205191.

As for materials, the language-promoting interactions questionnaire we developed, as well as the StimQ, are available on the link above. The Hebrew Web Communicative Development Inventory - CDI is available on WordBank, though note that currently the norms are not updated, and no norms for children under 12 months were yet uploaded: http://wordbank.stanford.edu/data?name=vocab_norms

The editor approved an exemption (18th January 2023) to materials sharing for the Ages and Stages Questionnaire, which is subject to copyright. A sample questionnaire is available on the ASQ website: https://agesandstages.com/wp-content/uploads/2015/02/asq-3-48-month-sample.pdf

Ethics statement

Ethics approval was obtained from the ethics committee of the University of Haifa, School of Psychological Sciences. All participants gave informed written consent before taking part in the study.

Authorship and Contributorship Statement

SBO and NH conceived of the study, designed the study and wrote the first draft of the manuscript. SBO collected the data. NH analysed the data. All authors approved the final version of the manuscript and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Acknowledgements

We would like to thank all participating families, as well as the University of Haifa Research Authority for a small research grant to SBO.

License

Morphosyntactic Analysis for CHILDES

Houjun Liu
Stanford University

Brian MacWhinney
Carnegie Mellon University

Abstract: Language development researchers are interested in comparing the process of language learning across languages. Unfortunately, it has been difficult to construct a consistent quantitative framework for such comparisons. Fortunately, recent advances in AI (Artificial Intelligence) and ML (Machine Learning) are providing new methods for ASR (automatic speech recognition) and NLP (natural language processing) that can be brought to bear on this problem. Using the Batchalign2 program (Liu et al., 2023), we have been transcribing and linking new data for the CHILDES database and have applied the UD (Universal Dependencies) framework to existing data to provide a consistent and comparable morphosyntactic analysis for 27 languages. These new resources open possibilities for deeper crosslinguistic study of language learning.

Keywords: morphology; grammatical relations; ASR; NLP

Corresponding author: Brian MacWhinney, Department of Psychology, Carnegie Mellon University, Pittsburgh, PA. Email: macw@cmu.edu

ORCID ID: https://orcid.org/0000-0002-4988-1342

Citation: Liu, H. & MacWhinney, B. (2024). Morphosyntactic analysis for CHILDES. *Language Development Research*, *4*(1), 233—258. http://doi.org/10.34842/j97r-n823

Introduction

Child language research involves three, partially separate, formats for data collection. The first focuses on the development of a single child or pair of children, often across several years. Work in this tradition includes classic diary studies from German (Stern & Stern, 1907), French (Bloch, 1921; Guillaume, 1927), Polish (Smoczynska, 2017; Szuman, 1959), Hungarian (Kenyeres, 1926; Ponori, 1871), Mandarin (Chao, 1951), Bulgarian (Gvozdev, 1949), Serbian (Pavlovitch, 1920) and other languages. It also includes diary and transcript studies of particular aspects of development such as phonology (Smith, 1973), grammatical morphology (Brown, 1973), lexicon (Tomasello, 1992), or all of the above (Leopold, 1939, 1947, 1949a, 1949b). This case-study work has helped us understand the diverse ways in which children acquire and use language to express their needs (Karniol, 2010).

A second data collection format measures and evaluates learning across groups of children within a single language. This type of analysis is particularly important for clinicians who need to diagnose, assess, and remediate language learning disorders. Data collection in this format includes standardized tests (Bishop, 1982; Goldman & Fristoe, 2000), language sample analysis (Garbarino et al., 2020), and language profiling (Bernstein Ratner & MacWhinney, 2023; Crystal et al., 1989; Scarborough, 1990).

A third data collection format examines development across languages. This work considers the ways in which variations in language structure and social input pose challenges or opportunities to the learner. For various reasons, this work has had a concentration of data from WEIRD (Western, educated, industrialized, rich, and democratic) participants (Henrich et al., 2010) along with an emphasis on monolingual acquisition. To broaden our crosslinguistic coverage, Slobin and colleagues (Slobin, 1985) have provided descriptions of linguistic and social development in a series of languages, including some from non-WEIRD communities. However, without quantitative tools to compare across these many languages, it has been difficult to generalize about patterns of language learning methods, structures, and challenges. The introduction of the MacArthur-Bates Communicative Development Inventory (Dale & Fenson, 1996) provided quantitative methods to bridge the WEIRD gap for the earliest stages of lexical development. That tool has now been validated for several Western languages (Frank et al., 2021), but extensions to less well-resourced languages and multilingualism (Tamis-LeMonda et al., 2024) will take additional time and effort.

The CHILDES data-sharing system (MacWhinney, 2000) offers another approach to extending child language research beyond WEIRD participants. CHILDES includes language samples from 49 languages, along with 41 corpora from children learning two or more languages, all contributed by researchers who are speakers of these languages. Although many of these families are WEIRD, there are also many from societies that are not Western, and not fully industrialized, rich, or democratic. Although nearly 40% of the data is from English, there are many large corpora from languages such as Mandarin, Spanish, German, French, and Japanese as well as a smaller number of large corpora from another 15 languages.

Creating child language corpora requires major commitments of researcher effort for recording, transcription, and analysis. However, recent advances in AI (artificial intelligence) and ML (machine learning) have led to marked improvements in ASR (automatic speech recognition)(Radford et al., 2023) and NLP (natural language processing)(Nivre et al., 2016) methods that can markedly facilitate this work. The use of ASR can greatly speed transcription (Liu et al., 2023), although recognition of child vocalizations before age 3 is still poor. When recording is done well, ASR can recognize adult input accurately enough to allow a transcript to be finalized after a much briefer period of hand correction. A further advantage is that ASR creates a transcript that is linked to the audio on both the utterance and single word level, thereby facilitating analyses of phonology, fluency, and total time talking. Moreover, the output can be structured directly in the CHAT (Codes for Human Analysis of Talk) format, thereby allowing analysis through the utilities built into the CLAN (Child Language Analysis) program (MacWhinney & Fromm, 2022). ASR methods can also be used to automatically link a complete, but unlinked, transcript to the corresponding media (audio or video) on the utterance and word level. This process is particularly useful for transcripts in the CHILDES database that have media, but which have not yet been linked to that media.

After a transcript has been created in correct CHAT format, we can then use NLP methods to automatically construct a complete morphosyntactic analysis. Both for newly collected data and for data in the current repository, creation of fully analyzed and tagged corpora involves the use of a series of processes which have now all been integrated into the Batchalign2 program (Liu et al., 2023). In the next sections, we will describe how these ASR and NLP methods are being applied to improve the use of CHILDES data across all three of the data analysis formats we have described with a special emphasis on facilitating crosslinguistic comparisons.

Automatic Speech Recognition

Once a language sample has been recorded, the next task is to create a transcript. Depending on the nature of the interaction, manual transcription of one hour of interaction can take from 10 to 16 hours (Bernstein Ratner & MacWhinney, 2020). To speed up this process, researchers can apply ASR methods using the Batchalign2 system (Liu et al., 2023) which outputs a transcript in the CHAT format required for inclusion in the CHILDES database. Batchalign2 offers access to two ASR systems: the Rev.AI ASR cloud service (Del Rio et al., 2022) or a local ASR model based on OpenAI Whisper (Radford et al., 2023). If IRB (Institutional Review Board) regulations do not allow transmission of data to a cloud service, users may prefer to use Whisper, although Rev.AI explicitly allows the user to determine that their data will not be stored on the Rev.AI cloud server. For English, Rev.AI output is a bit more accurate than Whisper due to the its use of a large amount of two-party conversations as training data (Del Rio et al., 2022). In addition, processing through Rev.AI is much faster than running with Whisper, particularly when local hardware is limited, but both options are good choices.

Another factor that favors use of Whisper is that the training data for the NLP models used in downstream analysis use native orthographies of each language (De Marneffe et al., 2021; Qi et al., 2020). Latinized transcripts must be converted back into the standard orthography for the language before downstream analysis. Because of this limitation, the significantly wider language and orthographic profile of the Whisper model (in particular, WhisperV3 available at https://huggingface.co/openai/whisper-large-v3) is advantageous for non-English languages not covered by Rev-AI. Therefore, most of the ASR work that we have used to cover all the languages described here (and in particular ones with non-latinized native orthography) is performed with the Whisper option.

Utterance Segmentation

Tagging for morphological categories and grammatical dependency structure requires accurate delineation of sentences or utterances. Segmentation of naturalistic spoken language data requires attention to features not found in written text (Fraser et al., 2015), such as incompletion, repetition, retracing, and other features. Sections 9.1 and 9.2 of the CHAT manual (https://talkbank.org/manuals/CHAT.pdf) provide a set of standards for utterance segmentation. For example, one important feature is that clauses joined only with coordinating conjunctions (and, or, but) are treated as separate *utterances*.

Because currently available tokenizers are all based on written language and because spoken language segmentation follows quite different rules and patterns, we have created novel tokenizers based on spoken language training data. To create the tokenizer for spoken English data, we turned to the TalkBank database, which contains many Gold Standard utterances segmented according to the rules mentioned above. The tokenizer (Liu et al., 2023) is trained via a token-classification task, which assigns each input text token as being the start (label 1), middle (label 0), a phrase which should be separated by a comma (label 5), or end of each utterance (label 2,3,4); in particular, there are three utterance-ending labels, each corresponding to the utterance being declarative, interrogative, or exclamatory respectively. The tokenizer uses a BERT-class model (Devlin et al., 2018) to generate semantic embeddings for language modeling, and a deep neural network (DNN) to perform token-level annotations.

Currently, Batchalign2 provides tokenizers for English and Mandarin. The English model was trained on the MICASE (The Michigan Corpus of Academic Spoken English) (Römer, 2019) corpus in CABank (https://ca.talkbank.org/access/MICASE.html), which includes transcribed data from 300 participants in a wide variety of interactions between students and faculty at the University of Michigan. The Mandarin model was trained on three corpora available on the TalkBank CHILDES database—Zhou Assessment (Li & Zhou, 2011), Chang Personal Narrative (Chang & McCabe, 2013), and Li Shared Reading. The ability to train new segmentation models based on segmented CHAT transcripts has been released along with the Batchalign2 software. In addition, work currently in progress by the HuggingFace diarization team (https://github.com/huggingface/diarizers) using the Pyannote framework (Bredin, 2023) with TalkBank data should be able to provide tokenizers for a wider variety of languages.

Text-Media Alignment

Apart from the processing of new recordings, ASR can also be used to link previously hand-transcribed transcripts to media for timing-aware analysis. Creation of these links allows us to improve the materials currently in CHILDES and other TalkBank repositories, many of which had no linkage between transcripts and media. Text-media alignment or linkage facilitates phonological analysis, analysis of fluency, study of the dynamics of international patterns, and playback through the TalkBank Browser. The Batchalign2 "align" command now supports this process by running a two-pass alignment of transcripts to media. This new process was not available in the previous version of Batchalign described in Liu et al. (2023) We provide here a high-level overview of this process. The first pass of this process involves performing rough, utterance time diarizations using ASR as a silver annotation reference. The second step involves extracting precise word-level timestamps through the analysis of the latent activations of audio-text cross-attention by using the Whisper ASR model.

Utterance Timing Recovery

We begin by assuming that the transcript to be linked has correctly segmented utterance text, but that it does not yet have any utterance time values. If the transcript has imprecise time values, we can use the CLAN CHSTRING command with the +cbullets.cut switch to remove them. We must then identify the relative time within the media in which an utterance occurred. This task is difficult to perform with classic alignment schemes, which face difficulty generating correct alignments among longer timestamps without some form of hierarchical or recursive scheme (Moreno et al., 1998), due to the exponential growth in number of possible alignments as sequence length increases.

To address this limitation, we take an optimistic, silver-labeling approach by using an

ASR-generated transcript (which can process the audio linearly by splitting it into segments) to obtain a silver transcript which we call the “backplate.” Because this ASR transcript has been generated directly from the audio, each of its tokens are linked against a relative timestamp within the audio file. By then aligning the transcript against the backplate, we can induce the timestamp in which each utterance in the gold standard transcript exists by reading the corresponding times on the backplate.

To perform the actual transcript-to-transcript alignment described above, we apply dynamic programming (Bellman, 1966) to create an alignment solution which minimizes the form-level Levenshtein edit distance (Levenshtein, 1965) between the gold transcript and the backplate. We can then calculate the level timings via direct computation using the first and last timestamps of aligned forms within an utterance labelled by the gold transcript, plus some time on each end to account for errors which will be tightened in the second step of the overall alignment procedure.

Although this procedure could theoretically also recover the timing of each individual token by aligning the backplate transcript against gold at a token level, this initial alignment is only practically feasible for utterance timing recovery. Instead, we assume that the overall time alignment for an utterance (as denoted by the timing between its first aligned token and the last aligned token) should be roughly accurate. Because we are doing utterance level alignment, any errors in the backplate (such as missing a filled pause, a very common error in ASR) which are within the bounds of an utterance are essentially irrelevant to this procedure. Even if a particular utterance is not properly transcribed in the backplate, we can infer its temporal alignment by knowing the values for the previous and following utterances. However, application of this procedure on the token level would result in missing time values for all forms which do not have precise alignments between the gold and backplate transcripts—reducing the quality of the resulting data.

Word-level Forced Alignment

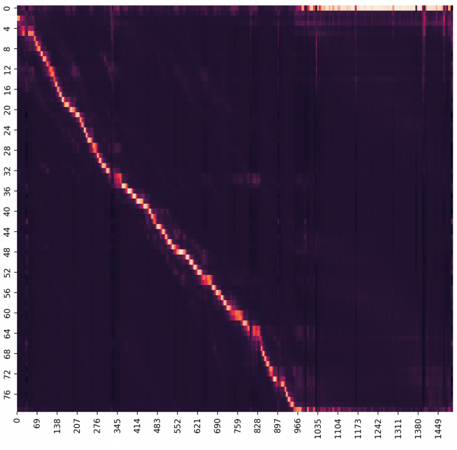
Next, to obtain word-level or token-level alignment, we perform an analysis of the Whisper ASR model attention activations to extract per-token audio-text alignment. Whisper is an encoder-decoder architecture model (Radford et al., 2023), whereby the encoder creates a latent embedding per sample (usually 16,000Hz) of the input audio sequence which is then used as input to the cross-attention (Niu et al., 2021) computation against the output text sequence.

The key motivation of our analysis follows closely to previous work in cross-attention activation analyses (Hou et al., 2019). We take advantage of the heuristic that the *highest audio-text cross attention scores (highest normalized value) are likely the most directly relevant pairings*. For speech analysis, this means that the most highly activated encoder time slice to decoder token activation is likely the best temporal alignment for the token. To take advantage of this fact, we run a single forward pass on the Whisper model per time-segmented utterance, providing the audio segment of the utterance (derived in the previous step of the utterance time diarization) as the encoder input and the gold utterance text as the decoder input. Then, we extract the last cross-attention activation matrix from model activations during this forward pass.

From this, we apply normalization procedures to ensure that the downstream processing is invariant of inherent inter-input variation—mean centering and median filter smoothing (Brownrigg, 1984)—to obtain a smoothed cross-attention matrix. After post-processing, such a matrix is given in Figure 1, clearly showing the relationship between text tokens (rows) and sequential timestamps (columns); taking successive highest values indices of this matrix along each axis reveals two sequences—one for time along each slice and another for transcript-token along each slice (as cross-attention scores correlate queries from the audio sequence against keys in the token sequence). The fact that Figure 1 displays a straight line indicates that all the words were correctly aligned in sequence. Finally, the actual alignment between these two sequences — which are already sorted in temporal order with alignments between them given by the matrix — can be computed through Dynamic Time Warping (DTW) (Berndt & Clifford, 1994) of these sequences together.

One notable step which is required for this procedure to function successfully is the need to impute the padding-token attention scores as the mean of the other scores. Empirically, the model attends to arbitrary positions in the audio whenever there is no speech at the end of the audio (i.e. the padding tokens at the end of the transcript), which would disrupt the DTW procedure. Hence, we set the across-attention scores of any timestamp against the padding token as the mean of all scores. This procedure is relatively quick to compute. Although DTW has O(nm) time complexity, the sequences are reasonably short, and they do not require perfect ASR performance because the gold transcript is provided directly to the Whisper decoder to compute attention. Through this scheme, we obtain a time alignment for each input word form which can be used in downstream analysis.

Figure 1. ***Cross-Attention matrix of audio-to-text activations showing temporal alignment between text tokens (rows) and timestamps (columns). Token 0 (padding) is shown to be highly active after the transcript ends during non-speaking segments, requiring later filtering.***



Universal Dependencies

Next, we will explain how Batchalign2 operates to produce morphosyntactic analyses. This work relies on the application of Universal Dependency (UD) models trained through the Stanza Python NLP package (Qi et al., 2020). This system, which can be used with over 70 languages (https://universaldependencies.org), is based on a consistent language-general set of codes for POS (parts of speech), GFs (grammatical features), and GRs (grammatical relations). Stanza models for each UD language can be downloaded for use by the Batchalign2 Python program which is freely available for download from https://github.com/talkbank. Before reviewing the details of the application of UD tagging to CHILDES data, we need to consider the previous state-of-the-art for tagging CHILDES transcripts.

Beginning in 1995, Brian MacWhinney, Roland Hausser, and Mitzi Morris created a system for word-level morphological coding called MOR (MacWhinney, 2008). This system relied on a series of hand-crafted declarative rules governing possible word analyses and a program called POST, created by Christophe Parisse (Parisse & Le Normand, 2000) for disambiguating alternative readings in context. The resultant analyses were entered on a %mor line in which each word on the main speech line is given its own morphological analysis. The manual for MOR is available at https://talkbank.org/manuals/MOR.pdf. Across the years, Leonid Spektor extended the MOR program and Brian MacWhinney refined the lexicon and rules to achieve a high level of accuracy and coverage. However, extending MOR to other languages represented a major challenge. Versions of MOR were created for French, Hebrew, Italian, Japanese, and Mandarin. However, these versions of MOR required a great deal of careful rule configuration by one or two people and learning how to build a new MOR grammar was difficult. Given this, extensions to the remaining 44 languages in CHILDES were outside the scope of the project.

Apart from word-level morphological analysis, creation of automatic programs for syntactic analysis across the 49 languages in CHILDES faced similar hurdles. Sagae and colleagues (Sagae et al., 2010) created a program called MEGRASP (maximum entropy grammatical relations syntactic processor) that uses the SVM (Support Vector Machine) method to tag CHILDES English and Spanish corpora for grammatical relation dependency structure. In principle, MEGRASP could be extended to cover additional languages. However, settling on consistent labels for the grammatical relations in each language and applying those labels to a large corpus of training utterances represented yet another major task that would have to be done one-by-one for all the languages in CHILDES.

Given the scope of the work needed to build MOR and MEGRASP analyzers for 49 languages and for languages that will be added to CHILDES in the future, we looked for alternative methods for building morphosyntactic analyses across languages.

Fortunately, the UD Project provides almost exactly what is needed. Relying on the latest AI/NLP technology, the UD community has been working to create taggers for 70 languages, including a majority that are outside of Indo-European. UD uses six open class POS (part-of-speech) tags (ADJ, ADV, INTJ, NOUN, PROPN, and VERB) and eight closed class POS tags (ADP, AUX, CCONJ, DET, NUM, PART, PRON, and SCONJ). It clusters GFs into seven lexical feature sets (PronType, NumType, Poss, Reflex, Foreign, Abbr, and Typo), nine nominal inflectional feature sets (Gender, Animacy, NounClass, Number, Case, Definite, Deixis, DeixisRef, and Degree) and ten verbal inflectional feature sets (VerbForm, Mood, Tense, Aspect, Voice, Evident, Polarity, Person, Polite, and Clusivity). Within each set, a further set of GF values is described. For example, Gender has the values Masc, Fem, Neut, and Com. Apart from this systematic listing of POS and GFs, UD provides a uniform nomenclature for grammatical relations (GRs) with six core arguments (nsubj, obj, iobj, csubj, ccomp, and xcomp), ten non-core dependents (obl, vocative, expl, dislocated, advcl, advmod, discourse, aux, cop, and mark), and ten coordination relations (conj, cc, fixed, flat, list, parataxis, compound, orphan, goeswith, and reparandum). The UD web pages provide complete descriptions of all these POS, GFs, and GRs and the documentation for each language shows how they map onto the language.

Preparing for UD Analysis

To align with the various format requirements of UD, Stanza, and Batchalign2, we first require that transcripts be in full compliance with the CHAT format as validated through the Chatter program which is available for download from https://talkbank.org/software/chatter.html. Because the CHILDES database had been validated using earlier versions of Chatter that failed to enforce some of the requirements of UD, we had to sharpen the specifications in Chatter and reapply the new version to the entire CHILDES database. That process involved a series of format fixes, such as systematization of spacing, use of new fluency codes, and elimination of use of the plus sign for marking compounds. To provide one-to-one alignment of text to audio, we also needed to eliminate use of repetition codes such as [x 3] for three repetitions of a word or phrase and we had to make overlap and retracing marking more consistent.

Once the data were in the required format, we could run the “morphotag” command in Batchalign2. Internally, this process creates data in the CONLL-U format which is then reformatted to the CHAT format to be written out in the %mor and %gra lines. The POS and GF information is formatted into the %mor line, and the GR information is outputted to the %gra line.

Matching the requirements of the UD grammars with the tokenization and transcripts in the CHILDES files faces problems that vary from language to language. One challenge found in nearly all the corpora is the use of eye-dialect to transcribe spoken forms. For example, in English some corpora may have used an apostrophe to represent conversion of final /ŋ/ to final /n/ as in *singin'* which then had to be converted to *singin(g).* Or German *hab'n* would be converted to *hab(e)n* for consistent recognition by the UD grammar. A form such as *tactor* could be converted to *t(r)actor,* whereas *practor* would be *practor [: tractor].* For languages such as French, Italian, and Spanish that had already gone through analysis by MOR, these standards were already in place, but for other languages word level forms had to be revised to match the standard.

For the Romance languages - French, Italian, Catalan, Portuguese, and Spanish - there were often issues relating to clitics and portmanteau forms. For example, the French corpora often inserted a space between proclitics and stems, as in *j' ai* rather than *j'ai* with the latter being the form expected in standard French orthography. Such divergences were easy enough to fix using global replacements. More complicated cases involved conversions such as *qu'est-ce-que* into *qu'est-ce que.* In each case, the goal of the conversions was to produce output that would match standard orthography, because this is how UD is trained and what it expects.

Another issue facing UD analysis involved how best to handle multi-word expressions (MWE) which the NLP literature refers to as multi-word tokens (MWT). For example, the French word for today is *aujourd'hui,* but without entering this form specifically as an MWT, Stanza's models would separate the front part as the prepositional phrase *au jour* (on the day) and then ending up as unable to tag the remaining segment *d'hui.* To address this problem, we introduced a modification in the Stanza pipeline that allowed for a specified set of MWTs to block over-analysis.

It was also necessary to make sure that the word-level transcription for each language matched the standard orthography used for that language, because UD grammars are trained on data in the standard orthography. This means that romanized transcripts for languages that use a non-Roman script need to be converted back to the standard orthography for that language. We are currently trying to deal with this problem by training a transformer based on human-checked gold-standard input.

Current State of UD Tagging

Here we summarize the status of the conversion and tagging process for the 27 languages in CHILDES that have available UD grammars. The 10 languages that have UD grammars, but which have not yet been processed with UD are identified with asterisks. The other 27 have been either fully or partially tagged. These UD taggings represent first drafts that have not yet been checked by native speakers and which will surely require further fine-tuning and use of the MWT method described above. At this point, no further conversion work for Chatter validation will be needed for these 27 languages, and they can all go smoothly through future automatic analysis when new versions of UD have been fine-tuned for each language.

1. Afrikaans: Given its limited morphology and the limited use of eye-dialect in transcription, application of UD to Afrikaans went smoothly.
2. *Arabic: The two current Arabic corpora use a romanization which will have to be converted to Arabic script.
3. *Basque: There are no obvious barriers to application of UD to Basque, but guidance from native speakers would make the result more reliable.
4. *Bulgarian: The Bulgarian romanization must be converted back to Cyrillic. Unfortunately, there are conflicting standards for romanization of Bulgarian and many digraphs are ambiguous, so this conversion will require further analysis.
5. Cantonese: Because the Cantonese corpora were transcribed in Hanzi, no script conversion was necessary. In addition, UD for Chinese languages handles word-level tokenization directly, so there is no need to add or remove spaces between words.
6. Catalan: Processing of Catalan was straightforward.
7. Croatian: Processing of Croatian was straightforward.
8. Czech: Processing of Czech was straightforward. However, the contributors of the Czech corpus had already created a carefully done %mor analysis which they prefer to keep in place without the UD tags.
9. Danish: Processing of Danish was straightforward.
10. Dutch: Processing of Dutch was straightforward.
11. English: CHILDES English transcripts are now tagged using UD. However, programs such as KidEval, IPSyn, FluCalc, and DSS rely on MOR tagging. So, we also maintain a version of the English data that is tagged by MOR.
12. Estonian: Processing of Estonian was straightforward.
13. French: The French database is quite extensive. However, after much detailed word-level repair, processing went smoothly.
14. German: The German corpora required extensive revision of eye-dialect forms. Once that was done, processing went smoothly. UD did a much better job than the previous MOR in its assignment of case/number/gender roles to modifiers and nouns, as well as in creating an accurate %gra line. This is because MOR rules were not able to condition case/number/gender assignment to articles and modifiers based on the following noun, whereas the DNN (deep neural network) architecture of Stanza is able to use the full DP context to make these assignments.
15. *Greek: Processing of Greek will depend on creation of a method for converting from the romanization back to the Modern Greek alphabet.
16. *Hebrew: Hebrew has already been processed by a MOR grammar. However, UD processing of Hebrew will require conversion from romanization to Hebrew right-to-left script and we have not yet finalized a method for doing this.
17. *Hungarian: The current Hungarian transcripts make extensive use of eye-dialect and phonological forms. Once these are modified, processing should be straightforward.

18. Icelandic: Processing of Icelandic required extensive modification of eye-dialect forms that will need to be re-checked. Otherwise, analysis was straightforward.
19. *Indonesian: The huge size of the Indonesian corpus and the extensive use of eye-dialect will require a fair amount of work for this corpus.
20. Irish: Processing of Irish was straightforward.
21. Italian: Processing of Italian was straightforward. Because Italian had earlier been analyzed by MOR, there were few word level problems, except for dealing with separation of clitics by spaces.
22. Japanese: Processing of Japanese has represented a unique challenge because of the use of three orthographies (Kanzi, hiragana, katakana), attempts to represent words in a mix of orthographies, and difficulties with word segmentation. Two of the Japanese corpora have been tagged, but others will need further orthographic work and fine-tuning of the UD tagger for Japanese.
23. Korean: Korean involved no script transformation and processing went quite smoothly.
24. Mandarin: Because Mandarin had already been processed through MOR, there were few irregularities in the transcripts. Also, Mandarin involved no script transformation and processing went quite smoothly.
25. Norwegian: Processing of Norwegian was straightforward.
26. Polish: Processing of Polish was straightforward.
27. Portuguese: After some repair for clitics, proclitics, MWEs, and format, processing of Portuguese was straightforward.
28. Romanian: Processing of Romanian was straightforward.
29. *Russian: Like Bulgarian, Russian will need conversion of romanization to Cyrillic. However, the extensive use of eye-dialect and phonological forms in the Russian corpora will make this difficult.
30. Serbian: Serbian UD allows for Roman orthography. As a result, processing of Serbian was straightforward.
31. Slovenian: Processing of Slovenian was straightforward.
32. Spanish: Most of the Spanish corpora had earlier been analyzed by MOR. For those corpora, processing was straightforward. However, there are several Spanish corpora that will need further work for eye-dialect, phonological forms, and other divergences.
33. Swedish: Processing of the Andren corpus was straightforward. However, work with the Lund corpus will require treatment of eye-dialect and phonological forms.
34. *Tamil: Processing of the Tamil transcripts will require conversion of the romanization to Abugida orthography.
35. *Thai: Like many other Asian languages, Thai orthography does not include spacing, which makes tokenization difficult. Current Thai transcripts all use romanization and there is no clear path for conversion to Sukhothai script.
36. Turkish: Processing of Turkish was straightforward. However, because UD morphology is non-analytic, the %mor line fails to capture the agglutinative nature of

Turkish word formation. A similar problem arises with Hungarian and Estonian.

37. Welsh: Processing of Welsh was straightforward, even though there are many forms that involve apostrophes for omissions. Apparently, these forms are already accepted in standard Welsh in the training set for UD.

Formal evaluation of the success of this initial application of UD to the child corpora will require input from workers in each of these languages. So far, we have been receiving this type of corrective input for Spanish, French, and Japanese. During the coming years, we will emphasize the importance of checking by native speakers and refinement of the taggers based on their input. Refinement relies on three methods: direct revision of output forms, inclusion of MWT forms in the pipeline, and creation of training sets for fine-tuning.

The Stanza website at https://universaldependencies.org/conll18/evaluation.html provides LAS evaluation scores for each of the taggers we have used. LAS (labelled attachment score) is computed as the harmonic mean of precision (P, i.e. correctly labelled arcs over arcs labelled) and recall (R, i.e. labelled arcs over all gold-standard arcs) which is 2PR/P+R. For the languages we studied, this score ranges between .89 and .93. Although there is clearly room for improvement in these taggers, the results are all in the useable range. However, these numbers are based on adult spoken and written input. We have so far seen that, when UD is applied to English child language corpora, it does a better job than MOR for the adult input, particularly for grammatical relations. However, like MOR, it has problems with tagging utterances from children younger than 3-years-old. This is a fundamental problem in the study of the first stages of grammatical development.

UD Accuracy for English

To evaluate the accuracy of the morphological and grammatical tagging by UD, we examined the Batchalign UD output for three transcripts from Roger Brown's Sarah corpus of North American English. The files were 020613.cha with 2190 words, 030507.cha with 3344 words, and 050002.cha with 1636 words. These were selected to represent early, middle and late segments of the Sarah corpus.

For part-of-speech and morphological feature analysis on the %mor line, the only error was the incorrect analysis of the word *o'clock* for which UD requires the version without the apostrophe. There was also a non-optimal analysis of the form *out_of* with an underscore in the phrase *out of the cave* in 020613. The joining of the two words into one with an underscore had been done to improve accuracy in the earlier MOR grammar. However, UD creates a better analysis when the words are written separately. The general principle here is that the modifications to standard orthographic practice that were done to improve accuracy with MOR should be undone to improve UD tagging. We are now working on this type of word-level improvement.

For analysis of grammatical dependency tagging, we used the GraphViz function in CLAN which allows the user to triple-click on the %gra line to view a graph of the grammatical dependency analysis with labelled arcs. In these transcripts, even at the older age, the child only produces very short sentences. However, the adults produce many long sentences with complex structures and the relations in these utterances are uniformly linked and tagged correctly. One problem that we noted was the treatment of initial *see* as the ROOT in a sentence such as *see this is nearly ready to fall.* It would seem better to link *see* to the ROOT through the DISCOURSE relation. Other initial communicators such as *well* or *sure* are linked to the utterance through the DISCOURSE relation and it seems that this would be the appropriate analysis also for initial *see*. This analysis is further supported by the fact that, when *see* occurs finally, it does use the DISCOURSE link. Apart from this, we only noted three other errors. One involved a failure in transcription to place angle brackets around retraced material. The other two involved transcription of two utterances on a single main line. When two utterances are placed on a single line, the analysis provided by UD is correct, but CHAT guidelines prefer placement of each utterance onto its own main tier line. In summary, the taggings produced by UD for these English files were extremely accurate for part-of-speech, lexical features, and grammatical relations.

Morphosyntactic Analysis

Here we describe in further detail the application of the neural analysis models provided by the Stanza (Qi et al., 2020) system, along with the modifications we make for characteristics of spoken language, child language, and language-specific forms.

Word Tokenization

The first step of analysis involves tokenizing each utterance in the CHAT transcript into tokens. Because the CHAT format (https://talkbank.org/manuals/CHAT.pdf) encodes tokenization by using whitespace delineated token groups to identify words, tokenization is frequently given natively in the transcript. However, for some languages token representations have little to do with word-level representations. In Japanese child language, for instance, two of the language's three writing systems—hiragana and katakana—are moraic-based units frequently employed to transcribe a child during L1 development (Ota, 2015) while the third—kanji, often used for actual word representations needed for morphosyntactic analysis, have little to do with phonology. Moreover, Japanese is not written with spaces. Because of this, whitespace-delineated token representations are not a reliable source of information for word representations.

For languages which have this limitation—and in particular, for our analysis of Japanese—we employ the more complex token segmentation scheme given in Stanza which involves formulating word-level tokenization as a token labeling task—ignoring any transcribed tokenizations and labeling each input *character* as belonging to the start, middle, or end of a token—before further processing each resulting "token group" via the downstream, semantic aware modules such as the Stanza lemmatizer. For instance, consider the Japanese phrase *karuto dantai* “cult group”:

カルト団体

The DNN tagger would first treat all constituent forms as separate and assign to each one a beginning and inside tags representing word boundaries. This creates the sequence:

B I I B I

Finally, separating the forms following the B tags, we obtain:

[カルト] [団体]

as the final word tokenizations, which we place back into the CHAT file as space-delimited tokens as follows:

カルト 団体

In this way, we recover a canonical tokenization for those particular languages based on the annotation style chosen by the working group of the target language in UD annotation; for Japanese, for instance, this may include some resulting orthographic Kanji formed by joining tokens from other syllabaries following the short-unit word (SUW) style (Den et al., 2008). We then use this canonical tokenization to "retokenize" the original CHAT transcript with this new tokenization. Once this initial re-tokenization is obtained, we can then proceed to the remaining analysis by the pipeline describe here.

Multi-Word Token and Form Correction

UD (De Marneffe et al., 2021) distinguishes between tokens—continuous character spans without internal delineation—and syntactic words used in analysis. This distinction is particularly relevant with respect to the treatment of multi-word tokens (MWTs)—a single continuous text span which contains multiple syntactic words, each with individual features and dependencies which need to be analyzed independently. Augmenting Stanza’s neural-only analysis, we use a lexicon and orthography driven approach to identify and expand three types of such MWTs.

Two types of such MWTs are usually automatically recognized by Stanza through the same tokenization procedure described in the section above: clitics and contractions. Clitics are independent syntactical forms attached to other words, such as in Spanish *despertarme* (*despertar* + *me*)—with the latter being a separate syntactic word which modifies the previous word which needs to be analyzed independently (i.e. modifying that I am who woke the object up); contractions are combinations of multiple words into one token, such as in English *I'm* (*I* + *am*).

If clitics and contractions are not automatically expanded by Stanza, we use a rules-based analysis of orthography to detect some of these common forms and manually expand them. This functionality is currently supported for detection of subject contractions in French and Italian (i.e. *t'aime* to *te* + *aime*), prepositional contractions (i.e. *jusqu'ici* to *jusque* + *ici*), and be-contractions in English (i.e. *you're* to *you* + *are*).

The third type of MWT not typically expanded by Stanza are single-unit, multi-word forms which are usually joined by an underscore in the CHAT transcription format, because they are a single semantic form and multiple syntactic words. For instance, the form *pirates_des_Caraïbes* (Pirates of the Caribbean) is one such form, broken into *pirates des Caraïbes*. Our pipeline uses a lexicon to detect and expand these forms. We implement this correction functionality as a custom step in the Stanza analysis pipeline which takes the "draft" tokenizations from Stanza as input and returns the correct tokenization and word expansions to downstream analysis functions in Stanza—ensuring that POS, GFs, and GRs will be analyzed on the corrected word.

Additionally, the neural tokenizer in Stanza will occasionally mark forms as MWTs when they are simply single-token single-word forms with a punctuation mark inside (i.e. the French word *aujourd'hui*); in those cases, we perform the opposite correction forcing Stanza to treat the resulting token a single word instead of an MWT. These cases are identified and corrected using a lexicon as well.

In final output into the CHAT transcription format, we follow the convention set forth by the CLAN MOR/MEGRASP system (MacWhinney et al., 2012) and join the morphology analyses of multi-word tokens together with a tilde (~), maintaining token-level alignment between the transcript and analysis yet being able to encode multiple words within a token.

Morphology and Dependency Analysis

After tokenization and MWT correction, we make no further adjustments to the Stanza morphology, dependency, and feature analysis of each language and simply run the remaining Stanza analysis pipeline with the corrected tokens. Because most Stanza models are trained via the Universal Dependencies dataset, some datasets, such as UD Dutch Alpino (Bouma et al., 2001), will be rich in annotated feature information whereas some others, such as UD Japanese GSD (Nivre et al., 2020), will have little to no GFs annotated. For Japanese, this is true in part because many of the GRs are expressed in separate morphology. Our UD analysis, therefore, carries the design choices of analysis made within these gold datasets. Once this information on POS, GFs, and GRs has been annotated by the Stanza system, we proceed to perform morphology-dependent extraction and correction of the resulting features as a final processing step.

Morphosyntactic Transcription and Feature Correction

After analysis by Stanza, we output the extracted GFs using an annotation format very similar to the one used in the MOR/MEGRASP system (described further in https://talkbank.org/manuals/mor.pdf) for the *%mor* and *%gra* lines in CHAT. Our overarching goal is to report the *maximal set* of GFs which 1) can be reported for each language and 2) provide additional information beyond the "default" case. In accord with these principles, the GFs for aspect, mood, tense, polarity, clusivity, case, type, degree, conjugation (form), and politeness are reported exactly as in the UD annotation specifications. Gender is reported for all tagged genders except "common neutral" (ComNeut); and number is reported for all except singular. For personhood, fourth and zeroth person are both reported as "fourth person". As in MOR, GFs are joined after the lemma by using a dash "-" and contractions and clitics are marked with ~, as in the earlier MOR standard.

Dependency Structure

In addition to creating a %mor line with its analysis of POS and GFs, Batchalign2 also produces a %gra line that encodes the GRs (grammatical relations) for each utterance. The creation of this GR analysis is the primary goal of the Universal Dependencies project. The encoding involves a directed acyclic graph in which words are connected through unidirectional arcs from the dependent word to its head. Each arc is labeled with a grammatical relation tag taken from the list summarized earlier. Using the GraphViz web service (https://github.com/xflr6/graphviz), one can double-click on a %gra line to produce a display such as the screenshot in Figure 2 which comes from a parental utterance in the Brown/Eve/020000b.cha file on line 44.

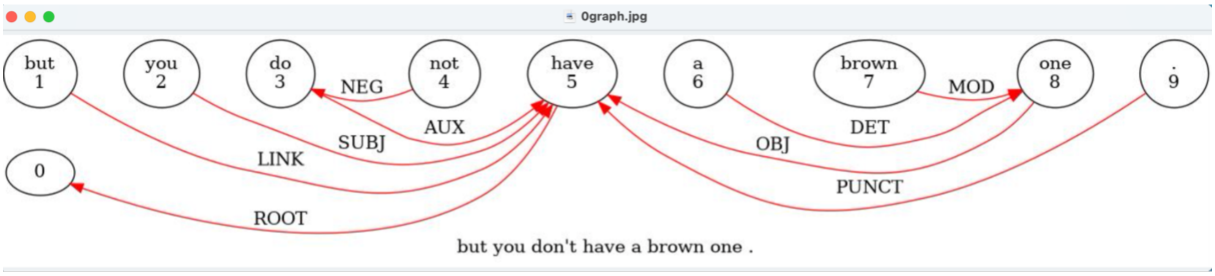


Figure 2. ***Dependency analysis by UD for an example utterance.***

This graph derives from processing of this utterance:

*MOT: but you don't have a brown one.
%mor: cconj|but pron|you-Prs-Nom-S2 aux|do-Fin-Ind-Pres-S2~part|not verb|have-Inf-S det|a-Ind-Art adj|brown-Pos-S1 noun|one.
%gra: 1|5|CC 2|5|NSUBJ 3|5|AUX 4|5|ADVMOD 5|8|ROOT 6|8|DET 7|8|AMOD 8|5|OBJ 9|5|PUNCT

In the %gra line, each word has two numbers and a GR. The first number is its serial position in the utterance and the second is the position of the word to which it is linked through a GR. After the two numbers comes the label on the GR. In Figure 1, for example, we see that the word *one* links to the verb *have* through the OBJ relation, that the word *brown* links to *one* through the MOD relation, and so on. This form of display is essentially the same as what was produced by MEGRASP (Figure 3), although the labels on the arcs are changed and in UD the word *not* is linked to the auxiliary *do* rather than directly to the verb.

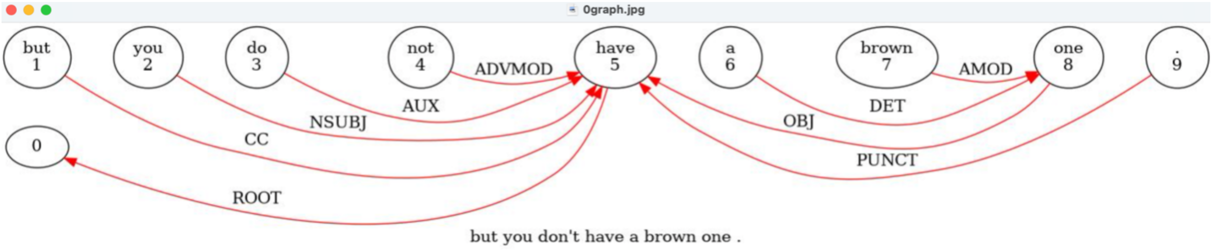


Figure 3. ***Dependency analysis by MEGRASP for an example utterance.***

Processing based on UD Analysis

Having tagged corpora in 27 of the languages in CHILDES for POS, GFs, and GRs, we are able to apply many of the TalkBank analytic tools that were earlier available only for English. This opportunity can go a long way toward reducing the WEIRD emphasis in child language studies. Most of these tools and frameworks will work directly, but some require further configuration. We can now use them to compute indices and profiles for the three data formats discussed earlier: longitudinal case studies, cross-sectional group studies, and crosslinguistic comparisons. In other words, having this morphosyntactic information available for all 27 languages benefits not only cross-linguistic comparison, but also the language-internal examination of development for individual children and clinically important comparison groups within each language. The tools that are available now or which will soon be available include:

1. Basic analysis commands: Researchers could make use of the 26 basic analysis

commands in CLAN on all languages prior to running of Batchalign2. However, because most of the languages previously had no %mor or %gra line, analyses were limited to the main speech tier. Now these same programs can run on these additional lines, making additional types of analyses possible.

2. KIDEVAL: This command combines 57 CLAN analyses into a single package. It includes tracking of the most common GFs in each language, repetitions, vocabulary diversity, error types, MLU (mean length of utterance), and other indicators. In a single command, KIDEVAL can be run on a single transcript or a whole folder of transcripts. It gives both the results for each child on each measure as well as a z-score for the extent to which the child matches a larger comparison group for that measure. The comparison group can be selected for age in 6-month intervals, participant type (TD, DLD, ASD, etc.) and recording type (narrative, free play, elicited). For this comparison to be meaningful, KIDEVAL needs a comparison sample of at least 25 cases. This is currently possible for Dutch, English, French, Japanese, Mandarin, and Spanish. Construction of comparison corpora for other languages that have sufficient comparison data is in progress.
3. DSS: DSS (Developmental Sentence Score) (Lee, 1974) is a profiling method that focuses on early learning of grammatical morphology and basic syntax in English. Given the new availability of a consistent set of POS, GF and GR tags, it will now be much easier to configure versions of DSS for additional languages.
4. IPSyn: IPSyn (the Index of Productive Syntax) (Scarborough, 1990) is similar to DSS. However, it includes measures of more advanced syntactic structures. Building on recent analyses (MacWhinney et al., 2020; Yang et al., 2021) we can create streamlined, automatic versions of IPSyn for multiple languages.
5. Vocabulary diversity: CLAN provides four measures of vocabulary diversity: TTR (type token ratio), NDW (number of different words), MATTR (moving average type token ratio) (Covington & McFall, 2010), and vocD (Malvern & Richards, 1997). Analysis through MATTR and vocD requires use of lemmas on the %mor line which is now possible across the 27 languages to which UD has been applied.
6. GF analysis: Although a basic level of GF analysis is built into KIDEVAL, there are many types of crosslinguistic analysis that will be best conducted using programs like FREQ on the %mor line across languages. For example, we can now look consistently at learning of tense marking across all these languages and observe how that feature is acquired in comparison with other features.
7. GR analysis: It is now possible to use GraphViz to visualize the syntactic structure for all 27 languages. In addition, Section 7.9.14 of the CLAN manual describes how to use FREQ with the UD %gra line to study the emergence of more complex relations, such as xcomp (a clausal complement without its own subject) or expl:pass (a reflexive marker of a middle or passive clause), as well as combinations of GRs.
8. Cross-tier analysis: We are currently building a new program called FLUPOS

for tracking features across multiple coding tiers, including the main line, %mor, %gra, and the %pho line for phonology. One particularly important application of FLUPOS will be to determine the degree to which disfluencies are proportionally higher with certain lexical, morphological, phonological, and syntactic configurations.

The combination of these new %mor and %gra tiers for these 27 languages, along with current analytic methods and ones we plan to build will provide us with a stronger quantitative foundation for crosslinguistic analysis of language development. We will be able to track the impact of language structure and input on the development of lexicon, morphology, and syntax in a set of languages that goes well beyond the limits of data from only WEIRD participants.

References

Bellman, R. (1966). Dynamic programming. *Science*, *153*(3731), 34-37. PMC

Berndt, D. J., & Clifford, J. (1994). Using dynamic time warping to find patterns in time series. Proceedings of the 3rd international conference on knowledge discovery and data mining,

Bernstein Ratner, N., & MacWhinney, B. (2020). TalkBank resources for psycholinguistic analysis and clinical practice. In A. Pareja-Lora, M. Blume, & B. Lust (Eds.), *Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences* (pp. 131-150). MIT Press. https://psyling.talkbank.org/years/2018/RatnerMacW.pdf

Bernstein Ratner, N., & MacWhinney, B. (2023). Assessment and therapy goal planning using free computerized language analysis software. *Perspectives of the ASHA Special Interest Groups*, *8*(1), 19-31. https://doi.org/10.1044/2022_PERSP-22-00156 PMC

Bishop, D. (1982). *The test of reception of grammar*. Medical Research Council.

Bloch, O. (1921). Les premiers stades du langage de l'enfant. *Journal de Psychologie*, *18*, 693-712. PMC
Bouma, G., Van Noord, G., & Malouf, R. (2001). Alpino: Wide-coverage computational analysis of Dutch. In *Computational linguistics in the Netherlands 2000* (pp. 45-59). Brill.

Bredin, H. (2023). pyannote. audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. 24th INTERSPEECH Conference (INTERSPEECH 2023), Brown, R. (1973). *A first language: The early stages*. Harvard University Press. https://doi.org/10.4159/harvard.9780674732469

Brownrigg, D. R. (1984). The weighted median filter. *Communications of the ACM, 27*(8), 807-818. PMC

Chang, C.-j., & McCabe, A. (2013). Evaluation in Mandarin Chinese children's personal narratives. *Studies in Narrative (SiN)*. PMC

Chao, Y. R. (1951). The Cantian ideolect: An analysis of the Chinese spoken by a twenty-eight-months-old child. *University of California Publications in Semitic Philolology, 1,* 27-44. PMC

Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics, 17*(2), 94-100. https://doi.org/10.1080/02687038.2012.693584 PMC PMC4569132

Crystal, D., Fletcher, P., & Garman, M. (1989). *The grammatical analysis of language disability. Second Edition*. Cole and Whurr.

Dale, P. S., & Fenson, L. (1996). Lexical development norms for young children. *Behavior Research Methods, Instruments, and Computers, 28*. PMC

De Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal dependencies. *Computational Linguistics, 47*(2), 255-308. PMC

Del Rio, M., Ha, P., McNamara, Q., Miller, C., & Chandra, S. (2022). Earnings-22: A practical benchmark for accents in the wild. *arXiv preprint arXiv:2203.15591*. PMC

Den, Y., Nakamura, J., Ogiso, T., & Ogura, H. (2008). A Proper Approach to Japanese Morphological Analysis: Dictionary, Model, and Evaluation. LREC,

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. PMC

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability and consistency in early language learning: The Wordbank project*. MIT Press.

Fraser, K. C., Ben-David, N., Hirst, G., Graham, N., & Rochon, E. (2015). Sentence segmentation of aphasic speech. Proceedings of the 2015 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies, Garbarino, J., Bernstein Ratner, N., & MacWhinney, B. (2020). Use of computerized language analysis to assess child language. *Language, Speech, and Hearing Services in Schools, 51*(2), 504-506. https://doi.org/10.1044/2020_LSHSS-19-00118 PMC 7225019

Goldman, R., & Fristoe, M. (2000). *The Goldman-Fristoe Test of Articulation - 2*. Pearson Assessments.

Guillaume, P. (1927). Les débuts de la phrase dans le langage de l'enfant. *Journal de Psychologie, 24,* 1-25. PMC

Gvozdev, A. N. (1949). *Formirovaniye u rebenka grammaticheskogo stroya*. Akademija Pedagogika Nauk RSFSR.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2-3), 61-83. PMC

Hou, R., Chang, H., Ma, B., Shan, S., & Chen, X. (2019). Cross attention network for few-shot classification. *Advances in Neural Information Processing Systems, 32*. PMC

Karniol, R. (2010). *Social development as preference management: How infants, children, and parents get what they want from one another*. Cambridge University Press.
Kenyeres, E. (1926). *A gyermek elsö szavai es a szófajók föllépése*. Kisdednevelés.

Lee, L. (1974). *Developmental Sentence Analysis*. Northwestern University Press.

Leopold, W. (1939). *Speech development of a bilingual child: a linguist's record: Vol. 1. Vocabulary growth in the first two years* (Vol. 1). Northwestern University Press.

Leopold, W. (1947). *Speech development of a bilingual child: a linguist's record: Vol. 2. Sound-learning in the first two years*. Northwestern University Press.

Leopold, W. (1949a). *Speech development of a bilingual child: a linguist's record: Vol. 3. Grammar and general problems in the first two years*. Northwestern University Press.

Leopold, W. (1949b). *Speech development of a bilingual child: a linguist's record: Vol. 4. Diary from age 2*. Northwestern University Press.

Levenshtein, V. I. (1965). Binary codes capable of correcting deletions, insertions, and reversals. *Doklady Akademii Nauk SSSR, 4,* 845-848. PMC

Li, L., & Zhou, J. (2011). *Preschool children's development reading comprehension of picture storybook: from a perspective of multimodal meaning making.* East China Normal University]. Shanghai.

Liu, H., MacWhinney, B., Fromm, D., & Lanzi, A. (2023). Automation of language sample analysis. *Journal of Speech, Language, and Hearing Research, 66,* 2421-2433. https://doi.org/10.1044/2023_JSLHR-22-00642 PMC

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk. 3rd edition.* Lawrence Erlbaum Associates. https://www.amazon.com

MacWhinney, B. (2008). Enriching CHILDES for morphosyntactic analysis. In H. Behrens (Ed.)*, Trends in corpus research: Finding structure in data* (pp. 165-198). John Benjamins. https://psyling.talkbank.org/years/2008/morphosyntax.pdf

MacWhinney, B., & Fromm, D. (2022). Language sample analysis with TalkBank: An update and review. *Frontiers in Communication, 7,* 865498. https://doi.org/10.3389/fcomm.2022.865498 PMC

MacWhinney, B., Roberts, J., Altenberg, E., & Hunter, M. (2020). Improving automatic IPSyn coding. *Language, Speech, and Hearing Services in Schools, 51*(4), 1187-1189. https://doi.org/10.1044/2020_LSHSS-20-00090 PMC 7842849

MacWhinney, B., Spektor, L., Chen, F., & Rose, Y. (2012). Best practices in the TalkBank framework. 8th International Conference on Language Resources and Evaluation (LREC), Istanbul. https://psyling.talkbank.org/years/2012/LREC-best.pdf

Malvern, D., & Richards, B. (1997). A new measure of lexical diversity. In A. Ryan & A. Wray (Eds.), *Evolving models of language* (pp. 58-71). Multilingual Matters.

Moreno, P. J., Joerg, C. F., Van Thong, J.-M., & Glickman, O. (1998). A recursive algorithm for the forced alignment of very long audio segments. ICSLP,

Niu, Z., Zhong, G., & Yu, H. (2021). A review on the attention mechanism of deep learning. *Neurocomputing, 452,* 48-62. PMC

Nivre, J., De Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., & Silveira, N. (2016). Universal dependencies v1: A multilingual treebank collection. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), Nivre, J., De Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F., & Zeman, D. (2020). Universal Dependencies v2: An evergrowing multilingual treebank collection. *arXiv preprint arXiv:2004.10643*. PMC

Ota, M. (2015). L1 phonology: phonological development. *The handbook of Japanese language and linguistics: Phonetics and phonology,* 681-717. PMC

Parisse, C., & Le Normand, M.-T. (2000). Automatic disambiguation of the morphosyntax in spoken language corpora. *Behavior Research Methods, Instruments, and Computers, 32*(3), 468-481. https://doi.org/10.3758/bf03200818 PMC
Pavlovitch, M. (1920). *Le langage enfantin: Acquisition du serbe et du francais par un enfant serbe*. Champion.

Ponori, T. E. (1871). A gyermeknyelvről. *Természettudományi Közlöny, 3,* 117-125. PMC

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations,

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. International Conference on Machine Learning,

Römer, U. (2019). *MICASE: Michigan corpus of academic spoken English.* https://doi.org/doi:10.21415/QT9V-2J96

Sagae, K., Davis, E., Lavie, A., MacWhinney, B., & Wintner, S. (2010). Morphosyntactic annotation of CHILDES transcripts. *Journal of Child Language, 37*(3), 705-729. https://doi.org/10.1017/S0305000909990407 PMC 4048841

Scarborough, H. (1990). Index of Productive Syntax. *Applied Psycholinguistics, 11*(1), 1-22. https://doi.org/10.1017/S0142716400008262 PMC

Slobin, D. (1985). *The crosslinguistic study of language acquisition. Volume 1: The data.* Lawrence Erlbaum Associates.

Smith, N. V. (1973). *The acquisition of phonology: A case study*. Cambridge University Press.

Smoczynska, M. (2017). The acquisition of Polish. In *The crosslinguistic study of language acquisition* (pp. 595-686). Psychology Press.

Stern, C., & Stern, W. (1907). *Die Kindersprache*. Barth.
Szuman, S. (1959). O wypowiedzianej oraz domyślnej treści wypowiedzi dziecka z pierwszychlat jego życia. *Przegląd Psychologiczny*. PMC

Tamis-LeMonda, C. S., Kachergis, G., Masek, L. R., Gonzalez, S. L., Soska, K. C., Herzberg, O., Xu, M., Adolph, K. E., Gilmore, R. O., & Bornstein, M. H. (2024). Comparing apples to manzanas and oranges to naranjas: A new measure of English-Spanish vocabulary for dual language learners. *Infancy*. https://doi.org/10.1111/infa.12571 PMC

Tomasello, M. (1992). *First verbs: A case study of early grammatical development*. Cambridge University Press.

Yang, J. S., MacWhinney, B., & Ratner, N. B. (2021). The Index of Productive Syntax: Psychometric Properties and Suggested Modifications. *American Journal of Speech-Language Pathology*, 1-18. PMC 9135028.

Data, code and materials availability statement

Batchalign2 is available from https://github.com/talkbank/batchalign2. The utterance segmentation models trained using the manner described in (Liu et al. 2023), as well as fine-tuned, language specific Whisper models, are additionally available for US English and Mandarin; these models are available at https://huggingface.co/talkbank. The tagged corpora for the 27 languages discussed here are available from https://childes.talkbank.org. All the corpora in CHILDES are open-access and have associated DOIs.

Authorship and Contributorship Statement

Houjun Liu is the author of the Batchalign2 program and Brian MacWhinney applied the program to analyze corpora in 27 languages. Both authors shared in the conception of the work, writing of this report, and approval of the final version. Both ensure that questions related to the accuracy or integrity of any part of the work will be appropriately investigated and resolved.

License

A novel corpus of naturalistic picture book reading with 2 to 3 year old children

Anastasia Stoops
Mengqian Wu
In Ho Ted Jung
Jessica L. Montag

University of Illinois Urbana-Champaign, USA

Abstract: Substantial literature suggests that reading to children is positively associated with language outcomes, but the causal pathways are less well understood. One possibility is that reading to children promotes language input that is particularly useful for some aspects of language learning. To better understand the language that is produced during picture book reading, we built a sharable corpus of caregiver-child interactions during book reading recorded in homes. Caregivers overwhelmingly read the book text. However, books varied in the language they generated, with some books promoting more conversational turns and extra-textual language, while others promoted more overall words, unique words, and longer utterances. Relative to other conversational contexts, books generally generated overall more words, more lexically diverse talk, and longer utterances. We see different profiles of language generated during book reading that are all plausibly linked with language skills. If a causal pathway exists between shared book reading and language outcomes, a sensible candidate may be that reading provides a varied range of linguistic experiences.

Keywords: Naturalistic language environments; Picture book reading; lexical diversity; Syntactic complexity.

Corresponding author(s): Anastasia Stoops, Department of Psychology, University of Illinois Urbana-Champaign, USA. Email: agusico2@illinois.edu.

ORCID ID(s): Anastasia Stoops https://orcid.org/0000-0002-0807-8377
Jessica L. Montag https://orcid.org/0000-0001-9446-1016

Citation: Stoops, A., Wu, M., Jung, I.H.T., and Montag, J.L. (2024). A novel corpus of naturalistic picture book reading with 2 to 2 year old children. *Language Development Research, 4*(1), 259—297. http://doi.org/10.34842/3kz6-4s17

Introduction

Caregivers often read to young children, including those who cannot yet read themselves. This behaviour, called shared book reading, is common in many, but not all, households and cultures around the world. Advice that caregivers should read to children is everywhere, particularly in western societies, where reading to children is popularly associated with a range of benefits.

The recommendation to read to children is not only in the popular culture; there is in fact substantial literature suggesting that reading to children is positively associated with language outcomes. Research has shown positive effects of book reading on a wide range of early language skills, including child's receptive and expressive language skills (e.g. Arterberry, et al., 2007; Demir-Lira, et al., 2019; Farrant & Zubrick, 2011; Fletcher & Reese, 2005; Karrass & Braungart- Rieker, 2005; Mol & Newman, 2014; Ninio, 1983; Payne, et al., 1994; Senechal & LeFevre, 2002) and subsequent literacy skills (e.g. Bus, et al., 1995; Deckner, et al., 2006; Dickinson & Tabors, 1991; Lonigan, et al., 2000; Scarborough, et al., 1991; Shahaeian et al., 2018). While scientific studies of reading have correlated exposure to picture books with positive language development outcomes, the causal pathways are less well understood. For recommendations surrounding reading to children to be in line with the scientific evidence, we must better understand the pathways by which reading to young children comes to be associated with improved language outcomes.

The goal of this study is to build a corpus of parent-child interactions during book reading sessions recorded in homes. We quantify features of the language that appears in these recordings, with an emphasis on understanding the independent contributions of caregiver-child conversation and book text read aloud. With this information, we can begin to understand the linguistic experiences that book reading may provide that might plausibly explain the associations with positive language outcomes.

One of the important reasons to better understand the unique linguistic (or other) experiences that shared book reading may provide is to establish whether there is a plausible pathway between shared book reading and language outcomes at all. Many studies fail to replicate the language-boosting effects of picture book reading, find only small effects (Davies et al., 2020; Noble et al., 2020; Noble et al., 2019; Sala & Gobet, 2017; Simons et al., 2016) or note problems with the generalizability of existing findings to lower-income or other marginalized families (Manz et al., 2010; Mol et al., 2008). One potential explanation is that correlations between shared book reading and language outcomes may reflect other factors that are causally independent of book reading. For example, caregivers who read more often to children are more likely to be white and tend to be wealthier than those who read less frequently

(Bradley, et al., 2001; Raikes et al., 2006; Yarosz & Barnett, 2001; Young, et al., 1998), so the effects of book reading on language outcomes may be attributable to other factors. We must remain open to the null hypothesis that the positive language outcomes might in fact be associated with the numerous other benefits that wealth and status impart on children. An investigation of the effect of shared book reading proper on language outcomes requires that the field have clearer explanations for why more frequent book reading might be associated with positive language outcomes.

When caregivers read to young children, interactions are not limited to simply reading the text of the book; caregivers and children engage in conversation as well. Caregivers may point to and label pictures, paraphrase text, comment and expand upon the text, ask and answer questions, and engage in other extra-textual speech (Deckner et al., 2006; Fletcher, et al., 2008; Kam & Matthewson, 2017; Ninio & Bruner, 1978; Whitehurst et al., 1988; see Read, et al., 2023 for a recent review). This conversation is often investigated as a source or language input for young children that is particularly useful for language learning (Demir-Lira, et al., 2019; Fletcher & Reese, 2005; Fletcher, et al., 2008; Hindman, et al., 2014; Justice & Ezell, 2000; Mol et al., 2008; Muhinyi & Rowe, 2019). One of the most highly investigated features of caregiver-child conversation during shared book reading is that it tends to contain more conversational turn-taking than other contexts of child-caregiver speech (Gilkerson et al., 2017; Sosa, 2016). A large research literature has identified frequent back-and-forth conversational turn taking, as a type of linguistic experience that is positively associated with language outcomes (Donnelly & Kidd, 2021; Gilkerson et al., 2018; Romeo et al., 2018), and shared book reading may be a particularly dense source of these conversational turns (Gilkerson et al., 2017). In addition to being a source of conversational turns, the speech itself may consist of more unique words and longer sentences (Crain-Thoreson, et al., 2001; Hoff-Ginsberg, 1991; Muhinyi, et al., 2020; Whitehurst et al., 1988) than other caregiver-child activities, such as free-play (Gilkerson et al., 2017; Sosa, 2016). The spontaneous conversation produced during shared book reading may contain many features that make it particularly useful for language learning.

Following the hypothesis that shared book reading may promote caregiver-child conversation, many interventions that aim to use shared book reading to improve language outcomes frequently target extra-textual talk. These interventions include dialogic reading, in which caregivers are encouraged to ask open-ended questions during book reading that encourage children to verbally respond so that the child becomes a more active participant in the book reading activity (Arnold, et al., 1994; Whitehurst et al., 1988). During dialogic reading, caregivers are instructed to use language that aims to elicit speech from children, i.e. rephrasing child's utterances with same/different voice or sentence structure, and asking open-ended questions, which are features of caregiver speech that have been shown to support child language development in other conversational contexts (Baker & Nelson, 1984; Cleave et al., 2015;

Farrar, 1990; Girolametto & Weitzman, 2002; Huttenlocher et al., 2010; Nelson, 1977). Dialogic reading is associated with gains in expressive language skills (Chacko et al., 2018; Lonigan & Whitehurst, 1998; Valdez- Menchaca & Whitehurst, 1992; Whitehurst, et al., 1994; Whitehurst et al., 1999), perhaps more than other reading methods (Flack, et al., 2018). Caregiver-child conversation may be a particularly useful source of language input, and if it is indeed associated with shared book reading, may indicate a plausible pathway by which book reading comes to be positively associated with language outcomes.

Another hypothesized pathway by which shared book reading may influence language outcomes is through exposure to the book text itself. When caregivers read aloud the text of picture books, they may be exposing children to unique words and sentences, including complex syntax, that might otherwise be rare. Picture books are well established to be more lexically diverse than other types of linguistic input that children may encounter. For example, picture books contain more unique words than child-directed speech (Hayes & Ahrens, 1988; Massaro, 2015, 2017; Montag et al., 2015). In line with these findings, recordings of caregivers and children interacting in book reading contexts indicate that linguistic input from shared book reading may be more lexically sophisticated than that of other contexts (Crain-Thoreson, et al., 2001; Salo, et al., 2016; Sosa, 2016; Weizman & Snow, 2001). However, these studies do not explicitly distinguish between book text read aloud and extra-textual talk, so the increase in lexical diversity could be attributed to book text read aloud or to other sources of speech, for example, labelling or talking about pictures.

In addition to the inventories of words, the text of picture books contains more instances of complex sentence structure than other sources of child-directed or child-available (speech that is produced in the vicinity of the child even if it may not be explicitly child-directed) speech. Studies that compare syntactic constructions present in picture books and child-directed speech find that picture books contain a variety of language structures that are rare in typical child-directed speech. For example, Cameron-Faulkner and Noble (2013) compared syntactic constructions from 20 best-selling picture books in the UK with a sample of British English child directed speech. Picture books contained more complete sentences (e.g. The boy ate a doughnut; The bat is flying) than child-directed speech, which tended to contain more fragments (e.g. on the table), commands (e.g. put it down), and copulas (e.g. It's very heavy; That's nice). The authors suggested that the picture book language could be an important input source for the development of both common linguistic constructions but also complex constructions that might be rare in child-directed speech. Likewise, Montag (2019) focused on American English and compared the frequencies of a set of complex syntactic constructions in a corpus of 100 picture books with a sample of child-directed speech from the CHILDES corpus (MacWhinney, 2000). The text in picture books had significantly higher frequencies of complex syntactic constructions including passive sentences and sentences containing relative clauses than child-directed speech. Similarly, Hsiao, Dawson, Banerji and Nation (2022) found that complex sentences such as those containing relative clauses are more frequent in child-directed written than spoken corpora, and that frequencies increased as the target age of the child increased. Such findings suggest that picture books could affect language development outcomes by exposing children to types of complex language that might be otherwise sparse in the child-directed input.

To complicate an investigation of the language input that picture books provide, picture books vary wildly in types of stories they tell, and the linguistic and visual formats in which they tell these stories. There is no reason that all books might provide similar linguistic input, or even vary from other sources of child-directed speech along similar dimensions. Different books promote different profiles of language input as a consequence of the story's genre and plot complexity (Price, et al., 2009; Leech & Rowe, 2014; Muhinyi et al., 2019; Read, et al., 2014; Saracho, 2017). Given the enormous variability across picture books, there may not be a single profile of book reading talk, but rather different types of books may promote different profiles of speech. For example, book genre such as fiction versus non-fiction books tend to elicit different profiles of speech from caregivers and children, with non-fiction books often eliciting more frequent and more lexically complex extra-text utterances (Anderson et al, 2004; Price et al., 2009, Weitzman & Snow, 2001). Likewise, Muhinyi et al. (2019) found that complex stories with false beliefs central to the plot elicited longer and more lexically complex caregiver utterances. Book format seems to matter as well, with wordless picture eliciting more caregiver-child conversation (Senechal, Cornell & Broda, 1995) and chapter books (versus picture books) eliciting less extra-text discussion from children (Leech & Rowe, 2014). In an entirely different vein, Read and colleagues (2014) found that caregivers' prosody varied when reading a rhymed than a non-rhymed version of the same animal story. Understanding variability across book types is necessary for developing a more complete picture of the language generated during shared book reading.

To better understand the language generated during naturalistic home book reading, how this input might be different from other sources of child-directed speech, and how this input might vary based on features of the book being read, we created a corpus of recordings of picture book reading sessions made by families in their own homes. We provided parents with 4 books that varied in the amount of text they contained and the syntactic complexity of that text. The full transcripts are available to other researchers as a book reading corpus through the CHILDES online repository (https://childes.talkbank.org/access/Eng-NA/StoopsMontag.html).

We first describe features of the language generated during book reading, and how different books elicited different profiles of speech. Specifically, we expect that different books should generate different profiles of child and caregiver speech, with some books generating more back-and-forth conversation and others more silent listening of the books. We then compare the language input generated during home book reading sessions with other sources of child-directed speech for age-matched children to understand similarities and differences across shared book reading and other contexts of child-directed speech. In line with other studies of shared book reading, we expect quantitative differences in various aspects of the speech generated during picture book reading and other conversational contexts.

Method

Participants

Families were recruited from the area surrounding the University of Illinois, Urbana-Champaign. The study was approved by the Institutional Review Board and all families gave their informed consent prior to the inclusion in the study.

Caregivers. Twelve families participated in the study. Family demographic information is included in Table 1. For all families, English was the primary language spoken in the home, Education is reported for 24 caregivers because all 12 families were two-parent households.

Table 1. ***Parent demographics***

Demographic Categories	Count or Mean (range)	
Race:	Both Parents White	7
	Both Parents Asian	3
	Asian-White	2
Education:	PhD	6
	MA	5
	BA	7
	AS	6
Income:	$200,000+	1
	$100,000-$200,000	4
	$75,000-$100,000	4
	$25,000-$75,000	3
# children's books at home:	150 (50-200)	
# non-children's books at home:	200 (50-1000)	

Children. The average age of the 12 children included in the study was 31 months (range: 27-37mo; 7 girls, 5 boys). The mean parent reported MLU computed from the MBCDI was 8.3 (range: 3.6-12) and the average MBCDI score was 486 (range 56-675). One child was diagnosed with speech delays and had been receiving speech therapy.

Materials

Four books were selected that varied along two dimensions: Book length, as quantified by the number of words in the book text, and the syntactic complexity of the book, as indexed by the number of a subset of rare sentence types: passive sentences and sentences containing relative clauses. Word counts ranged from 125 words, a book with a few words or one-to-two sentences every few pages, to 1211 words, a more complex narrative book with 4 or more sentences on each page. Rare or complex sentence counts ranged from 0-15. Word counts and rare/complex construction counts are shown in Table 2, as well as the number of reading sessions recorded of each book across all families. Examples of each complex sentence type are shown in Table 3. Audio recordings were made with the OLYMPUS VN-541PC digital audio recorder.

Table 2. ***New Book Summary***

Book Title	Word Count	Count of Complex Syntactic Constructions				# Recordings	# Families
		Subject Relative Clause	Object Relative Clause	Oblique Relative Clause	Passive Main Clause		
That is Not a Good Idea (Mo Willems)	125	0	0	0	0	21	10
When Dinosaurs Came with Everything (Elise Broach)	1018	0	1	1	0	18	10
Stellaluna (Janell Cannon)	1211	2	1	0	0	12	8
Oh the Places You'll Go! (Dr. Seuss)	939	5	4	4	2	9	8

Table 3. ***Syntactic Complexity Summary***

Syntactic Construction	Example
Subject Relative Clause:	More bats gathered around to see the strange young bat **who behaved like a bird**. (from "Stellaluna")
Object Relative Clause:	**The next thing I knew**, she had him cleaning the gutters (from "When dinosaurs came with everything")
Oblique Relative Clause:	**The places you'll go**!; You will come to a place where the streets are not marked. (from "Oh the places you'll go!")
Passive Main Clause:	**You'll be left in a Lurch** (from "Oh the places you'll go!")

Caregivers were asked to select three out of the four books that they did not own that their child was not familiar with, i.e., they have not read to their child before. Caregivers were asked to choose three books because we expected some families would have familiarity with some of the books and we wanted to keep the number of books constant across families. Further, we wanted to keep the recording demands on the families more reasonable with three versus four books. In addition to these three books, families were asked to also record themselves reading books they already owned at home. Families provided a total of 183 individual book reading episodes: 60 of novel books and 123 episodes of books that the family already owned.

The present report focuses on the descriptions and analyses of the 60 recordings (about 10 hours) of the novel books provided to the families. Novel book recordings were not equally distributed across books or families (Table 2) as some families contributed more or longer recordings than others (see Table 5).

Procedure

One parent from each family came to the lab for a one-time pre-study visit during which they were provided with the study materials. Parents selected three books that they did not own from the four available books. Families were provided with a digital recorder which they were instructed to keep at home for two weeks and record a minimum of 6 home-book reading sessions that included the books provided by the lab along with additional sessions including books that they owned at home. Families were not given any instructions about how to read or interact with the books aside from the experimenters emphasizing that the families should read the books the same way as they typically do at home. Experimenters instructed families on how to use the audio recorder and told families to keep the recorder in a pocket, or someplace not visible during recording. Families were instructed to record the first reading of the new books. After completing the recordings, families were instructed to return the digital recorder with the recorded reading sessions at the end of the two-week period via the postal mail service in a pre-paid envelope. During the visit, each parent completed 2 questionnaires: a paper-and-pencil MBCDI Words and Sentences and a brief survey of home reading practices.

Compensation. Families were given travel expense reimbursements, $40 compensation for the time taken to record book-reading sessions at home and kept the 3 books they selected during the visit to the lab.

Audio transcription and coding

Coding. 12 trained undergraduate research assistants used the ELAN software (Brugman & Russel, 2004) to diarize (tag speakers), segment (identify timestamp boundaries of utterances), and transcribe approximately 10 hours (575 minutes = 9.58 hours) of picture book reading of the four new books provided by the researchers. Additionally, each transcript was checked for accuracy by a research assistant who did not transcribe that file, so each research assistant transcribed some files and acted as a checker on other files. Transcription and annotation were done in the ACLEW DAS format (Casillas et al., 2017; Soderstrom et al., 2021), which is compatible with the CHAT and CLAN systems, with a few exceptions. First, we did not code vocal maturity (vcm) of child utterances because all target children produced words. Second, we included an additional tier under the adult speaker tier in which we coded whether utterances consisted of book text read aloud or other speech. Each minute of audio took about an hour to transcribe, and an additional half hour to check, yielding what we believe is an accurate and thorough corpus of naturalistic home book reading. The raw audio and transcripts will be available to other researchers in the CHILDES repository (https://childes.talkbank.org/). All other data and code is available at https://osf.io/b3egw/.

Measurements. Book text that was read aloud, and all speech produced by any individual captured in the audio recording was transcribed, including sibling and off-topic speech when present (e.g., cases when another speaker entered the room and asked something not related to the book reading session). From these transcripts, turn-taking, word counts per minute, counts of unique words uttered, and mean length of utterance (in words) of caregiver and child utterances were computed. In addition to computing overall counts for caregiver speech, we computed these variables of interest separately for caregiver speech consisting of book text read aloud or extra-textual speech.

We defined extra-textual speech as any caregiver speech (not sibling speech) produced during book reading that was not book text read aloud. However, adult speech not directed to the target child was not included. For example, if a caregiver entered the room and asked a question, and the picture-book-reading caregiver responded, none of this speech was included in our counts, though it was transcribed. The speech we defined as extra-textual includes talk about the stories or pictures, as well as talk not directly related to the story (e.g., instructions to turn the page, requests to sit down) or occasional talk unrelated to book reading. The amount of talk not directly related to the book content varied by family.

We computed the mean length of an utterance (in words) as an approximate measure of the grammatical complexity of an utterance (Hunt, 1970; Parker & Brorson, 2005). A segment was considered an utterance if it satisfied at least two out of three of the following criteria: (a) there is a silence/speech pause equal or longer than 2 seconds before it, (b) it presented terminal intonation contour, and (c) it presented syntax that makes a complete sentence (Bernstein & Brundage, 2013). Those cases that presented ambiguities were discussed by the first and last authors until a consensus was reached. The mean length of utterances was calculated by dividing the total number of words produced by a speaker by the total number of utterances produced by a speaker using a python script (see Supplemental Materials). The mean length of utterances for the speech from CHILDES were automatically estimated by the CLAN system (MacWhinney, 2000), which uses an identical method.

Results

We first describe the audio recordings that make up our picture book reading corpus, including the individuals that appear in the recordings and the number and length of the recordings. We then describe the content of the audio recordings, including the proportions of total words and complex sentences contained in the book text that were read aloud by caregivers. Finally, we compare the language contained in the audio recordings of picture book reading to other conversational contexts, drawn from existing recordings in the CHILDES corpus (MacWhinney, 2000).

Description of the audio recordings

Individuals in the Recordings. In nine of the 12 families, a female caregiver (the mother) read the picture books in all audio recordings. In two families both a male and female caregiver (mother and father) each contributed audio recordings of reading sessions, and in one family a male caregiver (the father) read the books in all audio recordings. Out of the 12 families 10 only had one child participating in the recording sessions and two families included an additional child – an older brother (4 and 5 years of age). The four-year-old brother participated in 2 out of the 6 total recordings and the 5-year-old participated in 9 out of 11 total recordings made by families. Though sibling speech was transcribed, it is not included in the current analyses.

Book Reading Recordings. The corpus consists of 60 individual book reading sessions summarized in Table 4 (mean per family: 5; range: 2-11). Families spent on average about 10 minutes reading one book (range 1-24 minutes). These results are comparable to and build upon the earlier reports that families spend between 3 and 15 minutes per book (Anderson-Yockel & Haynes, 1994; Cronan et al., 1996; Haynes & Saunders, 1998; Lyytenin et al., 1998).

Table 4. ***Book Reading Session Descriptive Statistics***

Families	Books	Cumulative duration (hour)	Reading time per book (min) (SD)
Family 1	5	0.88	10.60(4.98)
Family 2	7	1.31	11.29(2.63)
Family 3	11	1.21	6.64(4.03)
Family 4	7	1.18	10.14(4.41)
Family 5	3	0.52	10.33(5.69)
Family 6	3	0.22	4.33(2.31)
Family 7	4	0.40	6.00(4.69)
Family 8	3	0.47	9.33(3.21)
Family 9	8	1.97	14.75(6.59)
Family 10	3	0.62	12.67(5.51)
Family 11	2	0.13	4.00(2.83)
Family 12	4	0.67	10.00(3.46)
Total:	60	9.58	9.17(3.30)

We observe considerable variability in book reading duration both within and between families. Overall, families spent more time reading the longer books which contained more text. Figure 1 illustrates reading times of each book by each family, and clearly shows that overall families spent less time reading the book with the least amount of text (*That is Not a Good Idea*).

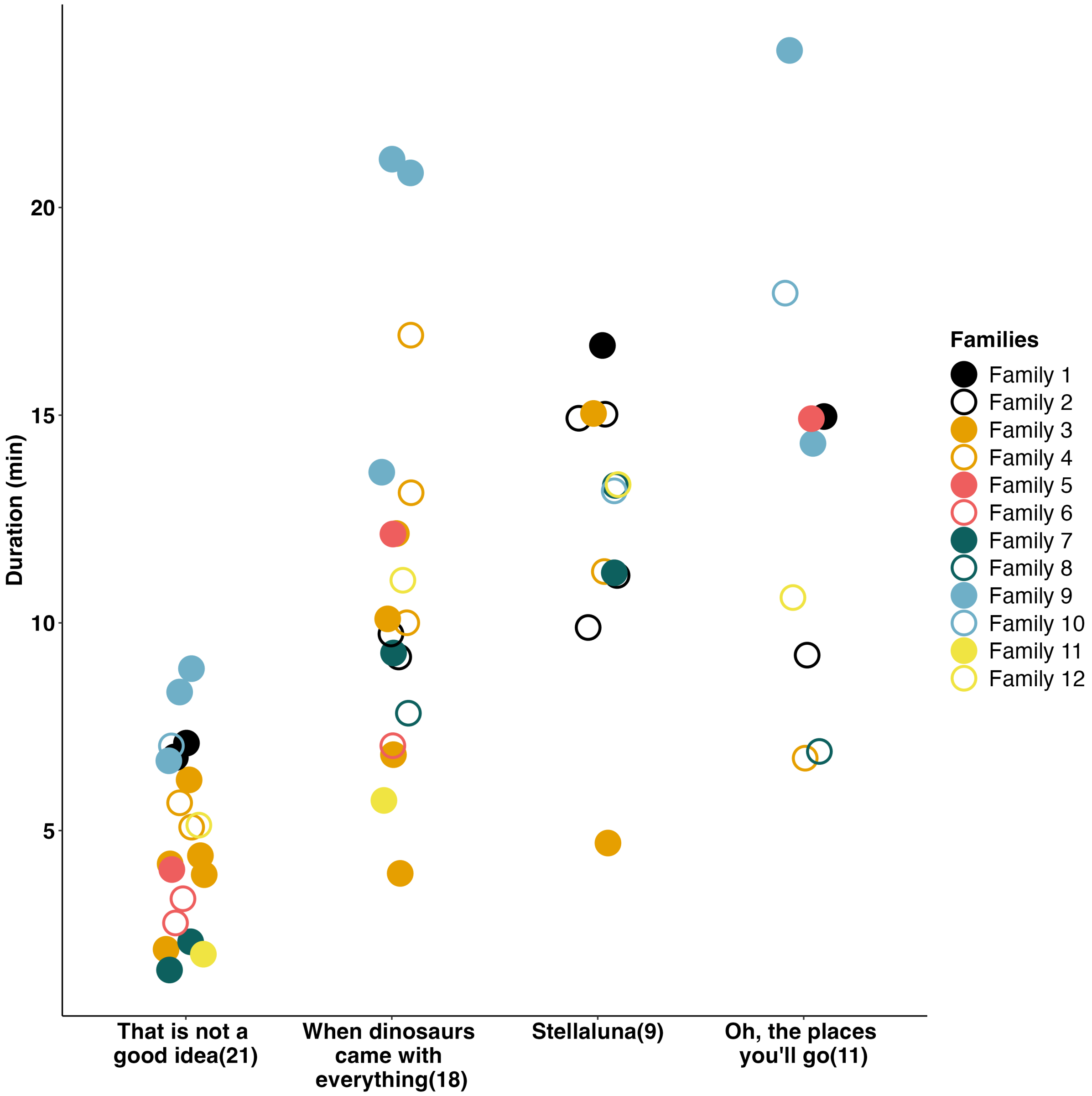


Figure 1. ***Reading duration by book and family; 1 point = one book, so each family may contribute multiple points to a single column; Recording counts by book are included in parenthesis***

Figure 1 also illustrates considerable variability across families, with some families consistently spending more or less time on a single book than others. For example, family 9 (blue filled circles) generally spent more time than average on each book and family 7 (green filled circles) generally spent less time. There was enormous variability across individual book reading episodes, and both features of the books and family individual differences contributed to overall reading times.

Analysis of the audio recordings

The first question we aimed to answer was whether families consistently read the book text aloud during shared book reading. If families do indeed read the book text aloud, differences in lexical diversity and syntactic complexity between the language in book text and typical child directed speech may be a plausible mechanism by which picture book reading may contribute to language outcomes.

Word Proportions. We first computed the overall proportion of the book text that was read aloud during the reading of each book. Caregivers occasionally re-read portions of text and these re-reads were counted only once. The proportions thus refer to the proportion of text read aloud, verbatim, at least once. Figure 2 indicates that caregivers overwhelmingly read all the text contained in the picture books. In only 7 reading instances across all 60 episodes did caregivers skip words. In four of these book reading episodes, parents summarized the text of the book and gave the child a warning before reading the book that they intend to summarize not to read word-for-word from the books. In the remaining three book reading episodes (all instances of "Stellaluna") parents summarized the plot from one to two pages for each of the reading episode without indicating to the child that they were summarizing. Only 3 families ever engaged in the summarizing behaviour while the remaining 9 families read all the words in every book they read (See Online Supplemental Materials Exhibit A for the visualization of word proportions by families).

Sentence Structure. In addition to the proportion of total words, we also measured how often the target complex syntactic constructions were read verbatim and unchanged by caregivers. Here we took a very conservative approach and noted any change that was made to the complex sentence, including the addition of extra words not in the text.

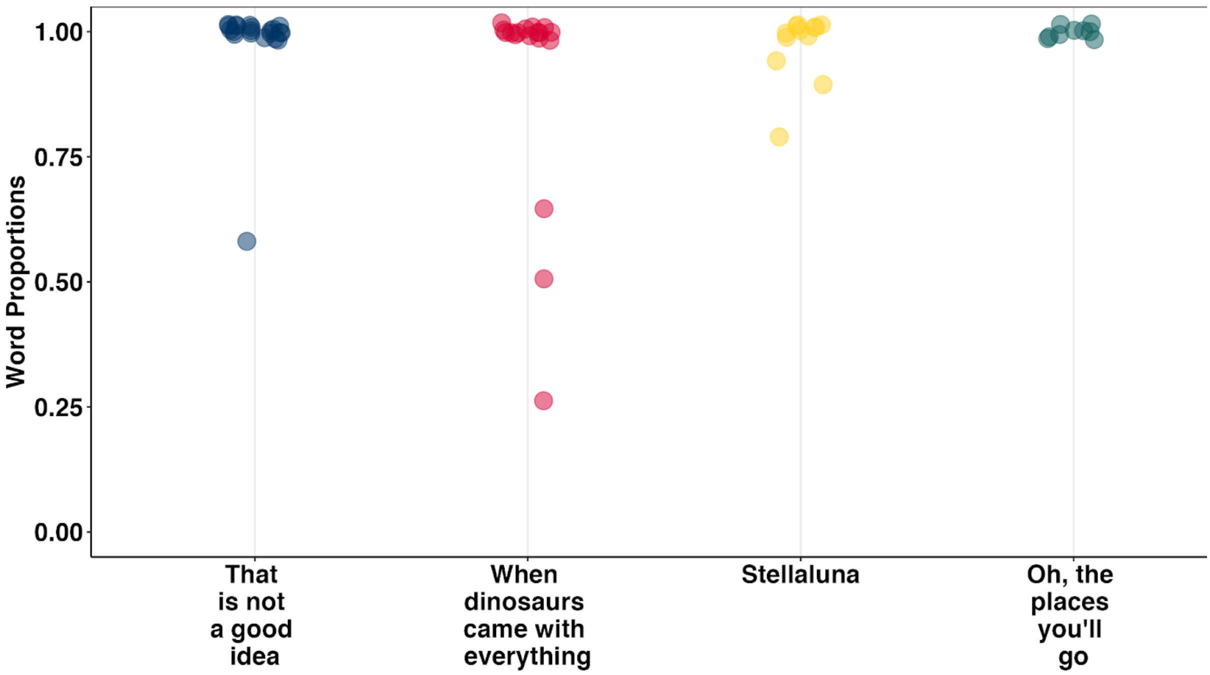


Figure 2. ***Proportion of words in the book read aloud verbatim; 1 point = 1 book, so each family may contribute multiple points to a single column***

The complex syntactic constructions in the books were indeed consistently produced by caregivers (Table 5). Out of 207 target constructions approximately 85% (175) were read from the book without any modification. The type of modification in the remaining 32 complex syntactic constructions are summarized in Table 5 (see Online Supplemental Materials Exhibit B or a complete list and the counts of syntactic constructions by modification types). Most modifications were additions before or after the target construction (25 out of 32 total), so the caregiver read the entire construction aloud but added a word or words of their own, sometimes a relative pronoun and sometimes a re-statement of, or commentary on, the complex construction. Only 5 instances were modifications of the syntactic constructions proper. Three of those five modifications were the addition of extra words in the construction and the remaining two were instances in which the caregiver did not read the construction. That means that the complex construction was produced intact 99% (205/207) of the time, and intact and unmodified in any way 98% (202/207) of the time. In our sample, the rare and complex sentences in picture books do indeed become a part of the linguistic input produced during shared book reading. Picture books may be an important source of complex syntax for children because adult caregivers seem to consistently read the complex language in the book text aloud.

Table 5. ***Summary of the modification types. Added words are indicated with an underline, omitted words are indicated with a strikethrough.***

Modification type	Example	N=32
Addition or omission before construction	And you may ~~not~~ find any you'll want to go down$_{\text{ORC}}$	4
Addition after construction	You'll be left in a Lurch$_{\text{Passive}}$ *Parent:* Oh.. his poor balloon got caught up in a tree	21
Repetition	*Parent:* You can steer yourself any direction you choose$_{\text{ORC}}$ You can steer yourself any direction you choose$_{\text{ORC}}$	2
Addition within construction	Stellaluna was terribly hungry – but not for the crawly things that Mama Bird brought$_{\text{ORC}}$	3
Omission	~~The places you'll go $_{\text{Oblique}}$;~~	2

Differences between book reading and other conversational contexts

To further understand the language generated during picture book reading, and how it may vary from other sources of child-directed speech, we compared aspects of the language produced during picture book reading to the language produced in other contexts. The present analyses aim to explore whether conversation generated during book reading is indeed characterized by conversation turns, large amounts of speech, and lexically diverse speech, relative to other contexts. We also investigate variability in turn taking and features of produced speech across different books.

We chose the transcribed Bates (1988) corpus available through CHILDES as the source of other conversation contexts to which we compared our picture book reading. First, we needed a comparable number of caregivers and typically developing children that matched the participants in our sample on age and gender. Second, we needed interaction clips comparable in length to our own book reading recordings that reflect different conversation contexts. The Bates corpus fit our criteria and allowed us to compare our audio recordings to those made in different contexts: snack time, free play, and another picture book reading event (*Miffy in the Snow*; 288 words, 0 rare/complex sentence types per our coding scheme). All these contexts were recorded in the laboratory. Children were approximately age and gender matched to those in our sample (all 28 months; 7 girls, 5 boys), and as in our sample, mostly from a middle-class background. Each child-parent dyad participated in all three of the events for 10 minutes each.

Turn-Taking. We define a turn as a back-and-forth speech-exchange between a child and an adult within 5 seconds, following the traditional methodological convention (Hart & Risley, 1989). We wrote a turn-taking counting algorithm in Python (included in the supplemental materials) that used the speaker tags and utterance timestamps to compute turn taking counts. Given the noisiness of naturalistic transcripts we were unsure whether our simple code would yield accurate turn counts so additionally, three raters manually counted turns for all 60 book reading episodes. Raters discussed their individual counts until a common agreement was reached. These counts overwhelmingly agreed with each other (0.94 interclass correlation coefficient), and we report the manual counts here. The Bates corpus transcript does not contain utterance timestamps, so we could not use own Python code to compute turn counts. However, the CLAN program available for CHILDES transcripts (CLAN, MacWhinney, 2000) can compute turn counts. To ensure that this CLAN algorithm uses similar criteria as own method, two independent raters sampled 10 random clips from Bates and counted turns manually. These counts were similar (0.91 interclass correlation coefficient) to the counts provided by the CLAN algorithm, so we report the algorithm counts here.

Turns per Minute. To better compare across speaking and reading episodes that varied in duration, raw count of turns for each episode were divided by the time of the episode to compute the number of turns per minute (Figure 3). We see considerable variability within each context, but two trends emerge. First, we do not find that book reading contexts contain more conversational turns than other contexts. The snack and free play contexts both elicited high counts of conversation turn and both the short book from CHILDES (*Miffy in the Snow*) and our corpus (*That is Not a Good Idea*) elicited more turns than the longer books. We do not believe that the overall higher rate of conversation turns in the Bates corpus can be attributed to methodological differences in how turn counts were computed, because we manually computed turn counts from the written transcript of each corpus using similar criteria. However, there is a potential confound such that the Bates corpus was recorded in a laboratory setting so both children and caregivers might have behaved differently than they would at home. That said, even within the Bates corpus, the book reading activity did not lead to more conversational turns than other contexts. At the very least, we can conclude that if there is a true effect that book reading promotes more turn taking than other contexts, this effect is small enough such that is it swamped by differences in recording context (home or lab). Our data is also consistent with the interpretation that book reading contexts do *not* systematically lead to more turn-taking (in dyads with children of about 2.5 years of age) than other conversational contexts.

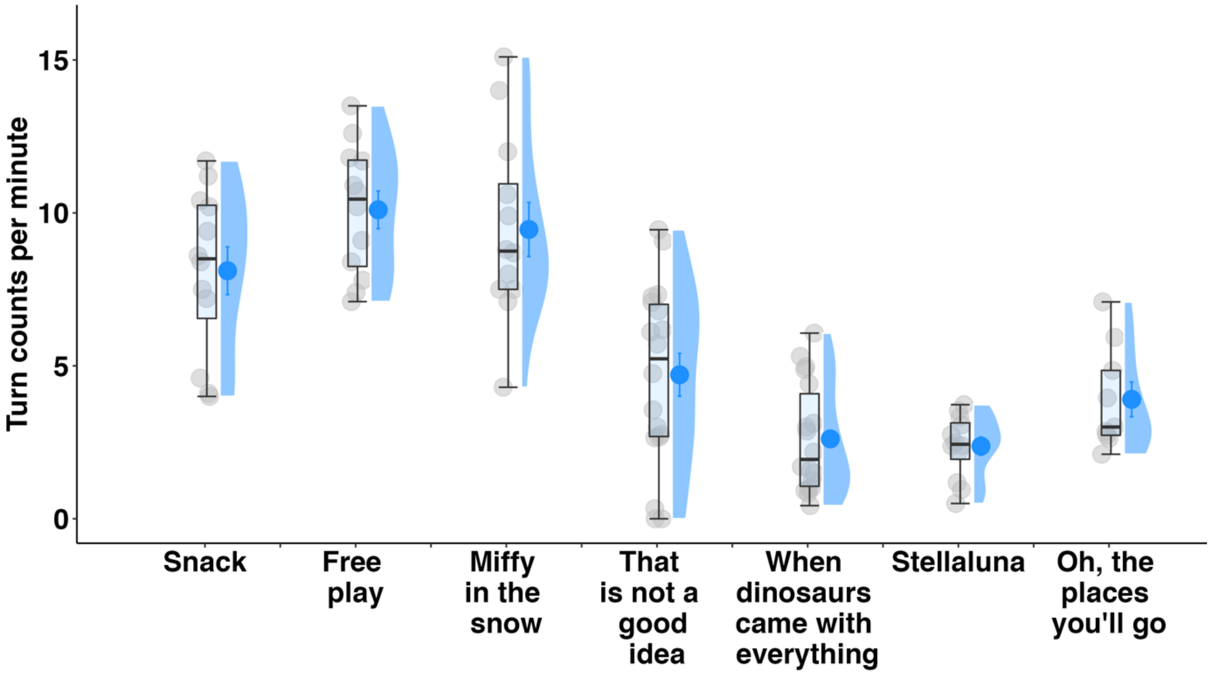


Figure 3. ***Turn taking per minute by book and family; 1 point = one reading session***

Second, different books generated different numbers of conversational turns. The short and syntactically simple books *Miffy in the Snow* (mean: 10; range 4-15) and *That is Not a Good Idea* (mean: 5; range 0-9.2) elicited the highest count of turns per minute. This is not surprising given that the text of these books is very simple with many pages consisting of only a few words, and much of the story is conveyed through the pictures, which depict events not otherwise described in the text. The three other books yielded similar rates of turn taking: *Oh, the Places You'll Go* (mean 3; range: 2-7)*, When Dinosaurs Came with Everything* (mean: 2.5; range 1-5) and *Stellaluna* (mean: 2.5; range 1.5-3). These books all contain more words of text and more detailed stories or narratives. In fact, the books that numerically generated the least conversational turns (*When Dinosaurs Came with Everything* and *Stellaluna*) tell complete stories from beginning to end that are surprising and complex: A boy accompanies his mother on a number of errands and receives a free dinosaur from each establishment they visit (*When Dinosaurs Came with Everything*), and a story about a bat who is temporarily separated from her mother and lives with a family of birds before finding her mother again (*Stellaluna*). One possible explanation of the observed data is that different books afford different reading styles. Some books, given their text or pictures, afford lots of back-and-forth conversation between children and their caregivers. Other books, especially those with complex stories conveyed through the text alone, may promote more silent listening of the story.

Age effect. To better understand how a child's own age might affect turn counts, we examined the turns counts per minute across four books of interest as a function of child age (Figure 4). The number of turns per minute for all the books *decreased* with age (r=-48, p<.001). The correlation with MBCDI score also yielded a negative correlation (r=-.38, p<.01). However, when the child who had the MBCDI score in the lowest 5% bracket was removed from the analyses the correlation between turns per minute and MBCDI score was not significant (r=-.03, p>1). Nonetheless, we do see some evidence that younger children, possibly due to their weaker language skills, or perhaps for another reason, engage in more conversational turn taking than older children. We anecdotally notice when listening to the audio that older children tended to genuinely enjoy passive listening to the story, particularly the longer and more complex stories. Perhaps the children whose age or language skills allow them to understand and appreciate the story prefer to silently listen, while children who cannot yet understand or appreciate a longer narrative (and their caregivers) use book reading as a more interactive, conversation-generating activity.

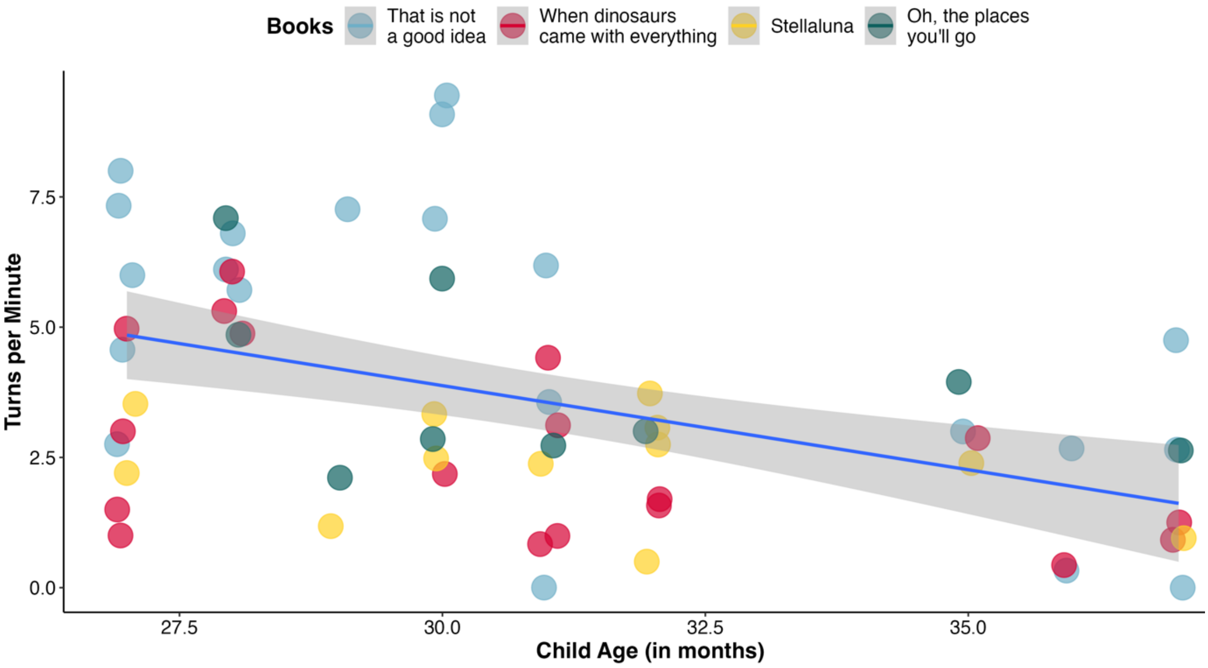


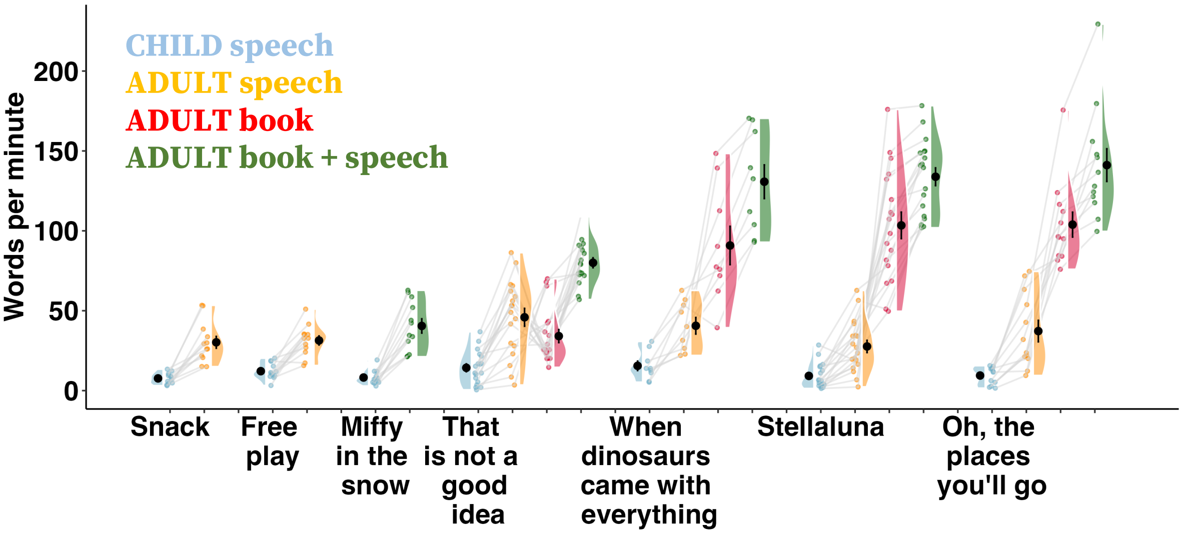
Figure 4. ***Turn taking per minute by child's age; 1 point = one book***

The number of turns highlight only one dimension of caregiver-child speech. The speech measures we report next: the total words per minute, the number of unique words, and mean length of utterance for children and their caregivers, across the seven contexts, provide a clearer picture of differences across conversational contexts. These counts also appear less sensitive to the potential confound (laboratory setting versus at-home recording) that may be present in the turn-taking counts.

Characteristics of caregiver and child speech

To understand how different conversational contexts affect speech characteristics of children and caregivers, we investigated various features of the speech produced in our picture book reading contexts and the Bates recordings. We investigated (1) the total number of words, (2) the number of unique words and (3) the mean length of utterances produced in different contexts. Importantly, for the picture book reading contexts, we look at the book text and extra-textual talk separately to understand the contribution of both to the overall language produced.

Figure 5 illustrates all three measures per individual reading session. In these figures, the blue points refer to child counts, the yellow refers to adult counts in extra-textual talk only, the red points refer to adult counts in read-aloud book text only, and the green points refer to counts in all adult speech (merging extra-textual talk and read-aloud book text). In non-book reading contexts, there are only blue and yellow points because there is no book text to read aloud. The *Miffy in the Snow* context is not broken down into book text and extra-textual talk because this distinction was not annotated in the written transcripts as we did in our own transcripts. Table 6 contains means and standard errors.



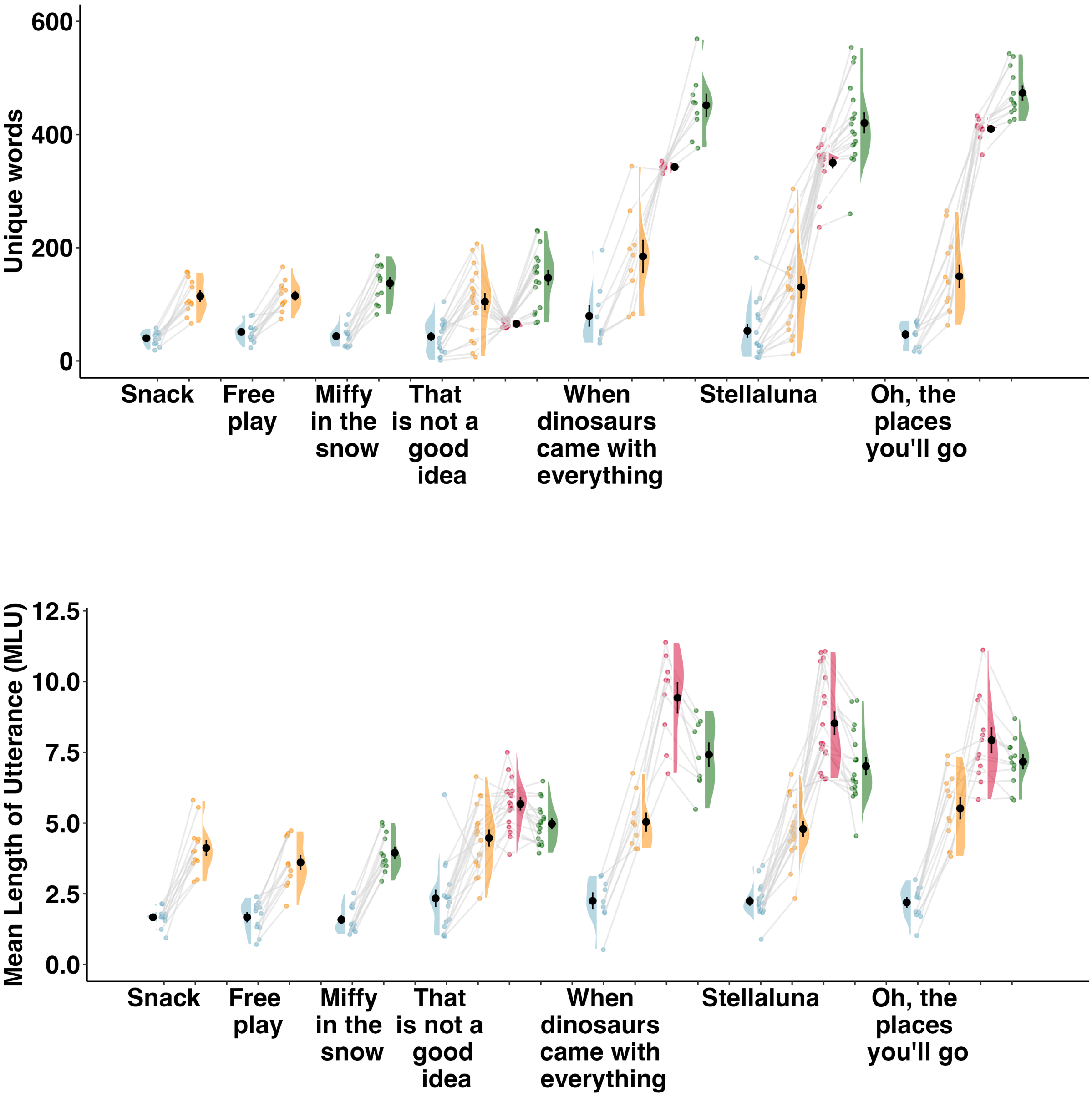


Figure 5. ***Words per minute, unique words and mean length of utterances (MLU) by different context; points connected with lines = one reading or speaking episode***

Table 6. Speech characteristic means with standard error (se) across different contexts

Measure	Speaker	Speech Type	Context: Mean (se)						
			CHILDES			Novel Corpus			
			Snack	Play	Miffy	Idea	Dino	Stellaluna	Places
Words per minute	Child	Speech	8 (.98)	12 (1.46)	8 (1.28)	18 (3.89)	9 (1.83)	9 (1.57)	16 (2.94)
	Adult	Extra-textual talk	30 (3.69)	31 (2.83)	na	46 (5.57)	28 (3.83)	37 (6.72)	41 (5.19)
		Book text read aloud	na	na	na	34 (4.08)	103 (8.29)	104 (7.79)	91 (11.99)
		Extra-textual talk + book text	na	na	41 (4.47)	80 (3.11)	128 (8.02)	141 (10.24)	131 (10.55)
Total words	Child	Speech	77 (9.84)	122 (14.59)	82 (12.84)	97 (17.92)	128 (35.55)	114 (20.25)	230 (73.15)
	Adult	Extra-textual talk	302 (36.91)	314 (28.33)	na	280 (47.82)	348 (70.86)	463 (95.00)	566 (116.71)
		Book text read aloud	na	na	na	145 (4.15)	1019 (37.40)	1214 (32.11)	1043 (23.42)
		Extra-textual talk + book text	na	na	405 (44.74)	425 (49.33)	1360 (117.22)	1677 (104.16)	1600 (109.87)
Unique words	Child	Speech	40 (3.47)	51 (5.08)	44 (5.00)	43 (6.82)	53 (10.94)	47 (6.13)	80 (17.68)
	Adult	Extra-textual talk	115 (8.92)	115 (7.38)	na	105 (13.97)	130 (18.49)	150 (19.05)	185 (28.03)
		Book text read aloud	na	na	na	65 (1.26)	350 (9.14)	410 (4.96)	343 (2.10)
		Extra-textual talk + book text	na	na	137 (9.70)	147 (12.09)	402 (23.50)	474 (12.21)	452 (19.04)
MLU	Child	Speech	1.67 (.10)	1.67 (.15)	1.58 (.14)	2.34 (.28)	2.24 (.14)	2.19 (.16)	2.25 (.28)
	Adult	Extra-textual talk	4.12 (.25)	3.61 (.25)	na	4.47 (.27)	4.79 (.25)	5.52 (.36)	5.04 (.32)
		Book text read aloud	na	na	na	5.68 (.21)	8.53 (.39)	7.92 (.43)	9.43 (.53)
		Extra-textual talk + book text	na	na	3.95 (.19)	4.81 (.23)	7.01 (.29)	7.17 (.24)	7.42 (.40)

Words per minute. Caregivers produced more words per minute during book reading than other activities. In three out of the four books we provided families the extra-textual talk consisted of more words per minute than other activities (28, 37, 41 and 46 words per minute vs. 30 and 31 for snack and play) but these rates of speech grow substantially with the presence of book text spoken aloud (80-141 words per minute). Caregivers in our sample overwhelmingly read all the book text, which contributed a substantial number of words to the interaction. The smallest contribution of text read aloud was found for the shortest book in our sample (*That is Not a Good Idea)* which incidentally also generated the most conversational turns. These findings are consistent with variability across books. Books that afford greater back-and-forth conversation generate more words of extra-textual talk, but longer books that afford more passive listening generate more caregiver words overall.

We observe small differences in the amount of child speech across books and non-book contexts. Some books, including *That is Not a Good Idea* generated more child words per minute than snack time or free play, while other books generated word counts approximately equal to those produced during non-reading activities.

Unique words. The fact that caregivers and children often produce more words per minute during shared book reading does not mean that these words are qualitatively different from the words produced in other contexts. To understand possible differences in the speech that is produced, we examine the number of unique words and the mean length of utterances elicited during book reading and other contexts. That is, we counted the number of unique words produced in each context. However, a methodological challenge arises such that while number of unique word types increases as the total number of word tokens increases, the rate of increase necessarily slows as the sample size gets larger given constraints of natural language (Heaps, 1978; Herdan, 1960; see also Malvern et al., 2004; McKee et al., 2000; Montag et al., 2018; Richards, 1987). Put simply, ratios of unique words to other words, such as type-token ratios are so confounded by sample size that they make poor estimates of unique word counts or lexical diversity measures. For this reason, we report only counts of unique words to avoid a measure of lexical diversity that is deceptively contaminated by total word counts. While unique word counts are not a measure of lexical diversity, they do provide some information about the range of vocabulary items that are present in a conversational context and how contexts might vary, even if they also vary in the total number of words produced. Figure 5 illustrates that caregivers produced more unique words during book reading than during other activities. Once again, this pattern was particularly true of the three longer, more complex books, where, again, the effect was largely driven by the presence of book text spoken aloud.

To illustrate the lexical diversity of speech samples in a way that is independent of total word counts we plot the number of unique words in caregiver speech that included both book text and extra-textual talk in similarly sized samples in Figure 6. As a workaround for the confound of the Bates corpus (all speech collected in a lab setting), we included additional 40 age-matched conversations from CHILDES corpus: Peter (Bloom, et al., 1974; Bloom et al., 1975), Adam and Sarah (Brown, 1973), and

Nina (Suppes, 1974) that were recorded at home (grey dots labelled CHILDES). These additional recordings may encompass a range of contexts and we neither selected nor excluded recordings on the basis of the conversational context. Figure 6 shows the number of unique words in samples that increase in increments of 10 words (i.e., first 10 words, then first 20 words, and so on), averaged across each transcript of the same context. The error bars refer to one standard deviation in the average number of unique words. The error bars become smaller and disappear and values appear noisier as the samples get longer because transcripts varied in length, so fewer transcripts are included in the means as the total word counts increases. The vertical spread of y-values at a single x-value can be interpreted as a difference in lexical diversity. For example, at 500 total words, snack and free play contexts (black and orange dots respectively) contain approximately115 unique words, the additional at-home CHILDES samples contain about 150 unique words while *Stellaluna* (yellow dots) contains nearly 100 more unique words (218 unique words).

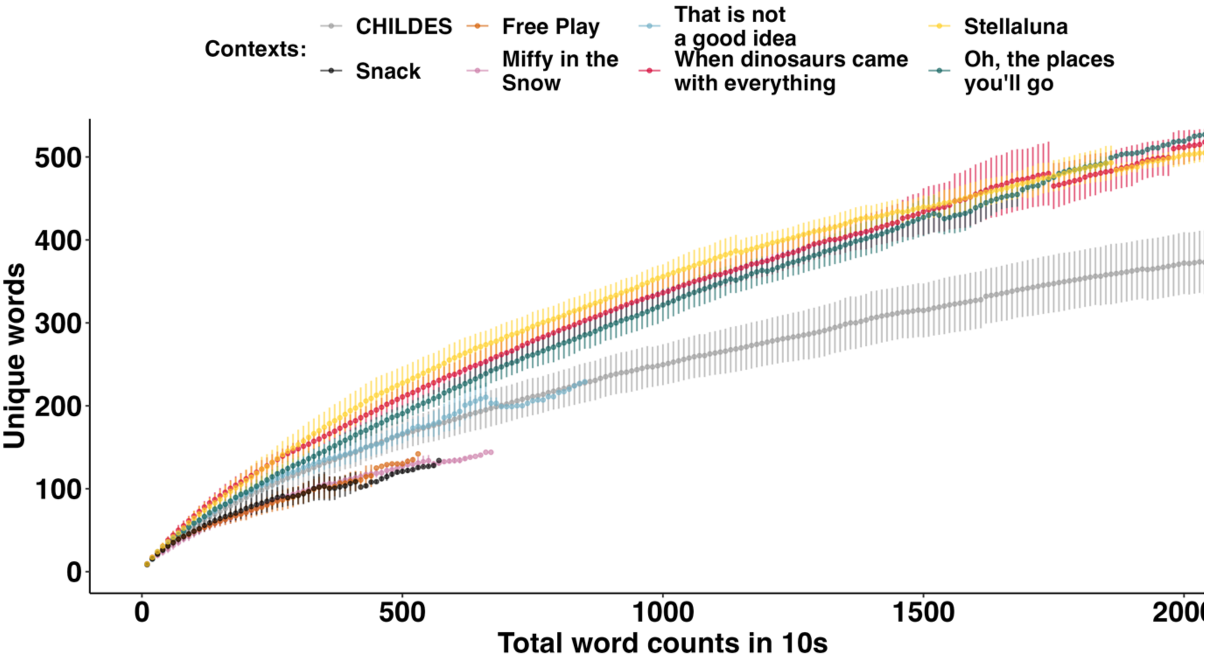


Figure 6. ***Unique word means over the total words in 10-word increments by different contexts. Error bars = 1 standard deviation***

We observe far more unique words in similarly sized speech samples during at-home shared reading events than in other contexts, including other at-home contexts. It is unclear why we count fewer unique words in the Bates corpus recordings than in other additional at-home CHILDES recordings. This difference may be due to the differences in population, recording, or experimental methodology. We also observe differences across the four books in our sample. The book reading sessions with longer books elicited more unique words, likely reflecting the large amount of lexically diverse book text that was read aloud. Further, the book that generated the most extra-textual words per minute (*That is not a good idea*) overall showed the least lexically diverse speech. Again, we see a trade-off, such that books that generated more speech, including more child speech per minute generated less lexically diverse caregiver speech.

Mean Length of Utterance. Caregivers produced longer utterances during book reading than other activities. Again, this pattern was particularly true of the longer, more complex books, and this pattern was largely driven by the presence of book text that is spoken aloud. The most syntactically complex books generated the longest mean length of utterances, which is consistent with the finding that caregivers indeed read book text aloud. Children also consistently produced longer utterances (on average about an extra half a word per utterance) during book reading at home than in other contexts, though we see only small differences across different books.

Discussion

We built a corpus of caregiver-child interactions during shared book reading recorded in homes to better understand the language that is produced during shared book reading. We found that caregivers overwhelmingly read the book text, including rare and complex sentence structures. We also found that books varied in the profiles of language they generated, with some books promoting more conversational turns and extra-textual language, while others promoted more overall words, unique words, and longer utterances. Further, relative to other conversational contexts, book reading generally generated overall more words, more lexically diverse talk, and longer utterances, but these tendencies were driven by the presence of book text read aloud, so they depended on characteristics of the book being read. Rather than a single profile of speech generated during book reading, different books may promote different profiles of caregiver-child interaction.

The goal of better understanding the language generated during shared book reading was to aid in establishing the plausibility (or implausibility) of causal pathways by which shared book reading might positively contribute to language outcomes. Hypotheses surrounding the reasons book reading may be associated with positive language outcomes often focus on features of the language or conversation generated during book reading, so evaluating these hypotheses requires a better understanding of the language generated during shared book reading. Our first key finding was that caregivers indeed consistently read the text of the picture books, so findings that picture book text is more lexically diverse and syntactically complex are indeed germane to the language environment generated during shared book reading. We replicate existing findings that the language produced during picture book reading is more lexically diverse than language produced in other contexts (e.g., Crain-Thoreson, et al., 2001; Demir-Lira et al., 2019; Hoff-Ginsberg, 1991; Mol & Newman, 2014; Salo, et al., 2016; Sosa, 2016; Weizman & Snow, 2001). Our work further clarifies these findings by showing explicitly that the increases in MLUs, speech rate (word count per minute), and unique words during shared book reading relative to other contexts are driven by the book text. The difference between speech in non-book contexts and speech in book reading contexts is largely driven by the presence of book text read aloud. Additionally, while child speech during book reading sessions did not elicit more conversational turns than during other contexts, children produced more words, more unique words and higher MLUs during longer book reading sessions than in any of the other contexts. These results suggest that child speech produced during book reading sessions perhaps is not part of back-and-forth conversation per se but rather in response to the book text read aloud that children heard.

With respect to the extra-textual talk, including turn-taking generated during shared book reading, our analyses suggest that the nature of this extra-textual talk depends a great deal on features of the book being read. Our book with the least amount of text, with pictures that tell aspects of the story that are not present in the text, generated the densest turn taking, while the books with more text and complex narratives generated the least turn-taking. This result replicates earlier findings showing that stories with less text facilitate more extra-textual talk per minute than stories with more book text (Chaparro-Moreno et al., 2017; Greenhoot, et al., 2014; Muhinyi & Hesketh, 2017; Petrie et al., 2021). While we do not want to overgeneralize our findings on the basis of only four different books, it is certainly the case that there is enormous variability in the text and pictures of picture books. Some of this variability likely reflects creative choices on the part of the authors and illustrators to vary how caregivers and children interact with the book, so it is not surprising that we observe variability across books in the profiles of speech and conversation that they generate.

More surprising, we find that shared book reading did not necessarily generate more conversational turns than other conversational contexts, like snack time or playtime. The Bates corpus, to which we compare our corpus, was collected in the lab while our shared book reading recordings were recorded at home, and this difference may have affected caregiver behaviour. One speculative possibility is that adult caregivers are not comfortable with silence in such formal unfamiliar environments as laboratory settings. Consequently, they produce large amounts of speech that is quite simple: more back-and-forth, but shorter, simpler utterances and more repetition. However, we do find that our shared book reading interactions generated fewer conversational turns than the activities in the Bates corpus, and even within the Bates corpus the picture book reading did not generate more conversational turns than eating a snack or playtime. That said, our book that generated the most conversational turns indeed generated more child words per minute than any other reading or non-reading context. More work documenting turn taking across different, ideally naturalistic contexts, is needed to draw stronger claims, but we find mixed evidence for the idea that shared book reading promotes turn-taking and child speech because there was substantial variation across books, families, and non-book contexts.

Given the emphasis on extra-textual talk in the picture book reading literature, as in dialogic reading and other approaches, it may be unexpected that we observe such low rates of extra-textual talk. There are many reasons for this potential discrepancy. One potential explanation is that there is in fact no discrepancy at all—our simplest book (*This is not a good idea*) generated similar rates of extra textual talk as did other episodes reported in the literature. In our simplest book, 58% of caregiver words were extra-textual. In a sample of 2–27-month-old children and their caregivers reading similarly simple books, Cline and Edwards (2017) report that 67% of word were extra-textual and in a sample of 18–30-month-old children and caregivers spontaneously reading books at home, Demir-Lira et al. (2019) find that 76% of utterances were extra-textual. If utterances in which caregivers read the book text are longer than extra-textual utterances (as they were in our analysis), these figures are broadly consistent with what we find. Our findings are not in contrast with existing results, but rather compliment and extend the literature to emphasize book effects, that the type of book that families are reading has enormous implications for the speech that caregivers and children produce.

Methodological differences may also underlie other observed discrepancies between our findings and other findings in the literature. The children in our sample were between about a year younger (Gilkerson et al. 2017; Lonigan & Whitehurst, 1998; Muhinyi et al., 2019; Muhinyi & Hesketh, 2017) or up to 3 years younger (Mol & Newman, 2014; Grolig et al., 2020; Payne et al., 1994) than many other studies that record caregiver-child interactions during book reading. Given clear age-related differences in caregiver speech during book reading (e.g., Patel et al., 2024; Senechal et al, 1995), the age of the children in our sample may contribute to the lower rates of caregiver utterances for some of our books. Another important methodological difference is that families used the audio recorder in their homes, with no experimenter present. Families were in a familiar location, were not being observed or videotaped, and were asked to keep the audio recorder out of sight, so it may have been easier to “forget” that they were being recorded and act more naturally, or at least differently, had the recording been more obvious. Finally, our sample is small and somewhat homogenous, so it is certainly possible that our sample demographics contributed to our observed results.

We interpret our results as suggesting that there may indeed be a plausible, causal relationship between picture book reading and language outcomes because we observed differences between various aspects of conversation and speech between picture book reading and other activities. Hypothesis about the utility of picture book reading generally focus on aspects of the language generated during book reading and how it may be different from other conversational contexts. While our analyses do not themselves advance claims of causality, our analyses suggest that these hypotheses surrounding the language generated during book reading are plausible, sensible candidate for mechanistic pathways by which picture books come to be associated with language outcomes. However, the pathway may be that reading provides a varied range of diverse experiences rather than any one feature.

The two non-mutually exclusive explanations for the positive effects of picture book reading, caregiver-child conversation and vocabulary and sentence structure of the book text, may both be correct, but in different contexts for different books. Some books may promote turn-taking and child speech, others varied vocabulary or rare syntax, still others facts about the world, and so on. Variability of experience with very different profiles of reading across books may be an important contribution of book reading to children's language environments. Variability of experience would also help explain why interventions that aim to alter caregiver reading behaviour may not be associated with better language gains than an active control group (e.g., Noble et al., 2020). Perhaps, it is not a specific style of reading or type of input that contributes to language outcomes but rather varied experience with a range of reading styles and language profiles.

We speculate that further support for the idea that a varied range of experiences underlies observed language benefits of shared book reading is found in our negative correlation between age and turn taking. If turn-taking or child productive language were a central goal of book reading, turn-taking should increase as a child's own language skills support such conversation. Perhaps children (and their caregivers) who had the language skills to understand the story preferred to listen to the story and engage in less conversation. If this is a common behaviour across families, it may be normatively true that shared book reading is not always accompanied by a great deal of extra-textual conversation, which is relevant when evaluating correlational studies that associate shared book reading with positive language outcomes. Our result is consistent with other reports of a negative effect of age on caregiver speech (Muhinyi et al., 2020), and may suggest a more complicated relationship between child age, extra-textual speech and other family factors that may contribute to caregiver extra-textual talk.

More generally, this work points to the importance of the developing child in creating their own language environment. Our age-dependent effects on conversation are exploratory, but we think the ways that picture book reading changes as a function of child age or language skills is a potentially interesting finding worth of future work and relates to existing work describing developmental cascades. The negative effect of age on caregiver-child turn-taking suggests that child characteristics may shape the nature of the book reading episode. This finding is consistent with work that finds that either implicitly or explicitly caregivers can accommodate their child's linguistic knowledge when producing utterances (Huttenlocher et al., 2010; Leung et al, 2021) or more broadly, that a child's own linguistic, motor, or other abilities can have effects on the aspects of their environment (e.g., Bradshaw et al., 2022; Karasik,Tamis-LeMonda & Adolph, 2014; Kretch et al, 2014; Oakes, 2023; Thelen & Smith, 1994). The role of the child in shaping their own language environment provides an additional complication for systematic investigations of shared book reading as an intervention, because the same book or book reading intervention may produce different language experiences for different children.

We hope that future work can build about the present work. We collected recordings from only a narrow age range of children, so future work is necessary to better understand how child age and other characteristics might interact with book characteristics over a larger age range. Further, studies with larger and more diverse samples that include child language and literacy development measures are necessary to more directly link book reading with language outcomes and generalize these findings across larger populations of children. Our immediate goal of transcribing and annotating the recordings to create a sharable corpus necessitated this small sample, but expanding the sample would be an obvious next step. Future work could also collect book reading and other non-book reading speech samples from the same caregiver-child dyads to address the limitations associated with comparing book reading and other caregiver-child episodes across different children and families.

A remaining empirical question is to what degree the variability across different book profiles we observed may or may not extend to books outside our sample, including books that are familiar to families. In the present work, we limited our analyses to four books that were novel to families. These books may not be a representative sample of picture books for any number of reasons. For example, reading styles may vary considerably when reading books that are familiar to children and caregivers. Caregivers might summarize the text more frequently because they are familiar with the plots, or they may summarize less frequently because they are more familiar with the text. Likewise, familiarity with the plot, text or pictures may affect (in any direction) the amount of extra-textual conversation with which a family engages (Fletcher & Finch, 2015; Read et al., 2023). We hope to answer these questions in ongoing work with the book reading events in our corpus in which caregivers and children read books they already owned at home to gain a more complete picture of naturalistic shared book reading.

References

Anderson, J., Anderson, A., Lynch, J., & Shapiro, J. (2004). Examining the effects of gender and genre on interactions in shared book reading. *Literacy Research and Instruction, 43*(4), 1-20.

Anderson-Yockel, J., & Haynes, W. O. (1994). Joint book-reading strategies in working-class African American and White mother-toddler dyads. *Journal of Speech, Language, and Hearing Research, 37*(3), 583-593.

Arnold, D. H., Lonigan, C. J., Whitehurst, G. J., & Epstein, J. N. (1994). Accelerating language development through picture book reading: replication and extension to a videotape training format. *Journal of Educational Psychology, 86*(2), 235.

Arterberry, M. E., Bornstein, M. H., Midgett, C., Putnick, D. L., & Bornstein, M. H. (2007). Early attention and literacy experiences predict adaptive communication. *First Language, 27*(2), 175-189.

Baker, N. D., & Nelson, K. E. (1984). Recasting and related conversational techniques for triggering syntactic advances by young children. *First Language, 5*(13), 3-21.

Bates, E., Bretherton, I., & Snyder, L. (1988). *From first words to grammar: Individual differences and dissociable mechanisms.* Cambridge, MA: Cambridge University Press.

Bernstein, N., & Brundage, S. B. (2013). A Clinician's Complete Guide to CLAN and PRAAT. *Recuperado de http://childes. psy. cmu. edu/clan/Clinician-CLAN. pdf.*

Bloom, L., Hood, L., & Lightbown, P. (1974). Imitation in language development: If, when and why. *Cognitive Psychology, 6,* 380–420.

Bloom, L., Lightbown, P., & Hood, L. (1975). Structure and variation in child language. *Monographs of the Society for Research in Child Development, 40,* (Serial No. 160).

Bradley, R. H., Corwyn, R. F., McAdoo, H. P., & Garcia Coll, C. (2001). The home environments of children in the United States part I: Variations by age, ethnicity, and poverty status. *Child Development, 72*(6), 1844-1867.

Bradshaw, J., Schwichtenberg, A. J., & Iverson, J. M. (2022). Capturing the complexity of autism: Applying a developmental cascades framework. *Child Development Perspectives, 16*(1), 18-26.

Brown, R. (1973). *A first language: The early stages*. Cambridge, MA: Harvard University Press.

Brugman, H., Russel, A. (2004). Annotating Multimedia/ Multi-modal resources with ELAN. In: Proceedings of LREC 2004, Fourth International Conference on Language Resources and Evaluation.

Bus, A. G., Van Ijzendoorn, M. H., & Pellegrini, A. D. (1995). Joint book reading makes for success in learning to read: A meta-analysis on intergenerational transmission of literacy. *Review of Educational Research, 65*(1), 1-21.

Cameron-Faulkner, T., & Noble, C. (2013). A comparison of book text and child directed speech. *First Language, 33*(3), 268-279.

Casillas, M., Bergelson, E., Warlaumont, A. S., Cristia, A., Soderstrom, M., VanDam, M., & Sloetjes, H. (2017). A new workflow for semi-automatized annotations: Tests with long-form naturalistic recordings of childrens language environments. In *Interspeech 2017* (pp. 2098-2102).

Chacko, A., Fabiano, G. A., Doctoroff, G. L., & Fortson, B. (2018). Engaging fathers in effective parenting for preschool children using shared book reading: A randomized controlled trial. *Journal of Clinical Child & Adolescent Psychology, 47*(1), 79-93.

Chaparro-Moreno, L. J., Reali, F., & Maldonado-Carreño, C. (2017). Wordless picture books boost preschoolers' language production during shared reading. *Early Childhood Research Quarterly, 40*, 52-62.

Cleave, P. L., Becker, S. D., Curran, M. K., Van Horne, A. J. O., & Fey, M. E. (2015). The efficacy of recasts in language intervention: A systematic review and meta-analysis. *American Journal of Speech-Language Pathology, 24*(2), 237-255.

Cline, K. D., & Edwards, C. P. (2017). Parent–child book-reading styles, emotional quality, and changes in early head start children's cognitive scores. *Early Education and Development, 28*(1), 41-58.

Crain-Thoreson, C., Dahlin, M. P., & Powell, T. A. (2001). Parent-child interaction in three conversational contexts: Variations in style and strategy. *New directions for child and adolescent development, 2001*(92), 23-38.

Cronan, T. A., Cruz, S. G., Arriaga, R. I., & Sarkin, A. J. (1996). The effects of a community-based literacy program on young children's language and conceptual development. American Journal of Community Psychology, 24, 251–272.

Davies, C., McGillion, M., Rowland, C., & Matthews, D. (2020). Can inferencing be trained in preschoolers using shared book-reading? A randomised controlled trial of parents' inference-eliciting questions on oral inferencing ability. *Journal of Child Language, 47*(3), 655-679.

Deckner, D. F., Adamson, L. B., & Bakeman, R. (2006). Child and maternal contributions to shared reading: Effects on language and literacy development. *Journal of Applied Developmental Psychology, 27*(1), 31-41.

Dickinson, D. K., & Tabors, P. O. (1991). Early literacy: Linkages between home, school and literacy achievement at age five. *Journal of Research in Childhood Education, 6*(1), 30-46.

Donnelly, S., & Kidd, E. (2021). The longitudinal relationship between conversational turn-taking and vocabulary growth in early language development. *Child Development, 92*(2), 609-625.

Ece Demir-Lira, Ö., Applebaum, L. R., Goldin-Meadow, S., & Levine, S. C. (2019). Parents' early book reading to children: Relation to children's later language and literacy outcomes controlling for other parent language input. *Developmental Science, 22*(3), e12764.

Farrant, M., & Zubrick, R. (2011). Early Vocabulary Development: The importance of Joint Attention and Parent-child Shared Book Reading.

Farrar, M. J. (1990). Discourse and the acquisition of grammatical morphemes. *Journal of Child Language, 17*(3), 607-624.

Flack, Z. M., Field, A. P., & Horst, J. S. (2018). The effects of shared storybook reading on word learning: A meta-analysis. *Developmental Psychology, 54*(7), 1334.

Fletcher, K. L., Cross, J. R., Tanney, A. L., Schneider, M., & Finch, W. H. (2008). Predicting language development in children at risk: The effects of quality and frequency of caregiver reading. *Early Education and Development, 19*(1), 89-111.

Fletcher, K. L., & Finch, W. H. (2015). The role of book familiarity and book type on mothers' reading strategies and toddlers' responsiveness. *Journal of Early Childhood Literacy, 15*(1), 73-96.

Fletcher, K. L., & Reese, E. (2005). Picture book reading with young children: A conceptual framework. *Developmental review, 25*(1), 64-103.

Gilkerson, J., Richards, J. A., & Topping, K. J. (2017). The impact of book reading in the early years on parent–child language interaction. *Journal of Early Childhood Literacy, 17*(1), 92-110.

Gilkerson, J., Richards, J. A., Warren, S. F., Oller, D. K., Russo, R., & Vohr, B. (2018). Language experience in the second year of life and language outcomes in late childhood. *Pediatrics, 142*(4).

Girolametto, L., & Weitzman, E. (2002). Responsiveness of child care providers in interactions with toddlers and preschoolers.

Greenhoot, A. F., Beyer, A. M., & Curtis, J. (2014). More than pretty pictures? How illustrations affect parent-child story reading and children's story recall. *Frontiers in Psychology, 5,* 738.

Grolig, L., Cohrdes, C., Tiffin-Richards, S. P., & Schroeder, S. (2020). Narrative dialogic reading with wordless picture books: A cluster-randomized intervention study. *Early Childhood Research Quarterly, 51,* 191-203.

Hart, B., & Risley, T. R. (1989). The longitudinal study of interactive systems. *Education and Treatment of Children,* 347-358.

Hayes, D. P., & Ahrens, M. G. (1988). Vocabulary simplification for children: A special case of 'motherese'?. *Journal of Child Language, 15*(2), 395-410.

Haynes, W. O., & Saunders, D. J. (1998). Joint book-reading strategies in middle-class African American and white mother–toddler dyads: Research note. *Journal of Children's Communication Development,* 20, 9–17.

Heaps, H. S. (1978). Information retrieval. Computational and theoretical aspects. New York: Academic Press.

Herdan, G. (1960). Type-token mathematics. Vol. 4. The Hague: Mouton.

Hindman, A. H., Skibbe, L. E., & Foster, T. D. (2014). Exploring the variety of parental talk during shared book reading and its contributions to preschool language and literacy: Evidence from the Early Childhood Longitudinal Study-Birth Cohort. *Reading and Writing, 27*(2), 287-313.

Hoff-Ginsberg, E. (1991). Mother-child conversation in different social classes and communicative settings. *Child Development, 62*(4), 782-796.

Hsiao, Y., Dawson, N. J., Banerji, N., & Nation, K. (2022). The nature and frequency of relative clauses in the language children hear and the language children read: A developmental cross-corpus analysis of English complex grammar. *Journal of Child Language,* 1-26.

Hunt, K. W. (1970). Syntactic maturity in schoolchildren and adults. *Monographs of the society for research in child development, 35*(1), iii-67.

Huttenlocher, J., Waterfall, H., Vasilyeva, M., Vevea, J., & Hedges, L. V. (2010). Sources of variability in children's language growth. *Cognitive psychology, 61*(4), 343-365.

Justice, L. M., & Ezell, H. K. (2000). Enhancing children's print and word awareness through home-based parent intervention. *American Journal of Speech-Language Pathology, 9*(3), 257-269.

Kam, C. L. H., & Matthewson, L. (2017). Introducing the infant book reading database (IBDb). *Journal of Child Language, 44*(6), 1289-1308.

Karasik, L. B., Tamis-LeMonda, C. S., & Adolph, K. E. (2014). Crawling and walking infants elicit different verbal responses from mothers. *Developmental Science, 17*(3), 388-395.

Karrass, J., & Braungart-Rieker, J. M. (2005). Effects of shared parent–infant book reading on early language acquisition. *Journal of Applied Developmental Psychology, 26*(2), 133-148.

Kretch, K. S., Franchak, J. M., & Adolph, K. E. (2014). Crawling and walking infants see the world differently. *Child Development, 85*(4), 1503-1518.

Leech, K. A., & Rowe, M. L. (2014). A comparison of preschool children's discussions with parents during picture book and chapter book reading. *First Language, 34*(3), 205-226.

Leung, A., Tunkel, A., & Yurovsky, D. (2021). Parents fine-tune their speech to children's vocabulary knowledge. *Psychological Science, 32*(7), 975-984.

Lonigan, C. J., Burgess, S. R., & Anthony, J. L. (2000). Development of emergent literacy and early reading skills in preschool children: evidence from a latent-variable longitudinal study. *Developmental Psychology, 36*(5), 596.

Lonigan, C. J., & Whitehurst, G. J. (1998). Relative efficacy of parent and teacher involvement in a shared-reading intervention for preschool children from low-income backgrounds. *Early Childhood Research Quarterly, 13*(2), 263-290.

Lyytinen, P., Laakso, M. L., & Poikkeus, A. M. (1998). Parental contribution to child's early language and interest in books. *European Journal of Psychology of Education, 13*(3), 297-308.

MacWhinney, B. (2000). The CHILDES Project: Tools for Analyzing Talk. 3rd Edition. Mahwah, NJ: Lawrence Erlbaum Associates

Malvern, D., Richards, B. J., & Chipere, N. (2004). Lexical diversity and language development: Quantification and assessment. Basingstoke, UK: Palgrave Macmillan.

Manz, P. H., Hughes, C., Barnabas, E., Bracaliello, C., & Ginsburg-Block, M. (2010). A descriptive review and meta-analysis of family-based emergent literacy interventions: To what extent is the research applicable to low-income, ethnic-minority or linguistically-diverse young children?. *Early Childhood Research Quarterly, 25*(4), 409-431.

Massaro, D. W. (2015). Two different communication genres and implications for vocabulary development and learning to read. *Journal of Literacy Research, 47*(4), 505-527.

Massaro, D. W. (2017). Reading aloud to children: Benefits and implications for acquiring literacy before schooling begins. *The American Journal of Psychology, 130*(1), 63-72.

McKee, G., Malvern, D., & Richards, B. (2000). Measuring vocabulary diversity using dedicated software. Literary and Linguistic Computing, 15, 323–338.

Mol, S. E., Bus, A. G., De Jong, M. T., & Smeets, D. J. (2008). Added value of dialogic parent–child book readings: A meta-analysis. *Early Education and Development, 19*(1), 7-26.

Mol, S. E., & Neuman, S. B. (2014). Sharing information books with kindergartners: The role of parents' extra-textual talk and socioeconomic status. *Early Childhood Research Quarterly, 29*(4), 399-410.

Montag, J. L. (2019). Differences in sentence complexity in the text of children's picture books and child-directed speech. *First Language, 39*(5), 527-546

Montag, J. L., Jones, M. N., & Smith, L. B. (2015). The words children hear: Picture books and the statistics for language learning. *Psychological Science, 26*(9), 1489-1496.

Montag, J. L., Jones, M. N., & Smith, L. B. (2018). Quantity and diversity: Simulating early word learning environments. *Cognitive Science, 42*, 375-412.

Muhinyi, A., & Hesketh, A. (2017). Low-and high-text books facilitate the same amount and quality of extratextual talk. *First Language, 37*(4), 410-427.

Muhinyi, A., Hesketh, A., Stewart, A. J., & Rowland, C. F. (2020). Story choice matters for caregiver extra-textual talk during shared reading with preschoolers. *Journal of Child Language, 47*(3), 633-654.

Muhinyi, A., & Rowe, M. L. (2019). Shared reading with preverbal infants and later language development. *Journal of Applied Developmental Psychology, 64*, 101053.

Nelson, K. (1977). First steps in language acquisition. *Journal of the American Academy of Child Psychiatry, 16*(4), 563-583.

Ninio, A. (1983). Joint book reading as a multiple vocabulary acquisition device. *Developmental Psychology, 19*(3), 445.

Ninio, A., & Bruner, J. (1978). The achievement and antecedents of labelling. *Journal of Child Language, 5*(1), 1-15.

Noble, C., Cameron-Faulkner, T., Jessop, A., Coates, A., Sawyer, H., Taylor-Ims, R., & Rowland, C. F. (2020). The impact of interactive shared book reading on children's language skills: A randomized controlled trial. *Journal of Speech, Language, and Hearing Research, 63*(6), 1878-1897.

Noble, C., Sala, G., Peter, M., Lingwood, J., Rowland, C., Gobet, F., & Pine, J. (2019). The impact of shared book reading on children's language skills: A meta-analysis. *Educational Research Review, 28*, 100290.

Oakes, L. M. (2023). The development of visual attention in infancy: A cascade approach. In *Advances in Child Development and Behaviour, 64*, 1-37.

Parker, M. D., & Brorson, K. (2005). A comparative study between mean length of utterance in morphemes (MLUm) and mean length of utterance in words (MLUw). *First Language, 25*(3), 365-376.

Patel, T. N., Marasli, Z. B., Choi, A., & Montag, J. L. (2024). An Online Survey of Picture Book Reading Practices with Children Between the Ages of 0 and 30 Months. *Language Learning and Development*, 1-27.

Payne, A. C., Whitehurst, G. J., & Angell, A. L. (1994). The role of home literacy environment in the development of language ability in preschool children from low-income families. *Early Childhood Research Quarterly, 9*(3-4), 427-440.

Petrie, A., Robert, M. A. Y. R., Fei, Z. H. A. O., & Montanari, S. (2021). Parent-child interaction during storybook reading: wordless narrative books versus books with text. *Journal of Child Language*, 1-28.

Price, L. H., Van Kleeck, A., & Huberty, C. J. (2009). Talk during book sharing between parents and preschool children: A comparison between storybook and expository book conditions. *Reading research quarterly, 44*(2), 171-194.

Raikes, H., Alexander Pan, B., Luze, G., Tamis-LeMonda, C. S., Brooks-Gunn, J., Constantine, J., ... & Rodriguez, E. T. (2006). Mother–child book reading in low-income families: Correlates and outcomes during the first three years of life. *Child Development, 77*(4), 924-953.

Read, K., Macauley, M., & Furay, E. (2014). The Seuss boost: Rhyme helps children retain words from shared storybook reading. *First Language, 34*(4), 354-371.

Read, K., Rabinowitz, S., & Harrison, H. (2023). It's the talk that counts: A review of how the extra-textual talk of caregivers during shared book reading with young children has been categorized and measured. *Journal of Early Childhood Literacy*, doi:10.1177/14687984231202968.

Richards, B. (1987). Type/token ratios: What do they really tell us? *Journal of Child Language, 14,* 201–209.

Romeo, R. R., Leonard, J. A., Robinson, S. T., West, M. R., Mackey, A. P., Rowe, M. L., & Gabrieli, J. D. (2018). Beyond the 30-million-word gap: Children's conversational exposure is associated with language-related brain function. *Psychological Science, 29*(5), 700-710.

Sala, G., & Gobet, F. (2017). Working memory training in typically developing children: A meta-analysis of the available evidence. *Developmental Psychology, 53*(4), 671.

Salo, V. C., Rowe, M. L., Leech, K. A., & Cabrera, N. J. (2016). Low-income fathers' speech to toddlers during book reading versus toy play. *Journal of Child Language, 43*(6), 1385-1399.

Saracho, O. N. (2017). Literacy and language: new developments in research, theory, and practice. *Early Child Development and Care, 187*(3-4), 299-304.

Scarborough, H. S., Dobrich, W., & Hager, M. (1991). Preschool literacy experience and later reading achievement. *Journal of learning Disabilities, 24*(8), 508-511.

Sénéchal, M., Cornell, E. H., & Broda, L. S. (1995). Age-related differences in the organization of parent-infant interactions during picture-book reading. *Early Childhood Research Quarterly, 10*(3), 317-337.

Sénéchal, M., & LeFevre, J. A. (2002). Parental involvement in the development of children's reading skill: A five-year longitudinal study. *Child Development, 73*(2), 445-460.

Shahaeian, A., Wang, C., Tucker-Drob, E., Geiger, V., Bus, A. G., & Harrison, L. J. (2018). Early shared reading, socioeconomic status, and children's cognitive and school competencies: Six years of longitudinal evidence. *Scientific Studies of Reading, 22*(6), 485-502.

Simons, D. J., Boot, W. R., Charness, N., Gathercole, S. E., Chabris, C. F., Hambrick, D. Z., & Stine-Morrow, E. A. (2016). Do "brain-training" programs work?. *Psychological Science in the Public Interest, 17*(3), 103-186.

Soderstrom, M., Casillas, M., Bergelson, E., Rosemberg, C., Alam, F., Warlaumont, A. S., & Bunce, J. (2021). Developing A Cross-Cultural Annotation System and MetaCorpus for Studying Infants' Real World Language Experience. *Collabra: Psychology, 7(1),* 23445.

Sosa, A. V. (2016). Association of the type of toy used during play with the quantity and quality of parent-infant communication. *JAMA Pediatrics,* 170, 132-137.

Suppes, P. (1974). The semantics of children's language. *American Psychologist, 29,* 103–114.

Thelen, E. and Smith, L.B. (1994) *A Dynamic Systems Approach to the Development of Cognition and Action.* MIT Press

Valdez-Menchaca, M. C., & Whitehurst, G. J. (1992). Accelerating language development through picture book reading: A systematic extension to Mexican day care. *Developmental Psychology, 28*(6), 1106.

Weizman, Z. O., & Snow, C. E. (2001). Lexical output as related to children's vocabulary acquisition: Effects of sophisticated exposure and support for meaning. *Developmental Psychology, 37*(2), 265.

Whitehurst, G. J., Epstein, J. N., Angell, A. L., Payne, A. C., Crone, D. A., & Fischel, J. E. (1994). Outcomes of an emergent literacy intervention in Head Start. *Journal of Educational Psychology, 86*(4), 542.

Whitehurst, G. J., Falco, F. L., Lonigan, C. J., Fischel, J. E., DeBaryshe, B. D., Valdez-Menchaca, M. C., & Caulfield, M. (1988). Accelerating language development through picture book reading. *Developmental Psychology, 24*(4), 552.

Whitehurst, G. J., Zevenbergen, A. A., Crone, D. A., Schultz, M. D., Velting, O. N., & Fischel, J. E. (1999). Outcomes of an emergent literacy intervention from Head Start through second grade. *Journal of Educational Psychology, 91*(2), 261.

Yarosz, D. J., & Barnett, W. S. (2001). Who reads to young children?: Identifying predictors of family reading activities. *Reading Psychology, 22*(1), 67-81.

Young, K. T., Davis, K., Schoen, C., & Parker, S. (1998). Listening to parents: a national survey of parents with young children. *Archives of Pediatrics & Adolescent Medicine, 152*(3), 255-262.

Data, code and materials availability statement

The data and analytical scripts used in the study are available at **https://osf.io/b3egw** Audio files and transcripts in ELAN and CHAT formats are available at https://childes.talkbank.org/access/Eng-NA/StoopsMontag.html

Ethics statement

The research reported was conducted in compliance with APA Ethical Standards regarding the treatment of human participants and the University of Illinois IRB protocols.

Authorship and Contributorship Statement

JLM conceived and designed the study and co-wrote and revised the manuscript with

AS. AS contributed to the design of the study, collected and analyzed the data and wrote the manuscript. MW and IHTJ analyzed portions of the data. All authors approved the final version of the manuscript and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Declaration of conflict of interests

We report having no competing interests.

Acknowledgements

We thank the undergraduate research assistant of Learning and Language Lab at the University of Illinois at Urbana-Champaign for transcription of the audio recordings. This research was supported by National Science Foundation grant (BCS-1749594), National Institute of Health (R03 HD096157), and a James S McDonnell Foundation Scholar Award to Jessica L. Montag.

License

Children with Developmental Language Disorder and typically developing children learn novel nouns more easily than novel verbs: An experimental comprehension and production study.

Paula Stinson
Julian Pine
University of Liverpool, UK

Abstract: Previous research suggests that nouns are generally learned more easily than verbs. However, few studies have investigated this issue amongst children with Developmental Language Disorder (DLD) or presented novel verbs and nouns in comparable training contexts. The present study therefore compared noun and verb learning in 18 Children with DLD and 36 Typically developing (TD) children aged from 3;1 (years; months) to 4;10. Participants were presented with two short cartoon videos that were dubbed with an audio script containing three novel nouns and three novel verbs (six novel nouns and six novel verbs in total). Children completed a comprehension and production task both immediately post-test and in a retention follow-up three to five days later. The TD children outperformed the children with DLD on both comprehension and production (though only in the retention test session and not the immediate test session). Although a noun advantage was observed, there was no evidence that its magnitude differs between TD children and children with DLD.

Keywords: Developmental Language Disorder (DLD); Word Learning

Corresponding author: Paula Stinson, University of Liverpool, 74 Bedford Street South, Liverpool, L69 7ZA, UK. Email: paulamcl@liverpool.ac.uk

Citation: Stinson, P., & Pine, J., (2024) Children with Developmental Language Disorder and typically developing children learn novel nouns more easily than novel verbs: An experimental comprehension and production study. *Language Development Research, 4*(1), 298—325. http://doi.org/10.34842/jmgp-2v74

Introduction

Developmental Language Disorder (DLD) is a condition that affects around 7% of the population (Norbury et al., 2016) with a prevalence of 8% for boys and 6% for girls (Tomblin et al. 1997). It is a neurodevelopmental disorder that is characterised by persistent and significant language difficulties that cannot be attributed to hearing loss or neurological damage. (Leonard, 2014: 3, Bishop, 2017). DLD is not a homogeneous disorder and individuals with this condition present with varied profiles of impairment in oral language and cognitive skills (Bishop, 2006). Areas of weakness in cognitive skills can include difficulties in attention, memory, problem solving and reasoning. Areas of weakness in oral language can include difficulties with phonology, morphology, syntax, semantics, and pragmatics. However, a key impairment in DLD is a deficit in the ability to learn new words, and this is the focus of the present study.

The Word-Learning Deficit in DLD

Word learning is a primary building block for acquiring language. Amongst Typically Developing (TD) children, there is evidence that early vocabulary size is a significant predictor of later grammatical development and literacy (Lee, 2010) and research suggests a link between vocabulary and academic achievement (Castles et al., 2018).

Children with DLD have consistent deficits in learning novel lexical labels (Alt & Plante, 2006) and storing new vocabulary compared with TD children (Gray, 2004, McGregor et al., 2011). Their vocabularies tend to show less breadth and depth in comparison to those of their age-matched peers (Dollaghan, 1998; Kail & Leonard, 1986, McGregor, Oleson, Bahnsen, & Duff, 2013) and numerous novel word learning studies have found that children with DLD require more encounters with a word before learning takes place (Alt, 2011; Alt & Plante, 2006; Alt, Plante, & Creusere, 2004; Gray, 2003, 2004; Gray, Pittman, & Weinhold, 2014; McGregor, Licandro, et al., 2013).

To learn a novel word, a child must build a phonological representation, a semantic representation, and make an association between the two (Chiat, 2001). A breakdown at any of these stages may impact the child's ability to recognise and further refine their word knowledge (Gray et al., 2020). McGregor et al. (2020) describe the process of learning a new word as involving encoding, re-encoding, and retention. Through this process, a child will learn a new word and store it in their memory where the lexical entry will be further built upon and refined. For some children, though, a breakdown in this process means that word learning is particularly difficult, although the reasons for this difficulty are still unclear.

One line of research suggests that children with DLD have difficulty with word learning due to impaired encoding (McGregor et al., 2013). There is evidence that children with DLD have difficulty encoding the phonological information in words (Bishop, North, & Donlan, 1996, Edwards & Lahey, 1998). In a series of studies with adolescents and adults with DLD, McGregor et al., (2013, 2017) developed the 'encoding deficit hypothesis'. Participants were trained on novel words and their associated referents, and it was found that children with DLD had poorer performance on the immediate post-training tasks, but retention of the word seemed intact after one week. However, the gap between TD participants and those with DLD widened over time, indicating a potential problem with the retention of phonological information. McGregor et al.'s subsequent studies controlled for confounds with retention and concluded that "encoding of word form is the primary bottleneck to word learning among people with DLD" (McGregor et al., 2020:14).

Bishop and Hsu (2015) found similar results. They compared eight-year-old children with and without DLD and found that the children with DLD showed significant difficulty with the word learning task post training, but after two weeks both groups performed at a similar level suggesting that the children with DLD may have had difficulty with encoding the information but not with retention (see also Leonard et al, 2019, for a similar finding with five-year-olds).

However, one potential problem with this conclusion is that in many studies that consider retention, the performance of both TD children and children with DLD is so poor that it is difficult to tell whether there is really no retention deficit in DLD, or just a floor effect in the data. For example, Jackson et al. (2020) looked at six-year-old TD children and those with DLD. Their study involved teaching the children eight novel words over a four-day period and considered encoding, re-encoding, and retention abilities in both groups. Their findings suggested that children with DLD have difficulty with word learning in comparison to their TD peers and were consistent with the idea that these difficulties were due to encoding rather that retention, but both groups performed so poorly on retention that it was impossible to rule out an additional retention deficit that was hidden by the floor effect in the data.

Noun and Verb Learning

The above research provides clear evidence for a word learning deficit in children with DLD. However, much of the current research on word learning has focussed on noun learning to the exclusion of other kinds of word learning (e.g., Kan & Windsor, 2010). Studies with TD children have shown that nouns tend to be easier to learn than verbs (e.g., Bornstein, 2005; Gentner, 1982). While there is some debate concerning the cross-linguistic data, (e.g., Choi & Gopnik, 1995; Tardif, Gelman, & Xu, 1999), the majority of research supports the idea that the noun advantage is a universal trend across languages (Au, Dapretto & Song, 1994; Bird, Franklin & Howard, 2001; Bornstein et al., 2004; Gentner & Boroditsky, 2008; Gillette, Gleitman, Gleitman & Lederer, 1999; Kauschke, Lee & Pae, 2007; Kim, McGregor & Thompson, 2000; Snedeker & Gleitman, 2004). Several theories as to the source of the noun advantage in early acquisition have been proposed. Some researchers have suggested that parental input and frequency play an important role (Barrett, Harris & Chasin,1991). Chan, Brandone and Tardif (2009) demonstrated that parents speaking a noun-privileged language such as English produced more nouns than verbs when speaking to their children. Goodman, Dale, and Li (2008) also suggested that frequency may play an important role in early acquisition. However, other research has shown that noun dominance cannot be attributed solely to frequency as nouns are learned more easily than verbs even when input frequency is controlled (Imai, Haryu, & Okada, 2005; Leonard, Schwartz, Morris, & Chapman, 1981; Merriman, Marazita, & Jarvis, 1993; Rice & Woodsmall, 1988). McDonough et al. (2011) found that English-speaking parents tend to request that their children repeat noun labels but act out verb meanings (Goldfield, 2000; Tardif et al., 2005), and that children prefer to attend to objects and map new names to objects rather than to actions. Some researchers have argued that nouns are more readily learned because the concepts to which they refer are more available to young learners than the concepts to which verbs refer (e.g., Byrnes & Gelman, 1991; Gentner, 1978; Gopnik & Meltzoff, 1986; Smiley & Huttenlocher,1995). Another possible explanation for the noun advantage is that, while objects are generally stable, actions are often fleeting (Gentner, 1982); thus, nouns tend to be more concrete, imageable, and more easily identifiable than verbs (McDonough et al., 2011). Salience and iconicity have also been shown to play a role (Hills, Maouene, Maouene, Sheya, & Smith, 2009; Perry, Perlman, & Lupyan, 2015; Roy, Frank, DeCamp, Miller, & Roy, 2015; Swingley & Humphrey, 2018). Other researchers have pointed out that the meaning of a concrete noun can often be inferred from the context in which it is uttered; however, the meaning of a verb depends more heavily on syntactic information and other linguistic cues. For example, in the study of Gillette et al., (1999), participants watched a mother-child interaction where both nouns and verbs were 'bleeped out' and were asked to guess the missing word. Participants were more able to guess the nouns than the verbs, which suggests that imageability played a significant role in the outcome.

A potential confounding factor when testing children's knowledge of noun and verb acquisition is that the way in which test items are typically presented may favour nouns over verbs. Nouns are typically presented in their stable state in both the testing session and the preceding training session, whereas in many word learning studies, verbs are presented in a stable state (i.e., using still pictures) in the testing session, but dynamically during training. In the present study, we address this potential confound by presenting the verbs as dynamic rather than static, using animations at test. This ensures that the actions to which the verbs refer are presented in the same way during testing and training.

The fact that children with DLD have been shown to have difficulties with word learning in general, combined with the general difficulty of verb learning, raises the question of whether children with DLD may find verb learning particularly challenging. Children with DLD may finding verb learning even more challenging than their TD peers because verb learning may require stronger abilities in phonology and semantics, and greater awareness of the links between these for effective learning, which is a known area of difficulty for these children (Wright, Pring & Ebbels, 2018). As Wright et al. point out, there are two possible reasons why phonology may impact on verb learning. First, in continuous speech, verbs are less stressed than nouns, making the phonological sequence more difficult to identify and store. Secondly, as verbs have more complex morphology, the phonological form of the verbs a child hears will be more variable than the phonological form of the nouns, which increases the complexity of the extraction process. In terms of verb semantics, as verbs only appear in particular sentence structures, a child can use the sentence structure a verb appears in to aid their hypotheses regarding the meaning of a new verb. However, Van der Lely (1994) argues that children with DLD may have more difficulties than TD children in using this kind of information. Other research has suggested that children with DLD may have difficulty with verb learning because the child's current verb lexicon, which tends to be reduced in comparison to their TD peers, will have less learned examples of 'verb types' and this will impact the ability to learn novel verbs (Windfuhr, Faragher & Conti-Ramsden, 2002).

In fact, there is already some evidence that verb learning may be a particular problem for children with DLD. For example, children with DLD have been shown to use a narrower range of verbs in their speech, and to overuse a small set of general all-purpose (GAP) verbs such as *go* and *do* (Rice & Bode, 1993) in comparison to TD children. However, the results of studies of novel noun and verb learning with this population have been mixed, with some studies finding that children with DLD had particular difficulty learning verbs (e.g., Oetting et al., 1995), while others have not (Rice et al. 1994, Rice, Buhr, & Nemeth, 1990; Rice & Woodsmall, 1988). Determining whether children with DLD have a particular problem with verb learning may thus have clinical implications for how these children are assessed and treated. For example, a greater difficult with verb than noun learning is likely to impact on the ease with which children with DLD master verb morphology which is often an area addressed in assessment and treatment in the clinic.

The Present Study

In view of the considerations discussed above, the aim of the present study is to compare noun and verb learning in children with DLD and age-matched controls in an ecologically valid word learning task. The study investigates the impact of encoding and retention difficulties on the word learning deficit by testing comprehension and production both immediately after presentation and three to five days later.

We use a design adapted from Rice et al. (1990), in which novel (i.e., non-word) nouns and verbs are embedded in the narrative script of a short video. Children are shown a short video with a dubbed audio script. Children are then tested in a format similar to the Peabody Picture Vocabulary Test- Revised (PPVT-R), in which they are asked to select the correct response from a choice of four pictures/animations (comprehension test) and to produce the word when shown the appropriate picture/animation (production test). This allowed us to investigate the extent to which both encoding and retention vary as a function of group (DLD/TD) and word class (noun/verb). We predicted: (1) that the children with DLD would perform significantly worse than the control group; (2) that both groups would perform significantly better on Nouns than Verbs: and (3) that the size of the noun advantage would be larger for the children with DLD than the TD children.

Methods

Participants

Participants were recruited from a range of nurseries across Northern Ireland. Children were identified by teachers and parents as appropriate for the study, in that they were monolingual English speakers of the relevant age group. The children did not have an existing diagnosis of DLD but were assigned to a group (DLD/TD) on the basis of the standardised tests conducted as part of the study. A power analysis was carried out (using GPower) prior to recruiting participants, assuming 0.8 power and an effect size of 0.6, on the basis of the following studies: Gray (2003, 2004 and 2006) and Rice, M. L., Buhr, J. C., and Nemeth, M. (1990). This resulted in us aiming to test ninety children: (forty-five children in the DLD group and forty-five children in the TD control group). Due to the difficulties identifying and recruiting children with DLD, in total, seventy-six children were tested, but five were excluded as they did not complete all the tasks. A further seventeen children were excluded because they had been identified by their class teachers as suitable for the study, but after assessment had been completed, did not meet the strict criteria for either the DLD or the control group.

The Groups were defined as follows:

1. The children with DLD scored 1.5 or more standard deviations below the mean on the Core Language score (CLS) of the Pre-School Clinical Evaluations of Language Fundamentals 2 (PS-CELF2); a composite of three sub-tests looking at Comprehension and Production of Language; Expressive Vocabulary, Word Structure and Understanding of Sentence Structure. For assignment to the DLD group, children were required to score within 1 Standard Deviation on a cognitive assessment: the Non-Verbal Index of the Kaufman Assessment Battery for Children II (K-ABC II)
2. The TD children were defined as scoring within 1 SD of the mean on the CLS and the Non-verbal assessment of the KABCII.
3. Children in both groups passed a hearing screening administered by the researcher, and parents reported no neurological or genetic conditions, or issues with the children's motor skills.

Application of these criteria resulted in a sample size of N=54, 36 TD children (control group), 18 children with DLD (experimental group). The children's ages ranged from 3;0 (years; months) to 4;8. Table 1 below provides more detailed descriptive data on the ages of the children.

Children were deemed by the researcher, a qualified Speech and Language Therapist, to have adequate phonology to participate in the study (i.e., to not have a disordered phonological profile). Children were accepted into the study if they presented with no phonological errors or if the errors that they made were developmentally appropriate as specified by Grunwell's (1987) Common phonological processes and their approximate ages of elimination in typical acquisition. In practice, no errors that would have made it difficult to determine the accuracy of responses were observed, with children either producing the target form or failing to respond.

Table 1. ***Details of age range***

	Age Range	**Mean**	**SD**	**No. of children aged 3:0yrs-3:5yrs**	**No. of children aged 3:6yrs-3:11yrs**	**No. of children aged 4:0yrs-4:5yrs**	**No. of children aged 4:6yrs-4:11yrs**
Combined TD and DLD	3yrs 0mths-4yrs 8mths 20 months	3yrs 9mths 45.93mths	5.63				
TD	3yrs 0mths-4yrs 8mths 20 Months	3yrs 10mths 46.03mths	5.82	11	8	12	5
DLD	3yrs 0mths-4yrs 5mths 17 Months	3yrs 9mths 45.72mths	5.24	5	3	9	1

Design and Procedure

The novel words were presented within two short Pixar cartoon videos which children watched on a laptop computer. Each cartoon was dubbed with an audio script (Appendix 1 and 2) containing three novel nouns and three novel verbs, meaning that each child heard a total of six novel nouns and six novel verbs. (See Figure 1 for details).

Six of the non-words were one syllable long and six were two syllables long. Each word was heard a total of four times.

Novel words were based on real nouns and verbs that would be familiar to children of this age group, according to the UK Communicative Development Inventory (UK-CDI). The words were manipulated to create non-words by altering the initial phoneme to a labial or alveolar sound (i.e., /p/, /b/, /m/, /n/, /t/ and /d/), which are sounds that children of the age group that was tested are typically able to produce.

The novel words were embedded in a relatively syntactically complex script as there is evidence that children use syntax to guide verb learning in a process known as syntactic bootstrapping (Fisher et al., 1991; Levin & Rappaport-Hovav, 2005; Pinker, 1989).

Novel Noun Words	Object	Novel Verb Words	Action	Video
Nall	Tin toy's hat	Diting	Tin toy walking and playing music	Tin Toy
Mot	Drum Type Toy	Tuddling	Baby waving their arms and legs	Tin Toy
Poffee	Beads	Bickling	Spinning in circles	Tin Toy
Dut	Small alien	Miping	Ship warping	Lifted
Bettle	Driving handlebars	Nuving	Man levitating	Lifted
Tellon	Big Alien	Povering	Big alien moving fingers	Lifted

Figure 1. ***Novel Nouns and Verbs Details***

The videos were divided into an A and a B group, where nouns and verbs in the testing were presented in a different alternating order (i.e., noun-verb-noun or verb-noun-verb). The order in which each condition was presented to the children was randomized.

The comprehension and production tasks were administered immediately after viewing and again (for the retention test) three to five days later. The comprehension task involved showing the child a choice of four pictures representing each noun or four short video clips (less than 2 seconds) for each verb (see Figure 2). The experimenter's probe was as follows:

Inv: *'Where's the Dut?'*
Child*: (points)*

Responses were recorded manually by the researcher; 1 as correct, 0 as incorrect/no response. The production task involved showing the child a picture of the novel noun or a short video clip of the novel verb and prompting the child as follows:

Inv: *'What's he doing?*
Child*: povering*

Verbs were presented in present progressive and past tense forms in the video as this is the most natural way to describe ongoing and completed actions. Responses were considered correct if the child produced the verb in the present (e.g., povers), progressive (e.g., povering), past tense (e.g., povered) or bare stem (e.g., pover) form.

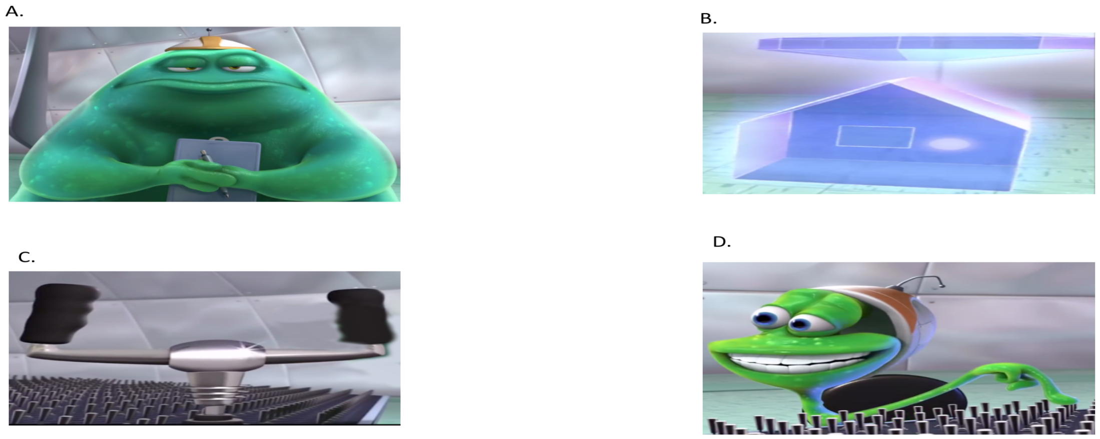


Figure 2 ***Comprehension Task Example***

Word Learning Task

Each child was tested individually in a quiet setting, with each session lasting approximately 30 to 45 minutes. Testing was divided into three sessions on three different days. The first two sessions were completed within 24 hours of each other, and the child was shown one video and completed the tasks on each day. The third session tested retention and was administered three to five days later.

On Day 1, children completed the sentence-structure and word-structure tests from the PS-CELF 2, the first two subtests from the K-ABC-II and twelve trials from the main study. On Day 2, children completed the expressive vocabulary test from the PS-CELF 2, the remaining two subtests from the K-ABC-II and a further 12 trials from the main study. On Day 3, Children were asked to complete all 24 retention trials without watching the previously seen videos. All responses were manually recorded by the researcher.

Analyses

The data were analysed using 'R' (version 1.4.1717; R version 4.1.0, R Core Team, 2018), with the packages lme4 (v1. 1-26; Bates et al., 2015) and yarr (v0.1.5; Philips., 2022). Mixed-effects models were used because the data had more than one source of random variability: participants and items. The dependent variable – for both the production and comprehension analyses - was whether the trial was completed correctly (1) or not (0). Predictor variables were Group (DLD/TD), sum coded as -0.5 and 0.5 respectively, and Part of Speech (Noun/Verb), sum coded as -0.5 and 0.5 respectively. Note that sum coding is crucial here in order to ensure that any effects of Group and Part of Speech can be interpreted as "ANOVA style" main effects.

Results

Comprehension

The data from the Immediate and Retention comprehension sessions were analysed using mixed-effects models where the dependent variable was response (1=correct, 0=incorrect) and the predictor variables were Group (DLD=-0.5, TD=0.5) and Part of Speech (Noun=-.05, Verb=0.5), both sum-coded. Age in months (raw, not scaled or centred) was included as a control predictor. Following Matuschek et al. (2017), we built models with all possible random effects structures that were justified given the data and chose the model with the lowest BIC value.

For the Immediate test session, the model with the following effects structure had the lowest BIC value and so was selected as the final model (note that this model does not include correlated random effects):

```
glmer(Response ~ Part_of_Speech*Group + Age + (1|Participant) + (1|Lexical_Item), data=subset(First, Test_Type=="Comprehension"), family=binomial, glmerControl(optimizer = "bobyqa"))
```

A summary of this model is shown in Table 2 (see the accompanying OSF site for the full model output, including estimated random effect variances). This analysis revealed a significant effect of Part of Speech, reflecting better overall performance for Nouns than Verbs (recall from the analysis section that, because sum coding was used, inferences regarding main effects can be made based directly from these fixed effect terms). However, the effect of Group (DLD/TD) was not significant. Nor was the interaction of Part of Speech x Group. We thus have no evidence that the Noun advantage is greater in children with DLD. A significant positive effect of age was observed, indicating that performance improves with age.

Table 2. ***Mixed-effects model for the Immediate comprehension session.***

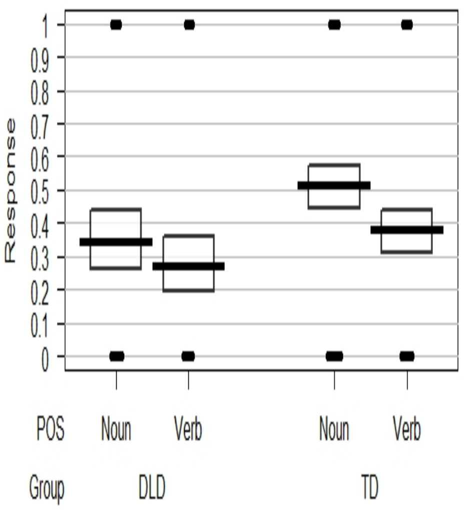
	Estimate	**Std. Error**	**z value**	**P Value**
Intercept	-2.58	0.98	-2.64	0.008 **
Part of Speech	-0.47	0.21	-2.20	0.028 *
Group1	0.40	0.23	1.77	0.078
Age	0.50	0.24	2.10	0.036 *
Part_of_Speech1:Group1	-0.21	0.362	-0.58	0.565

For the Retention comprehension test, the final, best fitting model had the same random effects structure as for the Immediate test session. A summary of this model is shown in Table 3 (again see the accompanying OSF site for the full model output). This time, the effect of Part of Speech was not significant, but a significant effect of Group indicated that the TD children outperformed the children with DLD. Crucially, the interaction of Part of Speech x Group was again not significant. We thus have no evidence to suggest that any Noun advantage (though none was observed in the Retention session) is greater for the children with DLD.

Table 3. ***Mixed-effects model for the Retention comprehension session.***

	Estimate	**Std. Error**	**z value**	**P Value**
Intercept	0.36	1.21	0.30	0.766
Part of Speech	-0.38	0.30	-1.27	0.203
Group1	0.98	0.29	3.42	0.001 ***
Age	-0.30	0.30	-1.00	0.317
Part_of_Speech1:Group1	0.30	0.39	0.77	0.441

These results are plotted in Figures 3 and 4. Figure 3 summarizes the proportion of correct comprehension responses for the children with and without DLD in the immediate test. Figure 4 summarizes the same data for the retention session (3-5 days later).



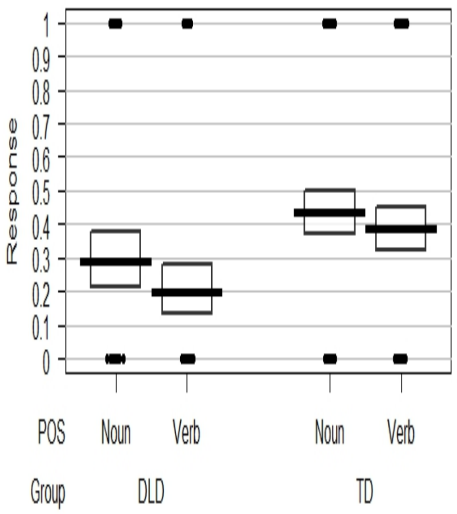


Figure 3: ***Proportion of correct choices by Part of Speech (Noun/Verb) and Group (DLD/TD) in the Immediate comprehension session***
Figure 4***: Proportion of correct choices by Part of Speech (Noun/Verb) and Group (DLD/TD) in the Retention comprehension session***

Consistent with the results of the mixed-effect analyses, these figures suggest a general advantage for Nouns over Verbs, which is greater in the immediate recall session and a general advantage for the TD children, which is greater in the retention session. They also provide no evidence of an interaction between Group and Part of Speech, though there is some suggestion that the difference in Verb learning between the two groups is greater in the retention than the immediate recall session.

Production

For the production data, mixed-effects models were run in the same way as for the comprehension data, though with the dependent variable as correct versus incorrect productions, and the same model structure was again optimal by BIC (see Tables 4-5). In this case, significant effects of Group (TD>DLD) were observed for the Retention production session (Table 5), though not the Immediate production session (Table 4). However, there was no effect of Part of Speech in either case. For both datasets, a significant positive effect of age was observed, indicating that performance improves with age. Crucially, though – just as for the comprehension data – the interaction of Part of Speech x Group was not significant in any analysis. We thus have no evidence to suggest that any Noun advantage is greater for the children with DLD.

Table 4. ***Mixed-effects model for the Immediate production session.***

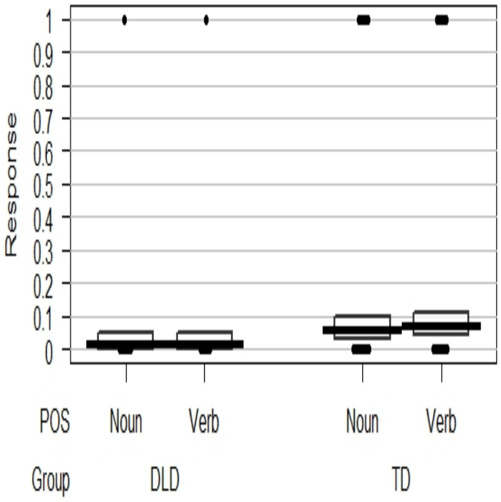
	Estimate	**Std. Error**	**z value**	**P Value**
Intercept	-13.98	3.75	3.73	0.000 ***
Part of Speech	0.19	0.82	0.23	0.815
Group1	1.08	0.88	1.24	0.215
Age	2.34	0.87	2.70	0.007 **
Part of Speech: Group1	0.25	1.49	0.17	0.865

Table 5. ***Mixed-effects model for the Retention production session***

	Estimate	**Std. Error**	**z value**	**P Value**
Intercept	-7.15	2.72	-2.63	0.009 **
Part of Speech	0.59	0.89	0.66	0.509
Group1	1.63	0.75	2.17	0.030 *
Age	0.74	0.64	1.16	0.247
Part of Speech: Group1	-1.11	1.31	0.85	0.396

These results are plotted in Figures 5 and 6. Figure 5 summarizes the proportion of correct production responses for the children with and without DLD in the immediate test session. Figure 6 plots the same data for the Retention production session.

Consistent with the results of the mixed-effects analyses, these figures provide no evidence of a Noun advantage in either group, but they also show that both groups were essentially at floor in both production sessions. The absence of such an effect is therefore unsurprising.



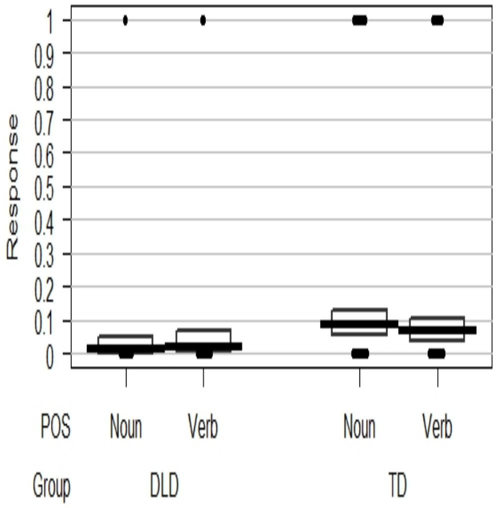


Figure 5: ***Proportion of correct production responses by Part of Speech (Noun/Verb) and Group (DLD/TD) in the Immediate comprehension session***

Figure 6*:* ***Proportion of correct production responses by Part of Speech (Noun/Verb) and Group (DLD/TD) in the Retention comprehension session***

Summary

Generally, as expected, the TD children outperformed the children with DLD in both comprehension and production; though, somewhat surprisingly, only in the Retention test session and not the Immediate test session. A (narrowly) significant Noun > Verb advantage (*p*=0.03) was seen in only one analysis: the Immediate comprehension session. Importantly, in no analysis was the Part of Speech x Group interaction significant. Thus, we have no evidence to suggest that any Noun advantage is greater for the children with DLD, in either production or comprehension, whether tested immediately, or in a later retention session.

Discussion

In this study we compared noun and verb learning in children with DLD and age-matched controls in an ecologically valid word learning task. We also investigated the impact of encoding and retention difficulties on the word learning deficit by testing comprehension and production both immediately after presentation and three to five days later. We predicted that the children with DLD would perform significantly worse than the TD children; that both groups would perform significantly better on nouns than verbs, and that the size of the noun advantage would be larger for children with than without DLD.

Our results showed that, as predicted, the children with DLD performed significantly worse than the TD children in the comprehension task and in the production task, though only in the retention session. This is in line with previous studies that have demonstrated that children with DLD have difficulty with word learning in comparison to their peers without DLD, though previous studies have tended to find differences in encoding rather than retention. The results of our study also confirmed our prediction that a noun advantage would be evident across both groups, though this was significant only in the immediate comprehension task. The study thus demonstrated that, other than lower overall performance, the children with DLD performed similarly to the TD children with respect to their comprehension of novel words. That is, despite the small sample size, both groups appear to show better comprehension of novel nouns than novel verbs, which adds to the existing body of literature suggesting a noun advantage.

Previous studies have identified deficits in word form encoding, but not retention. The opposite was found here. This may suggest that children with DLD have more difficulty with the retention than the encoding aspect of the task, but another explanation may be that in previous studies, the retention difficulties have simply been hidden by floor effects in the data. It is also possible that the immediate effect found in previous studies made it difficult to find an additional effect on retention, and hence that the absence of an immediate effect in the present study made the effect at retention easier to detect.

A further consideration is that in comparison to previous studies, the design of our study meant that there was a delay with the encoding assessment and so the study may be more accurately described as a study of short- and long-term retention in TD children and children with DLD, rather than as a study of encoding and retention. If this is the case, then it is possible that the greater differences between the TD children and the children with DLD in the delayed than the immediate recall task may actually reflect differences in encoding which only impacted retention over the longer term.

The present study provided further evidence for the noun advantage in English, but it did not show that the noun advantage is significantly greater in children with DLD than in TD children. Although the noun bias appeared to disappear in the retention session, this is likely a consequence of lower overall performance and the encoding versus retention issues outlined above; future studies should investigate whether it is still found when performance levels are higher. One way to increase performance might be to change the schedule according to which children are exposed to the novel words. For example, Childers and Tomasello (2002) found that production of novel words improved when training was spread over multiple days (see also Ambridge, Theakston, Lieven and Tomasello, 2006, for a similar finding for construction learning). A further factor that may have impacted the ability of both groups to learn novel words, may have been the syntactic complexity of the narrative. As previously discussed, the decision was made to present the words in this way, as research has shown that children can learn information about the meaning of verbs because the structure in which the verb participates provides information via a process called 'Syntactic bootstrapping'. Some studies have shown that children with language impairment may have syntactic bootstrapping difficulties and presenting the words in this way will have made the task more complex in terms of what the children had to hold in working memory and this may account for why both groups performed poorly on the task. Studies have also shown that working memory (Jackson et al., 2020) and syntactic bootstrapping_(Johnson & de Villiers,2009, Rice et al., 2000) may be particular areas of difficulty for children with DLD. While these factors may have differentially affected the performance of the children with DLD, they do not appear to have interacted with the type of word being presented as there was no evidence in this study for a greater noun advantage in the DLD children than in the TD controls.

A similar point can be made about the verbs used in the study, which varied both in syllable length in comparison to the nouns, due to the addition of inflection, and in the fact that they were presented in different tenses. Research has shown that children with language difficulties are prone to phonological and semantic impairments which may contribute to their word learning difficulties, including their difficulties with learning verbs (Black & Chiat, 2003). This may have contributed to the difficulty of the verb-learning task. It might therefore be advisable in future studies, to explicitly control for these factors in the design of the novel word stimuli. However, these factors do not appear to have differentially affected the performance of the children with DLD as they did not show a greater noun advantage than the TD controls. Verb learning in both groups may also have been impacted because nouns are always presented in the same state whereas verbs are presented with a range of endings depending on the context of the sentence. This at times may have increased the length of the word that the child had to hold in working memory and, may also have made it harder to learn the words as they were being heard in a range of different contexts.

An obvious limitation of the present study is that it suffers from lack of power due to difficulties recruiting children with DLD, which is common in this literature. A more definitive test of our predictions must therefore await future studies, which could use the effect size observed in the present study as the basis for a power calculation that would ensure a well-powered design. Our view, on the basis of the present results, is that – counter to our initial prediction – the noun advantage is probably *not* greater for children with than without DLD. Testing this prediction of a null effect would therefore require either a Bayes Factor analysis or frequentist equivalence testing, as well as a very large sample.

However, it is possible that a greater noun advantage could be detected using a different approach to that used in the present study. Recall from the Introduction that novel word learning studies have shown that children with DLD require more exposures to a word before it is learned (Alt, 2011; Alt and Plante, 2006; Gray, 2003,2004). It is therefore possible that the number of exposures needed to learn a word is a more sensitive measure than rates of correct comprehension and production per se, and hence that using this measure might reveal the kind of interaction between group and part of speech that was predicted, but not found, in the present study.

In summary, the main conclusion that can be drawn from the present study, is that TD children tend to show better overall performance in novel word learning tasks than children with DLD, and that a noun advantage can be seen in both groups, even with improved verb imageability, as compared to previous studies, though this effect was only found in the immediate recall condition. There is, however, no evidence in the present study that children with DLD show a greater noun bias in learning than age-matched TD children.

References

Alt, M. (2011). Phonological working memory impairments in children with specific language impairment: Where does the problem lie? *Journal of Communication Disorders,* 44(2), 173–185.

Alt, M., & Plante, E. (2006). Factors that influence lexical and semantic fast mapping of young children with specific language impairment. *Journal of Speech, Language and Hearing Research, 49*(5), 941-954.

Alt, M., Plante, E. & Creusere, M. (2004). Semantic features in fast-mapping. *Journal of Speech, Language and Hearing Research, 47*(2), 407-420.

Ambridge, B., Theakston, A. L., Lieven, E. V. M., & Tomasello, M. (2006). The distributed learning effect for children's acquisition of an abstract syntactic construction. *Cognitive Development, 21*(2), 174–193.

Au, T.K.F. , Dapretto, M. , Song, Y.K. (1994). Input Vs Constraints: Early Word Acquisition in Korean and English. *Journal of Memory and Language,*33(5),567-582.

Barrett, M., Harris, M., & Chasin, J. (1991). Early lexical development and maternal speech: A comparison of children's initial and subsequent uses. *Journal of Child Language,* 18, 21–40.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Bird, H., Franklin, S., & Howard, D. (2001). Age of acquisition and imageability ratings for a large set of words, including verbs and function words. *Behavior Research Methods, Instruments and Computers,* 33**,** 73–79.

Bishop D. V. M. (2017). Why is it so hard to reach agreement on terminology? The case of developmental language disorder (DLD). *International Journal of. Language and Communication Disorders,* 52, 671–680.

Bishop D. V. (2006). What Causes Specific Language Impairment in Children. *Current directions in psychological science, 15*(5), 217–221.

Bishop, D. V., & Hsu, H. J. (2015). The declarative system in children with specific language impairment: a comparison of meaningful and meaningless auditory-visual paired associate learning. *BMC psychology, 3*(3).

Bishop, D. V., North, T., & Donlan, C. (1996). Nonword repetition as a behavioural marker for inherited language impairment: evidence from a twin study. *Journal of child psychology and psychiatry, and allied disciplines, 37*(4), 391–403.

Black, M., Chiat, S. (2003). Noun–verb dissociations: a multi-faceted phenomenon. *Journal of Neurolinguistics,* 16(2-3) 231-250.

Bornstein, M., & Cote, L., (2005). Expressive vocabulary in language Learners from two ecological settings in three language communities. *Infancy,* 7(3),299-31.

Bornstein, M., Cote, L., Maital, S., Painter, K., Park, S., Pascual, L., Pecheux, M., Ruel, J., Venuti, P., & Vyt, A. (2004). Cross-Linguistic Analysis of Vocabulary in Young Children: Spanish, Dutch, French, Hebrew, Italian, Korean, and American English. *Child Development,* 75(4), 1115-1139.

Byrnes, J. P., & Gelman, S. A. (1991). Perspectives on thought and language: Traditional and contemporary views. In Gelman and Byrnes , 3-27.Castles, A., Rastle, K., & Nation, K. (2018). Ending the reading wars: Reading acquisition from novice to expert. *Psychological Science in the Public Interest, 19*(1), 5–51.

Chan, C. C. Y., Brandone, A. C., & Tardif, T. (2009). Culture, context, or behavioral control?: English- and Mandarin-speaking mothers' use of nouns and verbs in joint book reading. *Journal of Cross-Cultural Psychology, 40*(4), 584–602.

Chiat, S. (2001). Mapping theories of developmental language impairment: Premises, predictions and evidence. *Language and Cognitive Processes,* 16, 113-142.

Childers, J. B., & Tomasello, M. (2002). Two-year-olds learn novel nouns, verbs, and conventional actions from massed or distributed exposures. *Developmental Psychology, 38*(6), 967–978.

Choi, S., & Gopnik, A. (1995). Early acquisition of verbs in Korean: A cross-linguistic study. *Journal of Child Language, 22*(3), 497–529.

Dollaghan, C., & Campbell, T. (1998). Nonword repetition and child language impairment. *Journal of Speech, Language and Hearing Research,* 41, 1136–1146.
Edwards, J., & Lahey, M. (1998). Nonword repetitions of children with specific language impairment: Exploration of some explanations for their inaccuracies. *Applied*

Psycholinguistics, 19(2), 279–309.

Fisher, C. , Gleitman, H. & Gleitman, L. (1991). On the semantic content subcategorization frames. *Cognitive Psychology, 23,* 331-392.

Gentner, D. (1978). On relational meaning: The acquisition of verb meaning. *Child Development,* 49, 988-998.

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. *Language, 2,* 301-334.

Gentner, D., & Boroditsky, L. (2008). Early acquisition of nouns and verbs evidence from Navajo. In *Routes to Language: Studies in Honor of Melissa Bowerman,* 5-36.

Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition, 73*(2), 135–176.

Goldfield, B. A. (2000). Nouns before verbs in comprehension vs. production: The view from pragmatics. *Journal of Child Language, 27*(3), 501-520.

Goodman, J. C., Dale, P. S., & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. *Journal of child language, 35*(3), 515–531.

Gopnik, A., & Meltzoff, A. N. (1986). Relations between semantic and cognitive development in the one-word stage: The specificity hypothesis. *Child Development, 57*(4), 1040–1053.

Gray, S. (2003). Word-Learning by Pre-schoolers with Specific Language Impairment. *Journal of Speech, Language and Hearing Research,* 46(1), 56-67.

Gray, S. (2004). Word learning by pre-schoolers with specific language impairment: Predictors and poor learners. *Journal of Speech, Language and Hearing Research,* 47, 1117–1132.

Gray, S. (2006). The relationship between phonological memory, receptive vocabulary, and fast mapping in young children with specific language impairment. *Journal of Speech, Language and Hearing Research,* 49(5), 955-969.

Gray S, Brinkley S. (2011) Fast mapping and word learning by pre-schoolers with Specific Language Impairment in a supported learning context: Effect of encoding cues, phonotactic probability, and object familiarity. *Journal of Speech, Language and Hearing Research,* 54, 870–884.

Gray, S., Brinkley, S., & Svetina, D. (2012). Word learning by pre-schoolers with SLI: effect of phonotactic probability and object familiarity. *Journal of speech, language and hearing research, 55*(5), 1289–1300.

Gray, S., Lancaster, H., Alt, M., Hogan, T. P., Green, S., Levy, R., & Cowan, N. (2020). The Structure of Word Learning in Young School-Age Children. *Journal of speech, language and hearing research, 63*(5), 1446–1466.

Gray, S., Pittman, A., & Weinhold, J. (2014). Effect of phonotactic probability and neighbourhood density on word-learning configuration by pre-schoolers with typical development and specific language impairment. *Journal of Speech, Language and Hearing Research,* 57, 1011–1025.

Grunwell, P. (1987). *Clinical phonology,* 2nd Ed. Baltimore: Williams & Wilkins

Hills, T. T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. (2009). Longitudinal analysis of early semantic networks: Preferential attachment or preferential acquisition? *Psychological Science, 20*(6), 729–739.

Imai, M., Haryu, E., & Okada, H. (2005). Mapping Novel Nouns and Verbs onto Dynamic Action Events: Are Verb Meanings Easier to Learn Than Noun Meanings for Japanese Children? *Child Development, 76*(2), 340–355.

Jackson, E., Leitão, S., Claessen, M. and Boyes, M., (2020). Word learning and verbal working memory in children with developmental language disorder. *Autism & Developmental Language Impairments,* 6.

Johnson, V. E., & de Villiers, J. G. (2009). Syntactic frames in fast mapping verbs: Effects of age, dialect, and clinical status. *Journal of Speech, Language, and Hearing Research, 52*(3), 610–622.

Kail, R., & Leonard, L. B. (1986). Word-finding abilities in language-impaired children. *ASHA monographs,* (25), 1–39.

Kan, P. F., & Windsor, J. (2010). Word learning in children with primary language impairment: A Meta-analysis. *Journal of Speech, Language and Hearing Research,* 53, 739–756.
Kauschke, C., Lee, H.-W., & Pae, S. (2007). Similarities and variation in noun and verb acquisition: A crosslinguistic study of children learning German, Korean, and Turkish. *Language and Cognitive Processes, 22*(7), 1045–1072.

Kim, M., McGregor, K., & Thompson, C. (2000). Early lexical development in English- and Korean-speaking children: Language-general and language-specific patterns. *Journal of Child Language, 27*(2), 225-254.

Lee, J., (2010). Size matters: Early vocabulary as a predictor of language and literacy competence. *Applied Psycholinguistics,* 32(1),69-92.

Leonard L. B. (2014). Children with specific language impairment and their contribution to the study of language development. *Journal of child language, 41 Suppl 1*(0 1), 38–47.

Leonard, L. B., Deevy, P., Karpicke, J. D., Christ, S., Weber, C., Kueser, J. B., & Haebig, E. (2019). Adjective Learning in Young Typically Developing Children and Children With Developmental Language Disorder: A Retrieval-Based Approach. *Journal of speech, language and hearing research, 62*(12), 4433–4449.

Leonard, L., Schwartz, R., Morris, B., & Chapman, K. (1981) Factors influencing early lexical acquisition: lexical orientation and phonological composition. *Child Development*, 52, 882–887.

Levin, B., & Rappaport-Hovav, M. (2005). Argument realization. New York: Cambridge University Press

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, 94, 305-315.

McDonough, C., Song, L., Hirsh-Pasek, K., Golinkoff, R. & Lannon, R., (2011). An image is worth a thousand words: why nouns tend to dominate verbs in early word learning. *Developmental Science*, 14(2),181-189.

McGregor, K. K., Arbisi-Kelm, T., Eden, N., & Oleson, J. (2020). The word learning profile of adults with developmental language disorder. *Autism & developmental language impairments, 5,* 1–19.

McGregor, K. K., Berns, A. J., Owen, A. J., Michels, S. A., Duff, D., Bahnsen, A. J., & Lloyd, M. (2011). Associations between syntax and the lexicon among children with or without ASD and language impairment. *Journal of Autism and Developmental Disorders, 42*(1), 35-47.

McGregor, K. K., Gordon, K., Eden, N., Arbisi-Kelm, T., & Oleson, J. (2017). Encoding deficits impede word learning and memory in adults with developmental language disorders. *Journal of Speech, Language and Hearing Research, 60*(10), 2891-2905.

McGregor, K., Licandro, U., Arenas, R., Eden, N., Stiles, D., Bean, A., & Walker, E. (2013). Why words are hard for adults with developmental language impairments. *Journal of Speech, Language and Hearing Research*, 56, 1845–1856.

McGregor, K., Oleson, J., Bahnsen, A. and Duff, D. (2013). Children with developmental language impairment have vocabulary deficits characterized by limited breadth and depth. *International Journal of Language & Communication Disorders,* 48(3), 307-319.

Merriman, W. E., Marazita, J., & Jarvis, L. H. (1993). Four-year-olds' disambiguation of action and object word reference. *Journal of experimental child psychology, 56*(3), 412–430.

Norbury, C. F., Gooch, D., Wray, C., Baird, G., Charman, T., Simonoff, E., Vamvakas, G., & Pickles, A. (2016). The impact of nonverbal ability on prevalence and clinical presentation of language disorder: Evidence from a population study. *Journal of Child Psychology and Psychiatry*, 57, 1247–1257.

Oetting, J., Rice, M. and Swank, L., (1995). Quick Incidental Learning (QUIL) of words by school-age children with and without SLI. *Journal of Speech, Language and Hearing Research*, 38(2),434-445.

Perry, L. K., Perlman, M., & Lupyan, G. (2015). Iconicity in English and Spanish and its relation to lexical category and age of acquisition. *PLoS ONE, 10*(9), Article e0137147

Phillips, N., (2022). *YaRrr!: The Pirate's Guide to R. URL* [*https://bookdown.org/ndphillips/YaRrr/*](https://bookdown.org/ndphillips/YaRrr/)

Pinker, S. (1989). Learnability and cognition. Cambridge, MA: MIT Press.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/
Rice, M. L., Cleave, P. L., & Oetting, J. B. (2000). The use of syntactic cues in lexical acquisition by children with SLI. Specific Language Impairment. *Journal of speech, language, and hearing research : JSLHR, 43*(3), 582–594.

Rice, M. & Bode, J., 1993. GAPS in the verb lexicons of children with specific language impairment. *First Language*, 13(37 Pt 1), 113-131.

Rice, M. L., Buhr, J. C., & Nemeth, M. (1990). Fast mapping word-learning abilities of language-delayed pre-schoolers. *The Journal of speech and hearing disorders, 55*(1), 33–42.

Rice, M. L., Buhr, J. C., & Nemeth, M. (1990). Fast mapping word-learning abilities of language-delayed pre-schoolers. *Journal of Speech and Hearing Disorders, 55*(1), 33-42.

Rice, M., Oetting, J., Marquis, J., Bode, J. and Pae, S., (1994). Frequency of input effects on word comprehension of children with Specific Language Impairment. *Journal of Speech, Language and Hearing Research*, 37(1), 106-122.

Rice, M. L., & Woodsmall, L. (1988). Lessons from television: Children's word-learning when viewing. *Child Development, 59*(2), 420–429.

Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., & Roy, D. (2015). Predicting the birth of a spoken word. *PNAS Proceedings of the National Academy of Sciences of the United States of America, 112*(41), 12663–12668

Smiley, P., & Huttenlocher, J. (1995). Conceptual development and the child's early words for events, objects, and persons. In M. Tomasello & W. E. Merriman (Eds.), *Beyond names for things: Young children's acquisition of verbs,* (pp. 21–61). Lawrence Erlbaum Associates, Inc.

Snedeker, J., & Gleitman, L. R. (2004). Why It Is Hard to Label Our Concepts. In D. G. Hall & S. R. Waxman (Eds.), *Weaving a lexicon,*257–293.

Swingley, D. & Humphrey, C. (2018) Quantitative Linguistic Predictors of Infants' Learning of Specific English Words. *Child Development*, 89(4), 1247–1267.

Tardif, T., Gelman, S. A., & Xu, F. (1999). Putting the "noun bias" in context: A comparison of English and Mandarin. *Child Development, 70*(3), 620-635.

Tardif, T., Wellman, H. M., Fung, K. Y., Liu, D., & Fang, F. (2005). Pre-schoolers' understanding of knowing-that and knowing-how in the United States and Hong Kong. *Developmental Psychology, 41*(3), 562-573.

Tomblin, J. B., Records, N. L., Buckwalter, P., Zhang, X., Smith, E., & O'Brien, M. (1997). Prevalence of specific language impairment in kindergarten children. *Journal of speech, language and hearing research, 40*(6), 1245–1260.

Van der Lely, H. K. J. (1994). Canonical linking rules: Forward versus reverse linking in normally developing and specifically language-impaired children. *Cognition, 51*(1), 29–72.

Windfuhr, K.L., Faragher, B. and Conti-Ramsden, G. (2002). Lexical learning skills in young children with specific language impairment (SLI). *International journal of language & communication disorders,* 37(4), 415–432.

Wright, L., Pring, T., & Ebbels, S. (2018). Effectiveness of vocabulary intervention for older children with (developmental) language disorder. *International journal of language & communication disorders, 53*(3), 480–494.

Data, code and materials availability statement

The data that support the findings of this study and a pre-print of the manuscript are available from OSF. https://osf.io/3eqtk/files/osfstorage

Ethics statement

This study was approved by the University of Liverpool Ethics Committee. Informed written consent was obtained from the schools and caregivers, and the children also gave verbal assent.

Authorship and Contributorship Statement

Paula Stinson and Julian Pine conceived and designed the study and Paula Stinson collected the data and wrote the first draft of the manuscript. Both authors analysed the data, approved the final version of the manuscript and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Acknowledgements

Julian Pine is a Professor in the International Centre for Language and Communicative Development (LuCiD) at The University of Liverpool. The support of the Economic and Social Research Council [ES/L008955/1] is gratefully acknowledged. Thanks are also due to Ben Ambridge for his help with the statistical analysis.

Appendix A

Appendix 1- 'Tin Toy' Script

Object/ Action	Non- word
Noun 1- Beads	Poffee
Noun 2- Tin toy's hat	Nall
Noun 3- Drum	Mot
Verb 1- Baby waving their arms and legs	Tuddle
Verb 2- Tin toy walking and playing music	Dite
Verb 3- Spinning in circles	Bickle

Tin Toy

Visual	Narration
1. Camera spans over box and shows the room with toys.	1. Let's meet our new friend Tin Toy. He has just come out of his box! He's not the only toy there. There are some stacking rings and some Poffee (coloured beads).
2. Close up of Tin toy as he looks around the room.	2. Tin toy is wearing his special hat today. It's called a Nall (Tin toy hat). His Nall is his favourite thing to wear. He's having a look around the room. He can see the stacking rings and Poffee too!
3. Tin toy watches as baby enters the room. Baby sits by the toys and begins to move.	3. There's someone coming. It's a baby! Tin toy is happy to see the baby. Oh, look the baby is tuddling (waving legs and arms). Did you see him Tuddle? Tin toy was so surprised at that, it's made his Nall shake. The baby is fling again with the blue ring.
4. Baby lifts the beads from the floor.	4. Look what the baby is getting now. He's lifted the Poffee. Oh no, the baby Tuddled and the Poffee have broke.
5. Tin toy starts to move away from the baby.	5. Tin toy's not sure what to do. He stepped back and his Mot (drum) just banged. Look what Tin toy's doing, he's Diting (moving while paying music).
6. Tin toy moves around the room as the baby follows him	6. The baby is following him and watching him Dite. I hope his Nall stays on his

	head!
7. Tin toy spins around and continues to run away from the baby.	7. He's in such a hurry he's started to Bickle (spin in circles). Did you see how fast he Bickled? He's Diting as fast as he can now away from baby. Look at his Mot banging.
8. Tin toy heads towards the box and gets stuck.	8. Where will he Dite to? He's stuck in the box now. How will he get out? He's Bickled and got out of the box! The baby was surprised to see him Bickle.
9. Tin toy goes under the sofa and sees the other toys	9. He's safe under the sofa now. He's nice and quiet now that his Mot isn't banging. Look, there are other toys under the sofa with Tin toy.
10. The baby falls over and begins to cry.	10. Oh no. Now the baby's fallen over and he's crying. Tin toy goes to see if he's okay.
11. Tin toy goes out to the baby and the baby shakes him and throws him.	11. oh the baby's thrown Tin toy! I hope he hasn't broken his Mot.
12. The baby shakes inside the bag and Tin toy follows him as he leaves the room.	12. The baby's looking in Tin toy's box and what is he doing now? The baby is stuck! Oh dear, I hope he can get the bag off.

Appendix 2- 'Lifted' Script

Object/ Action	Non- word
Noun 1- Small alien	Dut
Noun 2- Big Alien	Tellon
Noun 3- Driving handlebars	Bettle
Verb 1- Man levitating	Nuve
Verb 2- Big alien moving fingers	Pover
Verb 3- Ship warping	Mipe

<u>Lifted</u>

Visual	Narration
1. Camera spans over a house at night. A man is sleeping inside, and a light is seen from the window	1. It's the middle of the night and everything is very quiet. There's a man sleeping soundly in his bed. Look there's a bright light and it's coming from a spaceship!
2. The man floats above his bed and moves around the room.	2. Watch, the man is starting to Nuve (levitate). He's nuved all the way out of his bed! Oh dear, he's nuved right into the wall ... again!
3. Inside the spaceship a small alien is working the control panel as a larger alien watches. The small alien chooses switches as the man is moved around inside his bedroom.	3. Look here's a Dut (small alien). It was him that was making the man Nuve. There's a Tellon (big alien) watching him. He doesn't look very happy. He's trying to find the right switch. Oh no! The man is moving all around the room.
4. The small alien tries other switches as the big alien watches.	4. The Dut is trying a different switch. That one wasn't right!
5. The small alien is frustrated as he tries many different buttons	5. Oh no. I don't think the Dut is doing very well. There are a lot of switches and he doesn't know which one to choose.

6. The big alien watches as the small alien consistently chooses the wrong switch. The man is moved around the room.	6. I don't think the Tellon thinks he's picking the right one, do you? Oh dear, that wasn't the right one, either. This must be very tricky. The man is going everywhere!
7. The man is transported outside his house and into a tree.	7. Maybe the book will help him find the right one. The man is stuck in the window now, oh no! He's gone straight into a tree.
8. The man travels from the tree up into the spaceship.	8 The man is going all the way up to the ship. Can you see?
9. The man falls from the spaceship towards the ground.	9. Uhoh! He's fallen back down again. The Tellon won't like that. Phew, they caught him before he hit the ground. Well done, Tellon!
10. Big alien takes the controls and returns the man to bed while restoring his room.	10. What's he going to do now? Maybe he'll pover. Look he's povering! He can pover very fast. Because he povered , now everything is going back the way it was.
11. The big alien begins to drive and when the smaller alien looks sad, he offers him the steering wheel.	11. He's using the Bettle (Handlebars) now to drive them home. Oh dear, the Dut is very sad. But look, he's giving him the Bettle so he can drive. He's so happy he can use the Bettle. He'll use the Bettle to drive them all the way home.
12. The spaceship is about to fly but falls back to earth and then flies into space. The man's house is destroyed but he is still sleeping in his bed.	12. The spaceship is getting ready to Mipe (warping) so they can go home. Uhoh, looks like they've fallen back to earth. Look the spaceship is Miping. They've Miped back into space They didn't even wake the man when they Miped!

License

Examining the incremental process of word learning: Word-form exposure and retention of new word-referent mappings

Sarah C. Kucker
Southern Methodist University, Dallas, TX, US

Bob McMurray
University of Iowa, Iowa City, IA, US
Delta Center, University of Iowa, US

Larissa K. Samuelson
University of East Anglia, Norwich, Norfolk, UK
Delta Center, University of Iowa, US

Abstract: This study examines the process of learning new word-object mappings and how repeated exposure to word-forms impacts retention. Infants 18- and 24-months-of-age were first exposed to new word-object mappings in a referent selection task. To examine the influence of extra word-form repetitions on retention, newly mapped word-forms were repeated in a preferential listening task prior to a delayed retention test. Retention was tested in an object selection task. Consistent with prior work, infants performed very well on novel referent selection yet demonstrated a novelty bias on known referent selection trials that was especially prominent in the younger age group. There were no differences in listening times across age groups during the preferential listening task. However, there was some evidence that longer listening time predicted retention. As a group, 24-month-olds showed above chance retention of word-object mappings created during referent selection – an ability rarely seen at this age. This suggests additional exposure to word-forms after mapping may increase learning, at least in 24-month-old children. These findings both replicate prior work on children's referent selection abilities and highlight the incremental and cascading nature of the processes that strengthen new word-object mappings over repetition and development.

Keywords: referent selection; word learning; auditory familiarization.

Corresponding author(s): Sarah C. Kucker, Department of Psychology, Southern Methodist University, Dallas, Texas, United States 75278. Email: skucker@smu.edu

ORCID ID(s): Sarah Kucker: https://orcid.org/0000-0003-2210-3599; Bob McMurray: http://orcid.org/0000-0002-6532-284X; Larrisa Samuelson: https://orcid.org/0000-0002-9141-3286

Citation: Kucker, S.C., McMurray, B., & Samuelson, L.K. (2024). Examining the incremental process of word learning: Word-form exposure and retention of new word-referent mappings. *Language Development Research, 4* (1),326—360. http://doi.org/10.34842/3ppz-g170

Introduction

Children make word learning appear quick, forming new word-referent mappings accurately and rapidly developing large vocabularies before reaching three years-old. However, learning even a single word is a time-extended process in which the mappings between word-forms and referents are strengthened over development (Carey, 2010). This process begins when a child first hears a word—in that first moment of exposure, the child must identify and encode the word-form from the speech stream and select the target from among multiple possible referents. This initial link, however, has been shown to be fragile: children often attend to the right object in-the-moment, but either fail to show evidence of retention a few minutes later (Bion et al., 2013; Horst & Samuelson, 2008) or show only limited retention in specific contexts (Axelsson & Horst, 2014; Kucker & Samuelson, 2012). In order to retain the new word-referent link, the initial mapping must be strengthened (Carey, 2010; Kucker et al., 2015). This can occur in a number of ways, including via repetition of the word-form, object, or word-object pair over exposures (McMurray et al., 2012; Mollica & Piantadosi, 2017).

For example, an associative computational model proposed by McMurray et al., 2012 (see also Kucker et al., 2018; Zhao et al., 2019) suggests that word-forms and objects are independently associated with conceptual or lexical representations (broader category of things to which the word-form might refer). In this account each subsequent exposure to a word-form, referent, or the pair, modifies the weights and connections between word-forms, referents, and conceptual representations in the lexical network, strengthening some connections and down-weighing or even pruning others. This means that relevant learning can occur both when a component of the mapping is not present, by reinforcing connections between that item (word-form or object) and the intermediate lexical layer (the nodes that connect word-forms and objects in the computational model), and by pruning spurious connections. Thus, mere exposure to a word-form or referent can reinforce individual connections, altering the network. Indeed, other work by Gathercole and colleagues (2006, 1997) suggests that the short-term memory of phonological forms (which are heightened with additional exposures) may play a particularly important role in word learning.

The current study probes this process, examining how a robust word-object mapping—one that can support longer-term retention—develops from real-time encoding, through repetitions, to retention. Notably, while our approach is based on an associative framework, the idea that repeated exposures are needed to solidify mapping is true of many other theoretical approaches (Carey, 2010; Hollich et al., 2000; Trueswell et al., 2013; Xu & Tenenbaum, 2007). For example, the propose-by-verify framework suggests that additional exposures either help infants verify or revise an initially proposed mapping (Trueswell et al., 2013). Thus, the current study aims to replicate and extend prior work on referent selection and retention by exploring how repeated exposure to auditory word-forms influences eventual retention at two ages (18- and

24-months). In particular, we hypothesize that repeated exposure to word-forms, particularly following an initial linking of a word and referent, may boost retention in the same way that prior work has shown to be true of object exposure (Kucker & Samuelson, 2012) or object-word pair repetition (Axelsson et al., 2012).

Developmental changes in word learning

Recent studies on word learning demonstrate reliable developmental changes in both in-the-moment referent selection and later retention abilities. For instance, as young as 14-months of age, children can identify a single referent of a single new word and retain it (Mervis & Bertrand, 1994). However, learning in more complex contexts with multiple competing referents shows an extended developmental trajectory that begins to emerge closer to 17-months (Lewis et al., 2020). In these paradigms (typically referred to as fast-mapping, mutual exclusivity, disambiguation, or referent selection), children are confronted with an array of objects, one of which is novel and the others familiar. They are then prompted with a novel word-form (e.g., get the cheem). In response, 17-month-old children look away from a known item and toward a novel object (Halberda, 2003). By 18 months, children can select the novel referent when prompted with a novel label (Bion et al., 2013; Horst & Samuelson, 2008; Kucker et al., 2018).

Thus, even young children can readily map novel word-forms to novel objects in-the-moment. But even at times when a correct initial mapping is not established, exposure to the word-referent pair can nonetheless create changes in the system that are the first step of learning. Indeed, work with both children and adult learners as well as computational models suggests that even when word learners choose the wrong item at first exposure, useful learning still occurs (Fitneva & Christiansen, 2011, 2017; Yurovsky et al., 2014). In one study, words that were not correctly mapped during initial exposure were subsequently mapped more quickly than brand new words upon a second exposure. This suggests that initial traces of learning were laid down at first exposure even if the behavioral responses did not show it (Yurovsky et al., 2014). Moreover, word learners encode information beyond the target during exposure, such as details of foil items and context (Wojcik & Saffran, 2013; Zettersten et al., 2018). Thus, even absent a correct referent selection, information about objects and labels that are simply present in the context is encoded by children in ways that can impact their future learning.

However, it is also clear that initial exposures are often not enough to support retention after a delay. For word-referent pairs encountered with competitors and named only once, retention is not robust until at least 30 months (Bion et al., 2013; Horst & Samuelson, 2008). Even past 30-months, mapping and retention continue to strengthen, supported by some of the same associative properties that direct attention toward the targets and away from distractors, strengthening and refining each associative path (Pomper & Saffran, 2019). Thus, children's ability to retain novel word-object mappings shows a rather protracted developmental course even in simple laboratory tasks. This process is likely to be especially pronounced from 18- to 24-months when vocabulary is exponentially increasing. Learning and retaining a new mapping requires that children form a robust representation of the object that is clearly distinguishable from other potential referents. They must form a similarly robust representation of the word form, and link both together. This system does not just rely on the exposure to the mapping (e.g., the word-form and object together) to achieve this robustness; rather research demonstrates that retention of new word-object mappings is also improved in the context of experiences that build stronger representation of the referents or word-forms alone (Vlach & Sandhofer, 2012).

The role of referent and/or label familiarity in retention

Associative learning is advanced through repeated exposure to word-forms and referents, either individually or together. A caregiver repeating word-forms in the presence of their referents can lead to learning that is usefully built upon in subsequent exposures, but exploration of unnamed objects and hearing words without referents can also build learning (see, e.g. Clerkin & Smith, 2022). Importantly, the relationship between objects and labels in this process does not need to be symmetrical (i.e. one does not necessarily learn objects and labels in equal ways or to equal degrees); recent work has shown that extra exposure to referents versus labels may impact word learning in different ways. Kucker and Samuelson (2012) demonstrated that even a short 1-2-minute familiarization with the referents of novel words prior to mapping boosted 24-month-old children's retention to levels higher than seen without familiarization. This suggests that familiarity with the referent supports stronger representations of the objects and has downstream effects on retention. This idea also fits with work by Clerkin et al. (2017) showing that the number of times infants have seen a referent, rather than heard its name, predicts which word-forms will be said first. Moreover, slightly older children (30-month-olds) who see multiple, variable, examples of the target object during referent selection retain the label for this new category (Twomey et al., 2014) as do similarly aged children given iconic or shape-based gestures alongside exposure to the label (Aussems & Kita, 2019; Capone & McGregor, 2005).

However, pre-familiarization with potential referents does not always improve performance. Kucker et al. (2018) showed that 18-month-olds given familiarity with novel referents prior to a single mapping trial did not show improved retention. This contrasts with the 24-month-olds of Kucker and Samuelson (2012) who benefited from familiarity, even with only a single mapping instance. One explanation for this is that for the younger 18-month-old children, referent selection is strongly driven by a novelty bias so robust that even a few minutes of familiarization does not diminish it (Kucker et al., 2018, 2020). Supporting this, these younger children fail to select *known* target items during referent selection when a novel foil item is present, even though they correctly select the same known items when no novel items are present. Together, these results suggest a shift in the impact of novelty and familiarity on referent selection and retention as children's vocabulary grows from 18- to 24-months. Moreover, object familiarity may have downstream effects on retention at older ages.

Other work suggests that exposure to auditory word-forms impacts subsequent learning. For instance, word-forms presented in isolation may prime word-referent mappings for toddlers (Willits et al., 2013), exposure to new phonological patterns influences the mapping of those sounds to objects (Breen et al., 2019) and in some cases, children's attention to auditory information may overshadow attention to visual stimuli (Robinson & Sloutsky, 2004). Moreover, word-form repetitions are frequent in the life of a child. Caregivers often use successive word repetitions in interactions (Schwab & Lew-Williams, 2016) and discuss absent objects (Gallerani et al., 2009), increasing a child's exposure to a word-form but without the referent. This is critical because auditory word forms are substantially different from objects. Whereas objects endure in time, word-forms are fleeting and need to be repeated to increase exposure.

This raises two central questions about toddlers' encoding of word-forms during referent selection tasks. First, how well do toddlers encode (and retain) the auditory word-form (independent of its referent) in the context of referent selection tasks when word-forms are heard once? Evidence suggests even young infants have this ability: 8-month-old infants retain repeated auditory word-forms for up to two weeks (Jusczyk & Hohne, 1997), and ten-month-old infants prefer pre-familiarized auditory stimuli (Robinson & Sloutsky, 2010), and recognize familiar words in speech (Jusczyk & Aslin, 1995). Further, work on toddlers' ability to discriminate mispronunciations of common words (Jusczyk & Aslin, 1995) and repeat novel forms presented in novel word representation tasks (Graf Estes et al., 2007; Hodges et al., 2016, see also Gordon et al., 2016 for evidence with 3-5 year-olds) suggests they quickly create auditory representations necessary for new word-object mappings.

However, less work has measured the robustness of these word-form representations in the context of word-object mapping, especially in children under 3-years. In many referent selection or fast-mapping studies assess learning by presenting the child with the word-form and asking them to select the referent (and not vice versa). However, children are rarely tested on their ability to retain the word-form independent of the mapping. One study with older children found that retention of new words by 3-year-olds was best supported by strong initial encoding, but that when newly learned mappings were lost, it was the word-forms that were most susceptible to decay (Munro et al., 2012). It is unknown if 18- to 24-month-old children's failure to retain new word-object mappings derives from the failure to form an auditory word-form representation if a word-form representation is formed but not linked to the referent, or if word-forms decay too quickly to support retention. What is needed is an independent test of word-form learning.

The second question is whether familiarity with word-forms promotes retention in the same way as object familiarity does. Literature on children's long-term episodic memory suggests it might—18-month-olds given verbal cues (narration of an event) prior to retrieval (but not during encoding) showed higher retention (Hayne & Herbert, 2004), and in the literature on memory in children, verbal cueing with task-relevant words can increase recall of prior facts and events (Bauer et al., 2007; Mateo et al., 2018). In each of the prior cases, children exposed to relevant verbal input performed better on subsequent memory tests. This suggests that retention of new word-referent pairs could theoretically be supported by word-form repetitions. However, few studies have tested this hypothesis on 18-24-month-old children who are in the midst of the vocabulary spurt. Studies that have, were limited to exposing children to word-forms *prior to* mapping, not after initial exposure.

In one example, Graf Estes, Evans, Alibali, et al., (2007) gave children an auditory statistical segmentation task (with no visual referents, but multiple repetitions of word-forms) prior to a referent mapping task. They found 17-month-old children were able to rapidly map novel objects to word-forms defined by high transition probabilities in the segmentation task. That is, by 17-months, exposure to word-forms may help with word-referent mapping, but its impact on retention is less clear. By 24-months, Kucker and Samuelson (2012) found that additional exposure to word-forms was not needed to aid in mapping—all children in this study reliably mapped new word-forms to referents regardless of exposure. However, word-form exposure did not improve retention at 24-months. Taken together, this work demonstrates there are developmental changes from 18-24-months in how word-form repetitions impact mapping, but the impact on retention is still unknown, especially for the younger children. Moreover, neither study tested the impact of word-form exposure after initial word-referent traces were laid down; a critical question given that initial mappings/exposures present the first step toward word learning. Thus, unknown is whether auditory familiarity after initial exposure may support retention in referent selection contexts, children's learning for word-forms, and at what ages such repetition may be beneficial for learning.

Current Study

Overall, the theoretical accounts and the literature suggest that repeated exposure may be helpful to support the retention of newly formed word-object mappings. However, while prior exposure to objects can boost retention of new mappings, the data is less clear with respect to the auditory component of new mappings. We know that infants can form initial auditory representations of word-forms heard in isolation. However, we do not know how well word-forms are retained from referent selection tasks or how repetition of word-forms impacts the process of word learning.

To examine these issues, we conducted a referent select and retention task, but inserted a probe of word-form encoding between referent selection and retention. A preferential listening task offered passive exposure to, and repetition of, the word-forms presented during referent selection. This task also provided a measure of children's recognition of word-forms from referent selection at this age (Willits et al., 2013). Multiple prior studies have used the referent selection and retention paradigm but without a preferential listening phase (Horst & Samuelson, 2008; Kucker et al., 2018); the study here paralleled those studies as closely as possible with the exact same stimuli, participant pool, and procedures used. This allowed comparison to similar groups of children who performed the same task without extra exposure to the word forms.

Our goals were three-fold. First, we used the preferential listening task to ask if children retain novel word-forms after initial exposure during the referent selection task. Second, we examined whether repetition of word forms after initial referent selection impacted later retention of word-referent pairs. Third, we examined how these effects change over vocabulary development by examining performance in two different age groups, 18- and 24-month-olds, selected because they had been the focus of the prior studies using the same paradigm and because the reviewed literature suggests ongoing changes in the processes supporting word learning in this period.

Methods

Participants

Two groups of children participated: 18-month-olds (N=33) and 24-month-olds (N=26), see Table 1. The sample size was based on prior work (e.g. Kucker & Samuelson, 2012) in which medium to large effects were found. Moreover, G*Power a priori power estimates for the ORs found in retention trials for Kucker & Samuelson (2012) suggest between 19-37 children would be needed to detect a large effect with a power of .95 in the current study. All children were monolingual English speakers and recruited from a Midwestern college town in the US. Ethnic/racial data and detailed SES information were not available for all children, but the majority for whom information was available identified as non-Hispanic, White and middle-upper class with at least one parent holding a college degree. Data for 3 additional children were dropped due to fussiness (2) and programming error (1). Informed consent was obtained prior to beginning the study and children received a small prize for participating.

The current sample was aimed to be representative of the population in terms of language ability and include children from the full spectrum of the curve. Although having a low expressive vocabulary (i.e. being a late talker) may be a risk factor for later developmental delays and DLD, simply being low on expressive vocabulary is far from a perfect predictor of later delays and many late-talking children catch-up to their peers and demonstrate normal vocabulary skills (Rescorla, 2011). Recent work has also suggested that there are no hard cut-off points for identifying a child as at-risk based solely on vocabulary size as vocabulary is a continuum (Dollghan, 2013; Kucker & Seidler, 2022). Thus, children with lower expressive vocabularies were not dropped from the current sample in order to provide a comprehensive assessment of this age, but children who were identified with significant developmental delays (e.g. autism, Down's syndrome) were excluded from participating in the first place.

Table 1. ***Demographics of sample***

Groups	N	Sex	Age	Expressive Vocabulary
18-month-olds	33	13 female	18; 26 (17;21-19;29)	81 (4-356)
24-month-olds	26	13 female	24; 18 (23;27-25;9)	300 (6-667)

Note, ranges shown in parentheses. Vocabulary according to total words on the MCDI-WS.

Stimuli

Two sets of objects were used during referent selection and retention: well-known familiar items and unfamiliar novel items (Figure 1). Labels for known items were known, on average, by 66% of 18-month-olds and 85% of 24-month-olds (LEX database; Dale & Fenson, 1996); novel items were unknown. All items were identical to those used in prior work (Horst & Samuelson, 2008; Kucker et al., 2018). Parents confirmed items were respectively known and novel and items were replaced as necessary. In addition, eight novel word-forms (from Horst & Hout, 2016) were used that conformed to the phonological rules of English but had no known referents (see Table 2). Four variations of each word-form (each clip 2 seconds long) were recorded in the experimenter's voice for preferential listening. Half of these word-forms were used in the Referent Selection (RS) trials and heard again in Preferential Listening and on the Retention trials; the other half were kept as novel and only heard in the Preferential Listening section. Order of trials were counter-balanced.

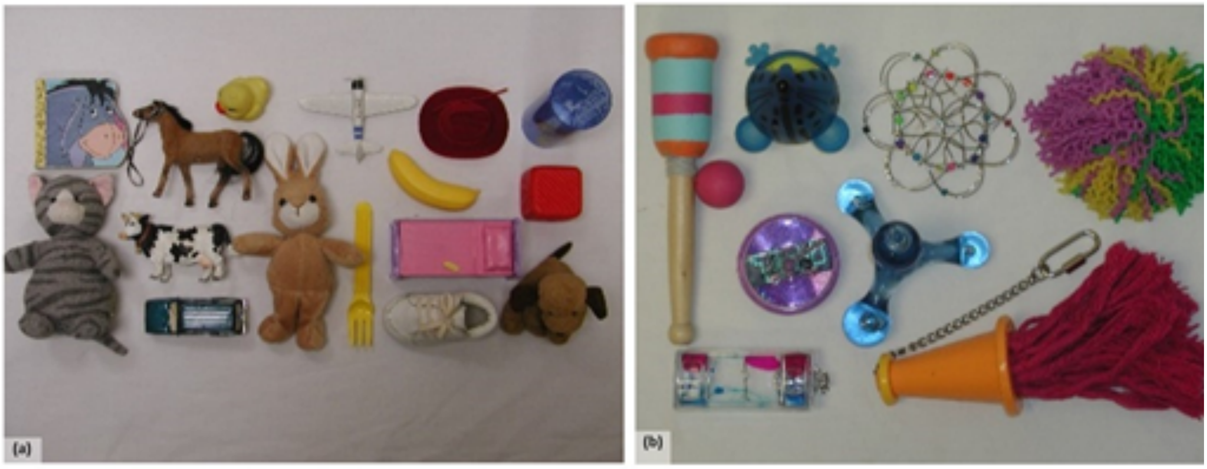


Figure 1. ***Known (a) and Novel (b) stimuli.***

Table 2. ***Novel Word-form stimuli***

Word forms	IPA	klattese	Phonotactic Probability	Neighborhood Density
Known				
Airplane	ˈɛɚˌpleɪ̯n	Erplen	.289	0
Banana	bəˈnɑːnə	b|n@nx	.311	0
Bed	bɛ́d	bEd	.162	22
Block	blɔ́k	blak	.158	4
Book	bʊk	bUk	.115	13
Bunny	bʌni	b^ni	.230	11
Cat	kæt	k@t	.238	27
Car	kɑɹ	kar	.232	17
Cow	kaʊ	kW	.102	9
Cup	kʌp	k^p	.169	12
Dog	dɔg	dcg	.086	7
Duck	dʌk	d^k	.145	7
Fork	foɹk	fork	.217	7
Hat	hæt	h@t	.185	25
Horse	hɔɹs	hcrs	.184	1
Novel				
Dite	dait	dYt	.152	19
Cheem	tʃîm	Cim	.090	8
Fode	foʊd	fod	.134	13
Lorp	lɔɹp	lorp	.198	1
Pabe	peɪb	peb	.140	6
Roke	ɹoʊk	rok	.153	19
Stad	stæd	st@d	.198	5
Yok	jɔ́k	yak	.122	8

Note, Phonotactic probability calculated from Vitevitch & Luce (2004). Neighborhood Density from child corpus from http://www.people.ku.edu/~mvitevit/PhonoProb-Home.html. Novel word-forms included both RS and NN words; which were RS and which were NN were randomized across children.

Procedure

The procedure for the warm-up, referent selection and retention phases were identical to that of prior work (Horst & Samuelson, 2008; Kucker et al., 2018). As in prior work, the child sat across a table from the experimenter in a booster chair or on their parent's lap. Parents were instructed to avoid interacting with their child, offering minimal, neutral encouragement only if needed. They completed the MacArthur-Bates Communicative Development Inventory: Words and Sentences (MCDI; Fenson et al., 1994) during the session, which was used to calculate total vocabulary size for each child. The procedure began with warm-up, then proceeded to the three phases of the test trials – referent selection, preferential listening, and retention (Figure 2). See online materials for full datasheets representing all objects, possible orders, and trials.

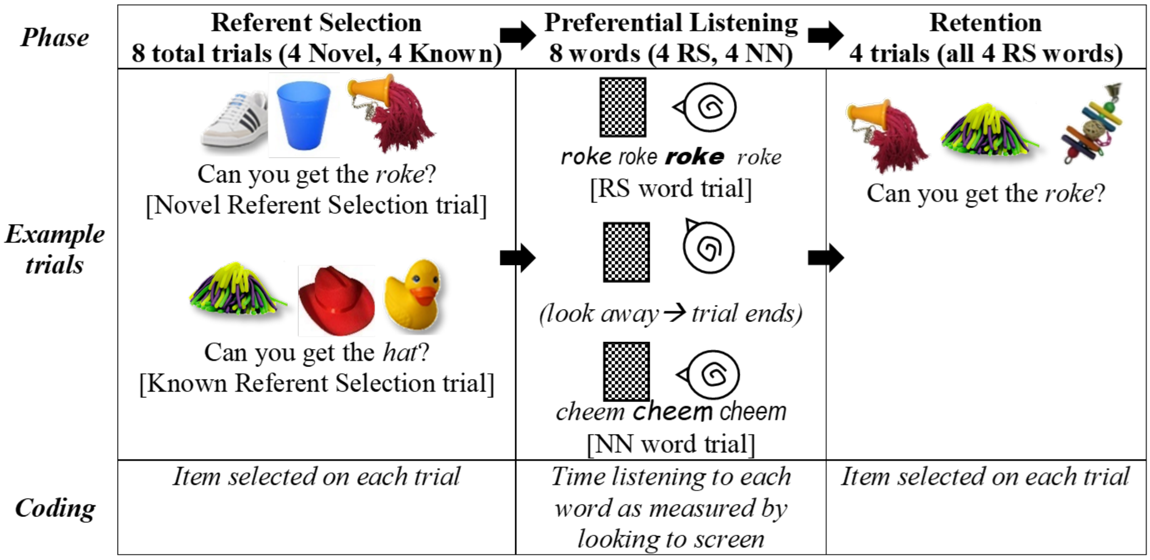


Figure 2. ***Schematic of the procedure***
Note: RS word-forms are those used on the novel referent selection trials, whereas NN word-forms are new novel word-forms presented only during preferential listening.

Warm-up

The warm-up period familiarized the child with the testing procedure. On each trial, three items were placed equidistant apart on a white tray. While maintaining eye contact with the child, the tray was placed on the table within view, but out of reach of the child, for three seconds. The experimenter then requested an item by name ("Can you get the hat?") and pushed the tray forward. Children were corrected or praised as needed (e.g. if the child chose the correct item, the experiment clapped and said, "Good job", whereas if the child chose the wrong item, the experimenter re-prompted once, then pointed to the correct answer). Target locations and prompts were randomized across trials and children, and each item was the target (with other familiar items as the foils) once for a total of three trials. Referent selection immediately followed.

Referent Selection

Referent selection consisted of eight trials with a similar procedure to warm-up, but without praise or correction. On each trial, two known items from warm-up and one never-before-seen novel item were present. On half the trials, children were asked to select a known item by name ("Can you get the hat?"); these are referred to as the Known RS trials. The other half of the trials (Novel RS) alternated and asked the child to select a novel item by name ("Can you get the roke?"). Children were prompted only once on each trial, consistent with prior work (Horst & Samuelson, 2008). Target items and locations were randomized across trials and children, and target items did not repeat.

Preferential Listening

Preferential listening took place in a curtained-off portion of an adjacent room immediately following referent selection. Children were seated on their parent's lap approximately 24" in front of a 42" flat screen monitor with speakers positioned on either side of the monitor. An infrared camera was positioned directly below the monitor and centered on the child. Auditory stimuli and a checkerboard pattern on the monitor were controlled via HABIT (Cohen et al., 2004). HABIT presents a simple, traditional habituation paradigm based on button presses by the experimenter to indicate the child is attending. The original HABIT software is now obsolete. However, an updated HABIT program is available from Oakes and colleagues (2019) that allows additional flexibility in software and stimuli. Parents wore headphones during the task to minimize interference.

The task began with five training trials using various sounds (e.g., bell chime, whistle). Using a head-turn procedure, a single sound was repeated as long as the child maintained attention at the screen displaying a black and white checkerboard pattern. Once the child turned away for two consecutive seconds, the trial ended, and a new sound was played following the same procedure.

Eight test trials immediately followed in the same manner using novel word-forms instead of sounds. To examine children's memory for word-forms presented during referent selection, we measured listening to both the novel word-forms heard during referent selection and to completely novel word-forms that were not previously presented during the study, with the expectation that a preference for one over the other would indicate learning. Head turns were recorded online by button presses from the experimenter and registered by the HABIT program.

This task tested four novel word-forms from referent selection (RS words) and four previously unheard new novel words (NN words). The specific words used for RS and NN were randomized across children. For each word-form, there were four different audio recordings produced by the same experimenter who conducted the referent selection phase. Clips were played in a random order. Order of test trials was randomized.

This procedure uses preferential listening as a test of recognition of the word-forms that appeared in the referent selection phase. Importantly, it also parallels the word-form familiarization procedure of Kucker and Samuelson (2012) in which the child, through their attention to the screen, chose which word-forms to hear again and how many times they would hear each word-form.

Retention

Retention immediately followed preferential listening. It was conducted in the same room as the warm-up and referent selection trials, using a similar task. Retention started with a single warm-up trial in the same manner as before to re-engage the child. To prevent repetition of referents, two retention trials followed (e.g. Horst & Samuelson, 2008). Each retention trial consisted of three previously seen novel items from referent selection: two items that had previously been named on a novel referent selection trial and one novel foil from known referent selection. Children were presented with all items on a tray as before and asked to select a single, previously-named item from a Novel RS trial by name ("Can you get the roke?"). No item repeated across the retention trials and the location and order of the target was randomized.

Coding

Children's final selections on referent selection and retention were coded by an experimenter blind to the hypothesis. See "choice" coding in online manual. Trials in which no item was a clear choice were marked as a no-response and not included in analyses. 40% of trials were recoded for reliability; agreement between coders was 100%. Experimenters achieved a 90% accuracy (via pre-recorded videos) on noting head turns in preferential listening prior to data collection.

Analysis

All referent selection and retention trials in which the child made a distinguishable choice were included in the analysis (>90% of trials). Preferential listening trials with less than 2000ms of listening (i.e. less than 2 repetitions of a word-form) were removed prior to analysis, as is standard in such tasks (Jusczyk & Aslin, 1995). A total of 27 (of 464) NN and 32 (of 464) RS preferential listening trials were dropped; all children had data from at least 2 (of 4) NN word trials and all but one child had at least 2 (of 4) RS word trials. The one exception was a child missing 3 RS word trials whose data was retained for the remaining word. One 18-month-old child missing retention data and one 24-month-old child without a completed MCDI were dropped from those respective analyses. Thus, all 59 children contributed data.

Generalized mixed models testing trial-by-trial performance and linear mixed models testing listening time were run separately for each experimental phase. Fixed factors included age-group (18- vs. 24-months-old, contrast coded respectively as -.5, +.5, then centered) as well as specific predictors relevant to the questions of each phase. In all cases, vocabulary was highly collinear with age (VIF's ≥5), so secondary exploratory analyses with models split by age group were run to examine the impact of vocabulary (centered) on performance (see Bion et al., 2013). Continuous predictors (vocabulary, looking time) were centered prior to inclusion as a predictor.

To assess performance in the RS and retention trials against chance, individual models for each age group were run with a random intercept of subject, and either a fixed effect of trial type (contrast coded; novel vs. known words for referent section), or no fixed effects (in the case of retention). Fixed effects were dummy coded with the trial type of interest set as 0 and the other as 1. The significance of the intercept was used to assess if accuracy within a condition was greater than chance (33%); because the default intercept assumes .5 as chance, an adjusted intercept was calculated by subtracting ln(1/2) (this value was used because chance was set to 33%: ln(.333/(1-.33)) = ln(1/2)) and dividing by the standard error to get a new Z score. A chance value of 33% was used here because children were given three items on each trial and prior work has suggested that children will consider all present items (Halberda, 2003); we had no reason to believe that children here would behave differently (indeed, as seen in the results, all objects present were chosen at least some portion of the time). Moreover, prior work with 3AFC paradigms, including those used as a comparison here, use 33% as chance (Gordon & McGregor, 2014; Horst & Samuelson, 2008; Warren & Duff, 2014).

Models were fit using R version 4.0.3 with the lme4 and lmerTest packages. We used the Laplace approximation for the glmer, and the Satterwaithe approximation to compute the degrees of freedom in lmerTest for linear models. The maximum random effect structure justified by the data was used according to AIC comparison (Seerdorff et al., 2019), which could include random intercepts and/or slopes of subject and item; in all cases a random intercept of subject was the best fit.

Results

Referent Selection

The goal of the referent selection trials was to test children's ability to select both a novel and a known item by name from an array. Mirroring prior work (Horst & Samuelson, 2008; Kucker et al., 2018), both age groups were well above chance at novel referent selection (Table 2, Model C; Figure 3): 18-months (96.05%, $p<.001$), 24-months (86.25%, $p<.001$), see Figure 3. These trials represent the initial exposure to the novel word-form as well as the first opportunity to link the label with a novel referent pair. Their behavior suggests that at a minimum, even young children's attention is directed toward the novel target when a novel label is present. This represents the children's initial exposure to the word-referent pair.

Table 2. ***Results of the models examining accuracy on referent selection trials***

Model & Predictors		***SE***	***Z***	***p***
A. **Main Model**				
Trial Type	1.444	.148	9.734	**<.001**
Age Group	.039	.162	.238	.812
Trial Type*Age Group	-.422	.134	-3.152	**.0016**
B. **Follow-up models adding vocabulary (exploratory)**				
18-month-olds				
Trial Type	1.715	.199	8.639	**<.001**
Total Vocabulary	.6596	.2240	2.945	**.003**
Trial Type*Vocabulary	-.546	.2244	-2.433	**.015**
24-month-olds				
Trial Type	.9473	.1894	5.001	**<.0001**
Total Vocabulary	.2359	.2155	1.095	.2737
Trial Type*Vocabulary	-.419	.185	-2.268	**.0233**
	adj	***SE***	***Z***	***p***
C. **Performance against chance**				
18-month-olds				
Novel Referent Selection	3.443	.398	8.651	**<.001**
Known Referent Selection	-.249	.258	-.967	.333
24-month-olds				
Novel Referent Selection	2.600	.331	7.867	**<.001**
Known Referent Selection	.695	.241	2.884	**.0039**

The variance for random effect of subject in Model A was .528, Model B 18mo was 0.00 and Model B 24mo was .322. Model C 18mo was .657 and 24mo was .417.
Note: Models C included only a fixed effect of trial type and a random effect of subject. Only the intercept, adjusted (βadj) to account for a chance level of 33%, was used to assess performance against chance.

However, performance on the known referent selection trials was not as strong. On the known referent selection trials, 24-month-old children accurately selected the target items above chance levels, 50.96% of the time, *p*=.004. Consistent with prior work (Kucker et al., 2018), younger 18-month-old children did not select known targets at levels different from chance, (33.09%, *p*=.333), instead selecting the novel foil item 67% of the time. Known foil items were only chosen 3% of the time. This suggests that children's responses to linguistic prompts can be swayed by the novelty or saliency of foil items (see also Pomper & Saffran, 2019), which may be due to the likely weaker prior knowledge in the younger group of children. This importantly replicates prior work showing that children perform well on novel referent selection, but that younger children can struggle to bring their vocabulary knowledge to bear in referent selection (Kucker, McMurray, & Samuelson, 2018).

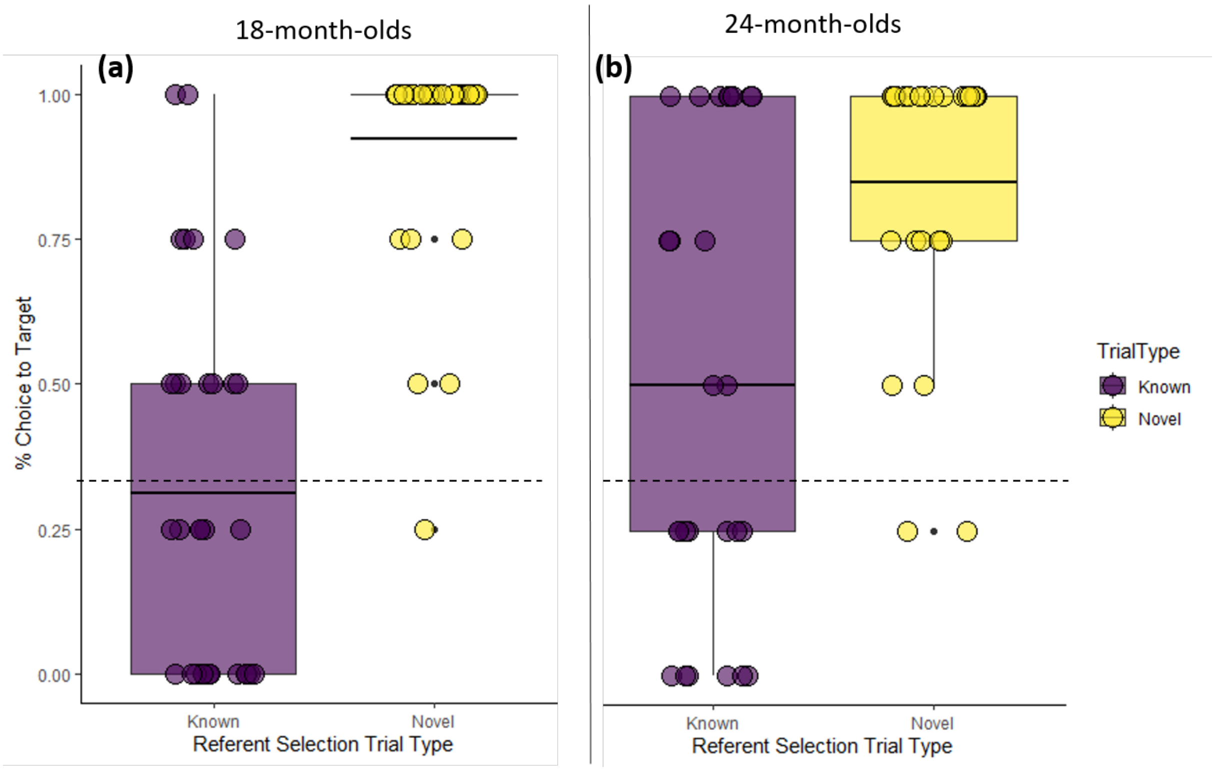


Figure 3. ***Average accuracy on RS trials for 18-month (a) and 24-month-old children (b).*** Dashed line represented chance (33%).

In order to further examine differences in performance by age group and across both trial types, a generalized linear mixed model (Table 2, Model A) of trial-by-trial performance was run with age group and trial type (Novel RS vs. Known RS, contrast coded respectively as +.5, -.5) as fixed factors. There was a significant interaction of age group and trial type as well as a significant main effect of trial type, suggesting that older 24-month-old children performed significantly better on the Known RS trials, but both ages performed equally well on Novel RS.

To understand the significant interaction, exploratory models were run for each age group, with trial-type and total vocabulary as fixed factors and a random intercept of subject (Table 2, Model B). The younger age group showed a significant effect of vocabulary and both ages showed significant effects of trial-type and significant interactions (Figure 4). In both age groups, children performed more accurately on Novel RS than Known RS and Known RS performance was positively correlated with vocabulary, but Novel RS performance was unaffected by vocabulary size (remained near ceiling).

Overall, performance on the Referent Selection trials mirrored prior work—children, regardless of age or vocabulary, accurately selected a novel item when given a novel

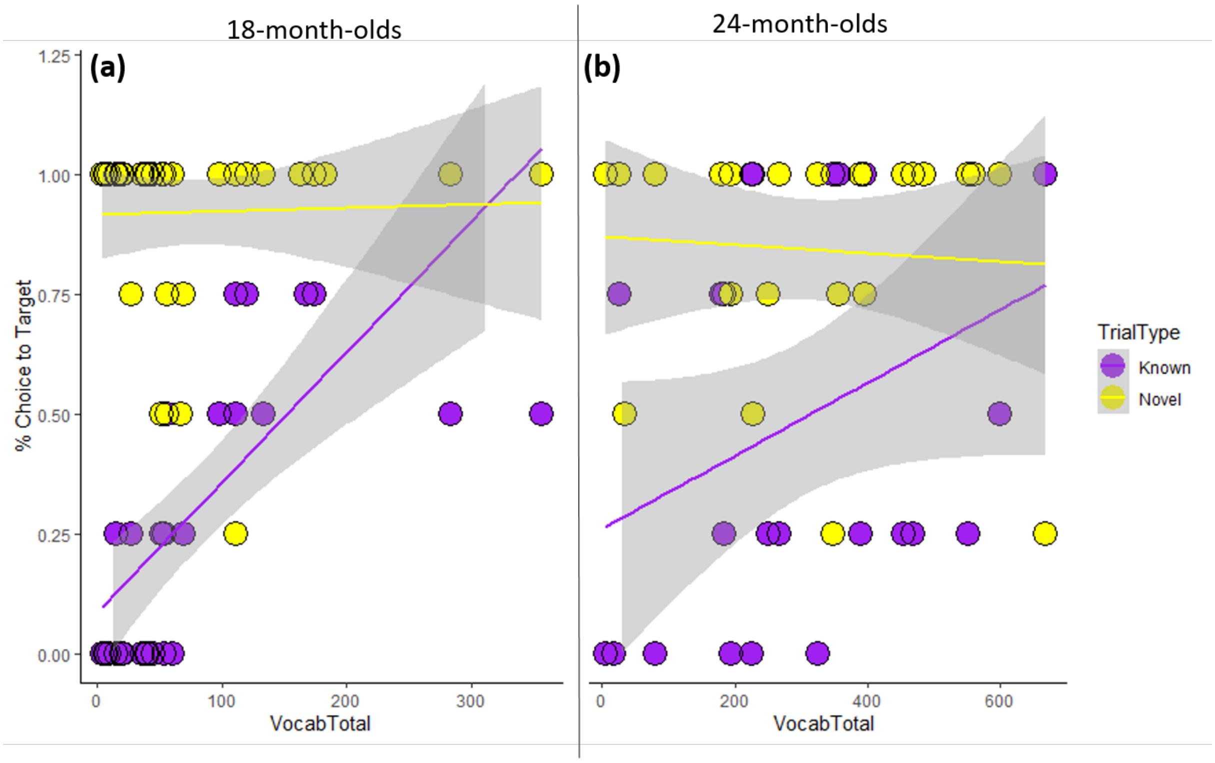


Figure 4. ***Correlations between average referent selection performance and vocabulary size for 18-month-old children (a) and 24-month-old children (b).*** Lines represent linear regressions and are for visualization purposes only.

label on Novel RS trials. Moreover, selection of a known item by name was predicted by a child's age (18-month-olds perform worse than 24-month-olds) and vocabulary (higher vocabularies perform better). Thus, real-time responding depends on the child's knowledge of the word (both form and referent) and their vocabulary level. Perhaps most pertinent is that the results here reproduce that of prior work (Horst & Samuelson, 2008; Kucker et al., 2018), showing differential performance between ages and trial types. The relatively poor performance of 18-month-old children on the Known RS trials is especially noteworthy as it calls in to question the mechanisms driving referent selection during Novel RS. Nonetheless, we know that exposure during these trials represent a critical first opportunity to lay down initial word-referent traces (McMurray et al., 2012), specially because children are clearly attending to the novel to-be-learned item (though if they also attended to the novel label is unknown, and one key question for the current study). Regardless of accuracy, at this point all children had been exposed to a set of known and novel word forms that co-occurred with specific referents (e.g., "cup" was heard when there was a cup present on the trial; "roke" was heard when its corresponding novel item was present). Thus, even if children did not select the correct item or did not listen to the label, it is still possible this exposure influenced their subsequent performance (Yurovsky et al., 2014).

Preferential Listening

The goals of preferential listening task were to 1) expose children to additional word-form repetitions before testing their retention, and 2) test children's memory for auditory word-forms. To assess preference for word-forms heard during referent selection (RS) compared to novel words (NN), a linear mixed model (Table 3) was run with age group and word-type (RS vs. NN, contrast coded as +.5, -.5). The best fitting model included a random intercept of subject.

Table 3. ***Result of the model predicting preferential listening performance***

Model & Predictors	***β***	***SE***	***t***	***p***
Word Type	-.0314	.0428	-.733	.464
Age Group	.166	.0960	1.727	**.091**
Word Type*Age Group	.0022	.043	.052	.959

Note: Random effect of subject variance was .346

There was a marginal main effect of age group (*p*=.091), suggesting 24-month-old children spent slightly more time listening overall to all word types, but there were no differences in listening by word-type or interactions of age with word type; see Figure 5. There was also no significant relationship between RS performance and listening time, see Supplemental Materials.

The lack of difference in listening times for RS words compared to NN words suggests that children could not differentiate between them and may not have retained the word-forms presented in the referent selection task. However, there was a lot of within child variability—individual children listened to some word-forms for only two seconds (only 2 repetitions of the word-form) and others for nearly a minute (up to 30 repetitions). Given that children in this study were slightly older than those in traditional head-turn tasks, this variability is not surprising. Familiarity vs. novelty biases are not always consistent in infants and can even be seen to switch from trial to trial, especially in children closer to this age (DePaolis et al., 2016; Fisher-Thompson, 2014; and see Mather, 2013 for a review). The within-subject variability thus likely obscured any potential evidence of word-form retention. As we describe further in the General Discussion, one possibility is children did learn words during Referent Selection, but that this preferential listening task was simply not sensitive enough to capture such learning. Another possibility is that children's representations of the novel words presented in the referent selection task were not robust enough to support differentiation between those words and new novel word-forms in the preferential listening task. These points notwithstanding, it is still the case that the preferential listening task provided the infants with additional exposure to the word-forms. Thus, we next asked whether this extra exposure impacted learning, and in particular if it supported retention of the new mappings.

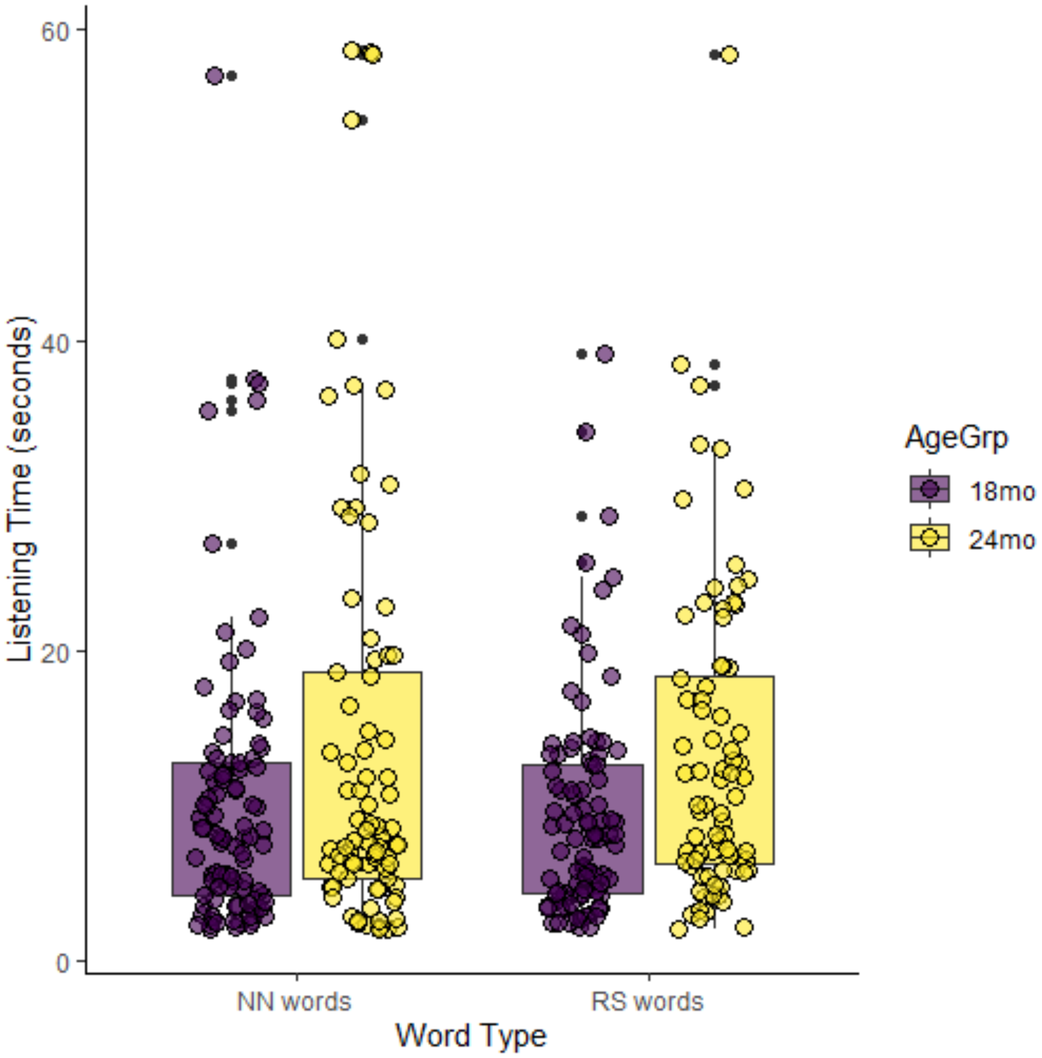


Figure 5. ***Listening time to NN words and RS words by 18- and 24-month-old infants***
Note: Points represent each child's listening for each individual word form (up to 8/participant), box represents mean listening time for each age group and word type.

Retention

The goal of retention was to ask if children recalled novel word-referent forms initially presented during referent selection. A key question is how retention was impacted by exposure to word-forms during the preferential listening phase. Consistent with prior work which found that 18-month-old children were at chance levels (33%) on retention trials (Kucker et al., 2018), 18-month-old children here were at chance on retention (40.3%, *p*=.245), suggesting they did not retain the novel word-referent pairs despite extra exposure. Children also chose the other named foil item at chance levels, 33.3% of the time, β_{adj}=0.00, SE=.26, z=.00, *p*=1.0, and the unnamed foil item the remainder of the time, demonstrating that they do consider all possible options available and showed no preference or evidence of knowing which items had labels.

However, contrary to prior work with 24-month-olds that had a retention rate of 36%, not significantly different from chance (Horst & Samuelson, 2008), 24-month-old children here were significantly *above* chance (54.9%, *p*=.0128; Table 4 Model B, Figure 6; see also Supplemental Materials Table S5). They also chose the other named foil item at levels significantly lower than chance, selecting it only 19.2% of the time, β_{adj}=.74, SE=.35, z=2.11, *p*=.035. Though there was no statistically significant effect of age group on retention (Table 4, Model A), 24-month-old children chose the labeled item more than foil items. Moreover, this is the same group of children who were marginally more likely to listen longer during preferential listening. This raises the possibility that additional exposures to multiple word-forms (both RS and NN) may have increased subsequent retention. However, further exploratory models predicting retention from a child's referent selection performance and/or listening preferences for specific word-forms were largely non-significant (see Supplementary material). Exploratory analyses (Table 4, Model B) suggested vocabulary was not a significant moderator. Thus, the retention effects are subtle, but noteworthy as they hint at one possible avenue for boosting learning of new words: repetition of word-forms.

Table 4. ***Results of the models examining retention performance***

Model & Predictors		***SE***	***Z***	***p***
A. Main Model				
Age Group	.317	.215	1.473	.141
B. Follow-up models adding vocabulary (exploratory)				
18-month-olds				
Total Vocabulary	.645	.5625	1.148	.251
24-month-olds				
Total Vocabulary	.0957	.342	.280	.779
	adj	***SE***	***Z***	***p***
C. Performance against chance				
18-month-olds	.301	.258	1.163	.2448
24-month-olds	.948	.381	2.488	**.0128**

Note: The variance for the random effect of subject for Model A was .319, Model B 18mo was 0.00, Model B 24mo was .874, and Model C was 0.00 for 18mo and 1.06 for 24mo.

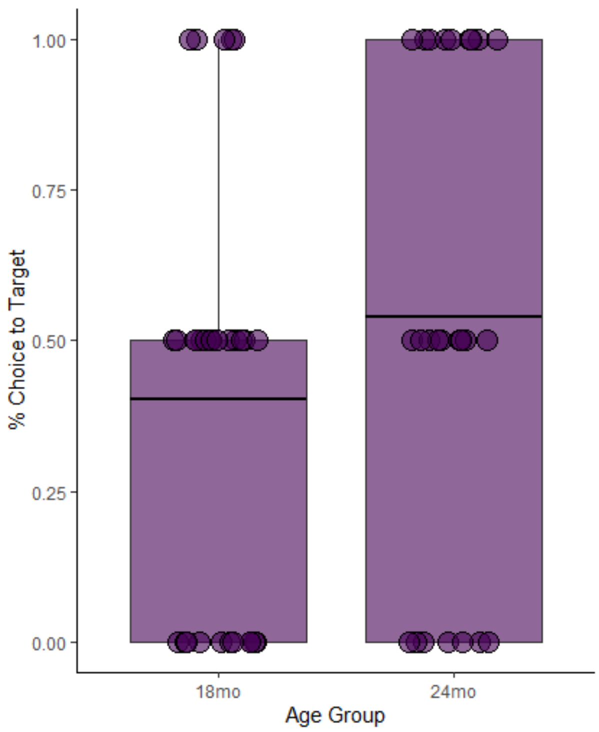


Figure 6. ***Average retention performance for 18- and 24-month-old children***

While the retention results were largely non-significant, they critically mirror prior work showing retention is very difficult at this age, and in particular that most children in this age-range do not retain words in this and similar paradigms (Bion et al., 2013; Horst & Samuelson, 2008; Kucker et al., 2018). However, 24-month-old children retaining words at above-chance levels shows possible evidence of downstream effects of word-form repetition on retention. While we did not find effects of individual words, it could be that general exposure to word forms hones the wider lexical network, thereby boosting retention without individual-level benefits. While these results should be taken with caution, they fit with theoretical accounts suggesting incremental changes in word-form representations from initial exposure to final retention for this age group.

Discussion

Word learning emerges over multiple cascading moments. A referent is selected in the moment after hearing a word-form and initial word-referent links are formed. These word-referent links are later reinforced and refined over exposures, ultimately leading to retention. The incremental nature of this process is partially supported by the current study. We replicated prior findings that young children reliably select novel referents on request but that 18-month-old children demonstrate a novelty bias when asked to select familiar, well-known referents. We also find support for the role of vocabulary in both known item selection and novelty bias by 18-month-olds.

While preferential listening did not reflect preferences for words presented during referent selection compared to new words (Goal 1), the 24-month-old age group (who performed better on the Known RS trials) listened marginally longer during the preferential listening phase. These older 24-month-old children also showed above chance retention after the additional exposure to word forms provided by the preferential listening task. This is notable as retention in this age group is not typically seen in similar paradigms without additional exposure (Goal 2). Neither of these last two findings held for 18-month-old children, suggesting possible changes in these effects during the period of early vocabulary development (Goal 3).

To learn a new word, a child has to make a robust association between a word-form and a referent but doing so is a time extended process. What is confirmed and replicated here is all children can easily select the novel referent when given a novel word-form, but younger 18-month-old children fail to select known referents when given a known name. Given that these younger children's vocabulary representations are likely less robust, this suggests that in-the-moment of referent selection strong novelty biases can override relatively weak lexical knowledge (see also Kucker et al., 2018). Thus, whatever novel words are mapped during referent selection may be driven by low-level perceptual processes and need additional reinforcement before supporting retention (see Mather, 2013). Indeed, 18-month-old children with high novelty biases did not show evidence of retention in the current study. As suggested by Kucker and colleagues (2018), the increased attention to novelty in referent selection may be a reflection of weaker lexical knowledge, and we know that children with weaker vocabulary skills have difficulty with mapping (Kucker & Seidler, 2022) and retention (Bion et al., 2013). This was likely true here as well as vocabulary was a significant predictor of RS performance in the 18-month-olds. However, vocabulary did not correlate with listening or retention, suggesting that at this age the strength of lexical knowledge may play less of a cascading, interactive role when it comes to word-form repetitions.

We hoped the preferential listening task would provide an intermediate test of word-form recognition, but there were no consistent differences in children's listening to word-forms from RS compared to novel words. However, there was also wide variability in listening times that likely masked systematic differences in memory. Preferences for novel vs. familiar words are known to shift between and within children in this age-range (DePaolis et al., 2016). In hindsight, interpreting preferential listening time as indicting "learning" is difficult (see also Cohen, 2004; Mather, 2013) especially at this age; does listening longer to the words from referent selection mean children did not finish encoding during RS, or does longer listening indicate they perceive the word-form as new? Differences in listening were thus ultimately not informative by themselves except to lend caution to future work using such a paradigm to capture differences at this age.

However, though it might be a weak measure of preference, the preferential listening task did allow individual children to control which word-forms they heard again. Children could self-select which word-forms they wanted more exposure to by continuing to look at the screen, much like the procedure used in Kucker and Samuelson (2012). As evidence of this opportunity, 24-month-old children did choose to listen slightly longer overall suggesting age differences in focus to auditory stimuli. We know that such additional exposure to word-forms can improve learning (Graf Estes, Evans, Alibali, et al., 2007; Hayne & Herbert, 2004) and indeed the 24-month-old group of children who listened more to the words from RS did demonstrate evidence of retention. Critically, retention for 24-month-old children here was at 51.3%, substantially higher than in prior comparison studies without a preferential listening period—in Horst and Samuelson (2008), 24-month-old children without familiarization showed 36% retention.

Thus, the results cautiously suggest the possibility that the extra exposures to word forms during the preferential listing task may have had downstream impacts on learning for some children. These results should be taken with caution though given the relatively small sample and limitations of the headturn preference task at this age. Moreover, it is important to point out that other work finds stronger benefits for additional object exposure. In Kucker and Samuelson (2012), children pre-familiarized with the novel objects retained over 70% of the time. The differential impacts of word-form and object familiarity are in some ways not surprising—word-forms are fleeting and harder to encode (Stager & Werker, 1997), more prone to decay over delay (Munro et al., 2012), and thus may require substantially more than just a half a dozen repetitions to have the same impact as one-minute of object familiarization. Indeed, other early word-learning work suggests that multiple exposures to the word-referent pair is necessary for robust mapping (Axelsson et al., 2012; Twomey et al., 2013); here children only heard the word-form and referent together once, which may also explain the spurious learning.

Alternatively, it is possible that referent selection, listening time, and retention are all related to a third individual difference factor that may be stronger in 24-month-old children compared to 18-month-olds. The literature suggests that vocabulary knowledge may be a possibility (Kalashnikova et al., 2016; Samuelson, 2021). However, given that neither listening time nor retention was related to vocabulary in the 24-month-old group, it is less likely that lexical ability is responsible here. This does not preclude other lower-level influences on word learning, however, and future work should further explore how attention, memory, and novelty play a role in word-form repetition and retention. One promising possibility is attraction to novelty which we know shifts over this same age range as vocabulary increases and executive function skills improve (Kucker et al., 2018; Samuelson, 2021).

Taken together, these results contribute to evidence of the moment-to-moment cascade of word learning and suggest that additional exposure to novel word forms after initial mapping may aid in retention. Variability between individual children in this process, and additional methodologies that can capture it, will be critical to examine in future work.

References

Aussems, S., & Kita, S. (2019). Seeing iconic gestures while encoding events facilitates children's memory of these events. *Child Development, 90*(4), 1123–1137. https://doi.org/10.1111/cdev.12988

Axelsson, E.L., Chuchley, K., & Horst, J.S. (2012). The right thing at the right time: Why ostensive naming facilitates word learning. *Frontiers in Psychology, 3*, https://doi.org/10.3389/fpsyg.2012.00088

Axelsson, E. L., & Horst, J. S. (2014). Contextual repetition facilitates word learning via fast mapping. *Acta Psychologica, 152*, 95–99. https://doi.org/10.1016/j.actpsy.2014.08.002

Bauer, P. J., Burch, M. M., Scholin, S. E., & Güler, O. E. (2007). Using cue words to investigate the distribution of autobiographical memories in childhood. *Psychological Science, 18*(10), 910–916. https://doi.org/10.1111/j.1467-9280.2007.01999.x

Bion, R. A. H., Borovsky, A., & Fernald, A. (2013). Fast mapping, slow learning: Disambiguation of novel word–object mappings in relation to vocabulary learning at 18, 24, and 30months. *Cognition, 126*(1), 39–53. https://doi.org/10.1016/j.cognition.2012.08.008

Breen, E., Pomper, R., & Saffran, J. (2019). Phonological earning influences label–object mapping in toddlers. *Journal of Speech, Language, and Hearing Research, 62*(6), 1923–1932. https://doi.org/10.1044/2019_JSLHR-L-18-0131

Capone, N. C., & McGregor, K. K. (2005). The effect of semantic representation on toddlers' word retrieval. *Journal of Speech Language and Hearing Research, 48*(6), 1468–1480. https://doi.org/10.1044/1092-4388(2005/102)

Carey, S. (2010). Beyond fast mapping. *Language Learning and Development, 6*(3), 184–205. https://doi.org/10.1080/15475441.2010.484379

Clerkin, E. M., Hart, E., Rehg, J. M., Yu, C., & Smith, L. B. (2017). Real-world visual statistics and infants' first-learned object names. *Philosophical Transactions of the Royal Society B: Biological Sciences, 372*(1711), 20160055. https://doi.org/10.1098/rstb.2016.0055

Clerkin, E.M., & Smith, L.B. (2022). Real-world statistics at two timescales and a mechanism for infant learning of object names. *PNAS, 119*(18), e2123239119. https://doi.org/10.1073/pnas.2123239119

Cohen, L. B. (2004). Uses and misuses of habituation and related preference paradigms. *Infant and Child Development, 13*(4), 349–352. https://doi.org/10.1002/icd.355

Cohen, L. B., Atkinson, D. J., & Chaput, H. H. (2004). *Habit X: A new program for obtaining and organizing data in infant perception and cognition studies (1.0).*

Dale, P. S., & Fenson, L. (1996). Lexical development norms for young children. *Behavior Research Methods, Instruments, & Computers, 28*(1), 125–127. https://doi.org/10.3758/BF03203646

DePaolis, R. A., Keren-Portnoy, T., & Vihman, M. (2016). Making sense of infant familiarity and novelty responses to words at lexical onset. *Frontiers in Psychology*, 7. https://doi.org/10.3389/fpsyg.2016.00715

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

Fisher-Thompson, D. (2014). Exploring the emergence of side biases and familiarity-novelty preferences from the real-time dynamics of infant looking. *Infancy, 19*(3), 227–261. https://doi.org/10.1111/infa.12051

Fitneva, S. A., & Christiansen, M. H. (2011). Looking in the wrong direction correlates with more accurate word learning. *Cognitive Science, 35*(2), 367–380. https://doi.org/10.1111/j.1551-6709.2010.01156.x

Fitneva, S. A., & Christiansen, M. H. (2017). Developmental changes in cross-situational word learning: The inverse effect of initial accuracy. *Cognitive Science, 41*, 141–161. https://doi.org/10.1111/cogs.12322

Gallerani, C. M., Saylor, M. M., & Adwar, S. (2009). Mother–infant conversation about absent things. *Language Learning and Development, 5*(4), 282–293. https://doi.org/10.1080/15475440902897604

Gathercole, S. (2006). Complexities and constraints in nonword repetition and word learning. *Applied Psycholinguistics, 27*(4), 599-613. doi:10.1017/S014271640606053X

Gathercole, S. E., Hitch, G. J., Service, E., & Martin, A. J. (1997). Phonological short-term memory and new word learning in children. *Developmental Psychology, 33*(6), 966–979. https://doi.org/10.1037/0012-1649.33.6.966

Gordon, K. R., & McGregor, K. K. (2014). A spatially supported forced-choice recognition test reveals children's long-term memory for newly learned word forms. *Frontiers in Psychology*, 5. https://doi.org/10.3389/fpsyg.2014.00164

Gordon, K. R., McGregor, K. K., Waldier, B., Curran, M. K., Gomez, R. L., & Samuelson, L. K. (2016). Preschool children's memory for word forms remains stable over several days, but gradually decreases after 6 months. *Frontiers in Psychology*, 7. https://doi.org/10.3389/fpsyg.2016.01439

Graf Estes, K., Evans, J. L., Alibali, M. W., & Saffran, J. R. (2007). Can infants map meaning to newly segmented words?: Statistical segmentation and word learning. *Psychological Science, 18*(3), 254–260. https://doi.org/10.1111/j.1467-9280.2007.01885.x

Graf Estes, K., Evans, J. L., & Else-Quest, N. M. (2007). Differences in the nonword repetition performance of children with and without Specific Language Impairment: A meta-analysis. *Journal of Speech, Language, and Hearing Research, 50*(1), 177–195. https://doi.org/10.1044/1092-4388(2007/015)

Halberda, J. (2003). The development of a word-learning strategy. *Cognition, 87*(1), B23–B34. https://doi.org/10.1016/S0010-0277(02)00186-5

Hayne, H., & Herbert, J. (2004). Verbal cues facilitate memory retrieval during infancy. *Journal of Experimental Child Psychology, 89*(2), 127–139. https://doi.org/10.1016/j.jecp.2004.06.002

Hodges, R., Munro, N., Baker, E., McGREGOR, K., Docking, K., & Arciuli, J. (2016). The role of elicited verbal imitation in toddlers' word learning. *Journal of Child Language, 43*(02), 457–471. https://doi.org/10.1017/S0305000915000240

Hollich, G. J., Hirsh-Pasek, K., Golinkoff, R. M., Brand, R. J., Brown, E., Chung, H. L., Hennon, E., Rocroi, C., & Bloom, L. (2000). Breaking the language barrier: An emergentist coalition model for the origins of word learning. *Monographs of the Society for Research in Child Development, 65*(3,), i-vi+1-135.

Horst, J. S., & Hout, M. C. (2016). The Novel Object and Unusual Name (NOUN) Database: A collection of novel images for use in experimental research. *Behavior Research Methods, 48*(4), 1393–1409. https://doi.org/10.3758/s13428-015-0647-3

Horst, J. S., & Samuelson, L. K. (2008). Fast mapping but poor retention by 24-month-old infants. *Infancy, 13*(2), 128–157. https://doi.org/10.1080/15250000701795598

Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology, 29*(1), 1–23. https://doi.org/10.1006/cogp.1995.1010

Jusczyk, P. W., & Hohne, E. A. (1997). Infants' memory for spoken words. *Science, 277*(5334), 1984–1986. https://doi.org/10.1126/science.277.5334.1984

Kalashnikova, M., Mattock, K., & Monaghan, P. (2016). Mutual exclusivity develops as a consequence of abstract rather than particular vocabulary knowledge. *First Language, 36*(5), 451–464. https://doi.org/10.1177/0142723716648850

Kucker, S. C., McMurray, B., & Samuelson, L. K. (2015). Slowing down fast mapping: Redefining the dynamics of word learning. *Child Development Perspectives, 9*(2), 74–78. https://doi.org/10.1111/cdep.12110

Kucker, S. C., McMurray, B., & Samuelson, L. K. (2018). Too much of a good thing: How novelty biases and vocabulary influence known and novel referent selection in 18-month-old children and associative learning models. *Cognitive Science, 42*, 463–493. https://doi.org/10.1111/cogs.12610

Kucker, S. C., McMurray, B., & Samuelson, L. K. (2020). Sometimes it is better to know less: How known words influence referent selection and retention in 18- to 24-month-old children. *Journal of Experimental Child Psychology, 189,* 104705. https://doi.org/10.1016/j.jecp.2019.104705

Kucker, S. C., & Samuelson, L. K. (2012). The first slow step: Differential effects of object and word-form familiarization on retention of fast-mapped words. *Infancy, 17*(3), 295–323. https://doi.org/10.1111/j.1532-7078.2011.00081.x

Kucker, S. C., & Seidler, E. (2022). The timescales of word learning in children with language delays: In-the-moment mapping, retention, and generalization. *Journal of Child Language,* 1–29. https://doi.org/10.1017/S0305000921000817

Lewis, M., Cristiano, V., Lake, B. M., Kwan, T., & Frank, M. C. (2020). The role of developmental change and linguistic experience in the mutual exclusivity effect. *Cognition, 198,* 104191. https://doi.org/10.1016/j.cognition.2020.104191

Mather, E. (2013). Novelty, attention, and challenges for developmental psychology. *Frontiers in Psychology,* 4. https://doi.org/10.3389/fpsyg.2013.00491

McMurray, B., Horst, J. S., & Samuelson, L. K. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological Review, 119*(4), 831–877. https://doi.org/10.1037/a0029872

Mervis, C. B., & Bertrand, J. (1994). Acquisition of the Novel Name-Nameless Category (N3C) Principle. *Child Development, 65*(6), 1646. https://doi.org/10.2307/1131285

Mollica, F., & Piantadosi, S. T. (2017). How data drive early word learning: A cross-linguistic waiting time analysis. *Open Mind, 1*(2), 67–77. https://doi.org/10.1162/OPMI_a_00006

Munro, N., Baker, E., McGregor, K., Docking, K., & Arculi, J. (2012). Why word learning is not fast. *Frontiers in Psychology*, 3. https://doi.org/10.3389/fpsyg.2012.00041

Oakes, L.M., Sperka, D., DeBolt, M.C., & Cantrell, L.M. (2019). Habit2: A stand-alone software solution for presenting stimuli and recording infant looking times in order to study infant development. *Behavior Research Methods, 51*(5), 1943-1952. http://dx.doi.org.proxy.libraries.smu.edu/10.3758/s13428-019-01244-y

Pomper, R., & Saffran, J. R. (2019). Familiar object salience affects novel word learning. *Child Development, 90*(2), e246–e262. https://doi.org/10.1111/cdev.13053

Robinson, C. W., & Sloutsky, V. M. (2004). Auditory dominance and its change in the course of development. *Child Development, 75*(5), 1387–1401. https://doi.org/10.1111/j.1467-8624.2004.00747.x

Robinson, C. W., & Sloutsky, V. M. (2010). Effects of multimodal presentation and stimulus familiarity on auditory and visual processing. *Journal of Experimental Child Psychology, 107*(3), 351–358. https://doi.org/10.1016/j.jecp.2010.04.006

Samuelson, L. K. (2021). Toward a precision science of word learning: Understanding individual vocabulary pathways. *Child Development Perspectives, 15*(2), 117–124. https://doi.org/10.1111/cdep.12408

Schwab, J. F., & Lew-Williams, C. (2016). Repetition across successive sentences facilitates young children's word learning. *Developmental Psychology, 52*(6), 879–886. https://doi.org/10.1037/dev0000125

Seerdorff, M., Oleson, J., & McMurray, B. (n.d.). *Maybe maximal: Good enough mixed models optimize power while controlling Type I error.*
Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature, 388*(6640), 381–382. https://doi.org/10.1038/41102

Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive Psychology, 66*(1), 126–156. https://doi.org/10.1016/j.cogpsych.2012.10.001

Twomey, K. E., Ranson, S. L., & Horst, J. S. (2014). That's more like it: Multiple exemplars facilitate word learning. *Infant and Child Development, 23*(2), 105–122. https://doi.org/10.1002/icd.1824

Vitevitch, M.S. & Luce, P.A. (2004) A web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, and Computers, 36,* 481-487.

Vlach, H.A. & Sandhofer, C.M. (2012). Distributing learning over time: the spacing effect in children's acquisition and generalization of science concepts. *Child Development, 83*(4), 1137-44. DOI: 10.1111/j.1467-8624.2012.01781.x

Warren, D. E., & Duff, M. C. (2014). Not so fast: Hippocampal amnesia slows word learning despite successful fast mapping. *Hippocampus, 24*(8), 920–933. https://doi.org/10.1002/hipo.22279

Willits, J. A., Wojcik, E. H., Seidenberg, M. S., & Saffran, J. R. (2013). Toddlers activate lexical semantic knowledge in the absence of visual referents: Evidence from auditory priming. *Infancy, 18*(6), 1053–1075. https://doi.org/10.1111/infa.12026

Wojcik, E. H., & Saffran, J. R. (2013). The ontogeny of lexical networks: Toddlers encode the relationships among referents when learning novel words. *Psychological Science, 24*(10), 1898–1905.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review, 114*(2), 245–272. https://doi.org/10.1037/0033-295X.114.2.245

Yurovsky, D., Fricker, D. C., Yu, C., & Smith, L. B. (2014). The role of partial knowledge in statistical word learning. *Psychonomic Bulletin & Review, 21*(1), 1–22. https://doi.org/10.3758/s13423-013-0443-y

Zettersten, M., Wojcik, E., Benitez, V. L., & Saffran, J. (2018). The company objects keep: Linking referents together during cross-situational word learning. *Journal of Memory and Language, 99*, 62–73. https://doi.org/10.1016/j.jml.2017.11.001

Zhao, L., Packard, S., McMurray, B., & Gupta, P. (2019). Similarity of referents influences the learning of phonological word forms: Evidence from concurrent word learning. *Cognition, 190*, 42–60. https://doi.org/10.1016/j.cognition.2018.12.004

Data, code and materials availability statement

De-identified participant-level data for preferential listening and referent selection, datasheets used for data collection, instructions for coding children's choices in referent selection, sound files used in preferential listening, and R scripts used for analysis are available at https://osf.io/9t8fk/. Images of items used as stimuli in referent selection are seen in Figure 1. The Editor agreed an exemption (3rd August 2023) to materials sharing for the MBCID as it is subject to copyright. It is available from the publisher at: https://brookespublishing.com/webcdi/

Ethics statement

Ethics approval was obtained from the ethics committee of the University of Iowa. All participants gave informed written consent before taking part in the study.

Authorship and Contributorship Statement

LKS and BM conceived of the study and jointly designed the study with SCK. SCK wrote the first draft of the manuscript, collected the data, and analyzed the results. LKS and BM revised the manuscript and aided in analysis. Funding for the project was provided by LKS. All authors approved the final version of the manuscript and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Acknowledgements

This work was supported by the *Eunice Kennedy Shriver National Institute of Child Health and Human Development* (5R01HD045713), National Institutes of Health, U.S. Department of Health and Human Services. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Supplementary Materials

Additional details pertaining to participants, data, analyses, and results are below.

Participants and data cleaning details

All referent selection and retention trials in which the child made a distinguishable choice were included in the analysis. Over 90% of trials were kept. All preferential listening trials in which the child listened for at least 2000 milliseconds were kept for analysis. For preferential listening, a total of 27 (of 464, 6%) NN and 32 (of 464, 7%) RS preferential listening trials were dropped; all children had data from at least 2 (of 4) NN word trials and all but one child had at least 2 (of 4) RS word trials. The one exception was a child missing 3 RS word trials whose data was retained for the remaining word. One 18-month-old child missing retention data and one 24-month-old child without a completed MCDI and were dropped from those respective analyses. Thus, all 59 children contributed data.

Additional Results

Additional exploratory analyses were run to examine the impacts of vocabulary and relations in performance across tasks. A final set of analyses compared the results here to that of prior work (Horst & Samuelson, 2008; Kucker et al., 2018), see Table S5.

Preferential Listening

In preferential listening, there were no differences in listening for word type; 18-month-old children listened to words from RS an average of 8.27 seconds (SD=4.55) and NN words 9.19 seconds (SD=6.49), 24-month-olds listened to RS for 11.34 seconds (SD=7.47) and NN words for 11.56 seconds (SD=10.28).[1] However, there was also significant variability in children's listening during preferential listening (Figure 5 in main text); given the already established variability in referent selection (RS) performance (Figure 4 in main text), this raised the question of whether listening time might be related to how children did during RS. That is, children who performed better during RS might have been expected to retain the novel word forms better. To test this, further exploratory analyses were run predicting listening time from prior RS performance (Table S1, Model B, below). This linear mixed model included fixed effects of accuracy on Known RS (centered), accuracy on Novel RS (centered), age group, and word type. The results were largely non-significant, however there was a marginal interaction of Age Group and Known RS. Analyses of vocabulary for each age group were also non-significant (Table S1, model C, below). Thus, there were no differences

[1] One word repetition was heard every 2 seconds. Thus, 18-month-olds heard RS words an average of 4 times and NN words 4.5 times, and 24-month-olds heard both RS and NN words an average of 5.5 times.

in overall listening times for type of word, and only a hint that individual older children who did better on Known RS may have listened to more word-form repetitions during this phase.

Table S1. ***Result of the model predicting preferential listening performance***

Model & Predictors	*β*	*SE*	*t*	*p*
A. Main model				
Word Type	-.0314	.0428	-.733	.464
Age Group	.166	.0960	1.727	**.091**
Word Type*Age Group	.0022	.043	.052	.959
B. Adding RS performance as predictor (exploratory)				
Novel RS accuracy	-.1887	.115	-1.648	**.107**
Known RS accuracy	.0383	.0626	.611	.544
Age Group	.157	.093	1.692	**.098**
Word Type	-.0204	.0434	-.469	.640
Novel RS*Age Group	-.0232	.116	-.201	.842
Novel RS*Word Type	-.0207	.0527	-.393	.694
Known RS*Age Group	.113	.062	1.829	**.0746**
Known RS*Word Type	-.033	.0295	-1.132	.259
Novel RS*Age*Word Type	-.001	.053	-.025	.980
Known RS*Age*Word Type	-.048	.029	-1.644	**.1013**
C. Follow-up models adding vocabulary (exploratory)				
18-month-olds				
Word Type	-.0425	.060	-.708	.480
Total vocabulary	-.0618	.135	-.458	.651
Word Type*Vocabulary	-.0422	.060	-.701	.484
24-month-olds				
Word Type	-.033	.0664	-.502	.617
Total Vocabulary	.0156	.151	.103	.919
Word Type*Vocabulary	-.035	.0667	-.519	.604

Retention

Children's ability to retain the novel word-referent mappings from referent selection were tested in the final phase of the experiment. In addition to the main effects of age group, exploratory analyses examined the impact of vocabulary size on performance (Table S2, Model A; Figures S1 and S2, below). A final set of analyses explored how preferential listening performance related to retention (Table S3, below).

Table S2. ***Results of the models examining retention performance***

Model & Predictors		***SE***	***Z***	***p***
A. Main model				
Age Group	.317	.215	1.473	.141
B. Adding RS performance as predictor (exploratory)				
Age Group	.224	.215	1.043	.297
KnownRS performance	.215	.147	1.468	.142
NovelRS performance	-.229	.258	-.887	.375
Age*Known RS	-.087	.142	-.616	.538
Age*Novel RS	.461	.254	1.815	**.070**^{m}
C. Follow-up models adding vocabulary (exploratory)				
18-month-olds				
Total Vocabulary	.645	.5625	1.148	.251
24-month-olds				
Total Vocabulary	.0957	.342	.280	.779
	$_{adj}$	***SE***	***Z***	***p***
B. Performance against chance				
18-month-olds	.301	.258	1.163	.2448
24-month-olds	.948	.381	2.488	**.0128**

Note: Models C included only a random effect of subject. Only the intercept (which was adjusted for chance at 33%) was used to assess performance in this model.

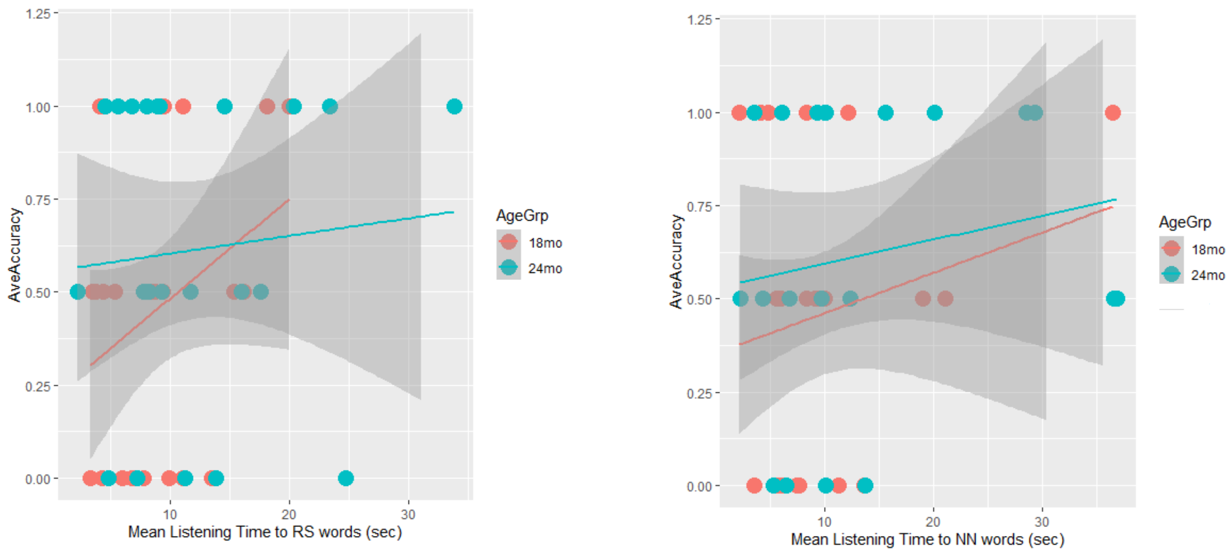


Figure S1. Average retention performance and listening time to RS words (left) and novel words (right), according to age group

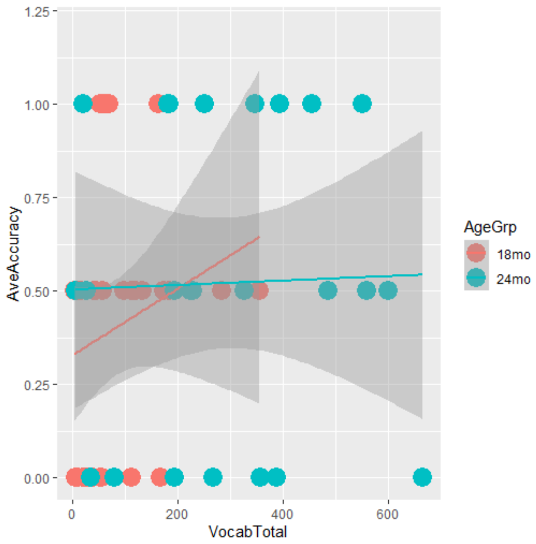


Figure S2. ***Retention performance as predicted by vocabulary size in 18- and 24-month-old infants***

Table S3. ***Results of the model predicting retention from Average listening preferences***

Model & Predictors		***SE***	***Z***	***p***
A. Predicting retention from average listening preferences				
Age Group	.2095	.2536	.826	.409
Ave RS listening time	.585	.721	.811	.418
Ave Novel listening time	.851	1.085	.784	.433
RS listening*Age Group	-.7663	.695	-1.103	.270
Novel listening*Age Group	.302	1.062	.284	.777
B. Follow-up models adding vocabulary (exploratory)				
18-month-olds				
Ave RS listening time	.825	.504	1.635	**.102**
Ave Novel listening time	-.174	.510	-.341	.733
Total Vocabulary	.440	.406	1.084	.279
RS listening*Vocabulary	.564	.545	1.035	.301
Novel listening*Vocabulary	-.901	.806	-1.118	.264
24-month-olds				
Ave RS listening time	-.927	1.237	-.749	.4454

Ave Novel listening time	3.269	2.229	1.467	.142
Total Vocabulary	-.343	.485	-.708	.479
RS listening*Vocabulary	.201	1.829	.110	.913
Novel listening*Vocabulary	-3.092	2.196	-1.408	.159
C. Predicting retention from listening time to specific words				
Age Group	.307	.240	1.278	.201
Listening time	.376	.3197	1.175	.240
Age Group*Listening	.073	.321	.228	.819
D. Follow-up models adding vocabulary (exploratory)				
18-month-olds				
Listening Time	.306	.369	.830	.407
Total Vocabulary	.252	.317	.795	.426
Listening*Vocabulary	.1995	.461	.433	.665
24-month-olds				
Listening Time	.306	.369	.830	.407
Total Vocabulary	.252	.317	.795	.426
Listening*Vocabulary	.1995	.461	.433	.665

Comparison with prior work

In order to compare to prior work using identical procedures but without a preferential listening period, between groups t-tests were run. The 18-months-olds of the current study were compared to that in Experiment 1 of Kucker et al. (2018) and 24-months compared to Experiment 1a of Horst and Samuelson (2008). There were no differences in 18-month-olds Known RS or Retention performance, though children here did perform better on Novel RS. For 24-month-old children, those here were marginally less likely to select the target on both Known and Novel RS trials (and both at levels still above chance). There was no difference in retention, though 24-month-old children in the current study were above chance. See Table S4, below.

Table S4. ***Proportion of correct trials***

	Known RS	Novel RS	Retention
18-month-olds			
Kucker et al. (2018), E1	.31 (.31)	.78 (.27)*	.33 (.36)
Current Study	.33 (.34)	.96 (.18)*	.36 (.35)
Between groups comparison	*t*(64)=.300, *p*=.765 [-.18, .14]	*t*(64)=3.231, *p*=.002 [-.29, -.07]	*t*(59)=.362, *p*=.718 [-.22, .15]
24-month-olds			
Horst & Samuelson (2008), E1a	.72 (.25)*	.69 (.18)*	.36 (.23)
Current Study	.51 (.42)*	.86 (.24)*	.51 (.41)*

Between groups comparison	*t*(40)=1.93, *p*=.061 [-.01, .43]	t(40)=2.644, *p*=.012 [-.30, -.04]	t(39)=1.35, *p*=.180 [-.38, .08]

Note: Means shown with standard deviation in parentheses. 95% CI for the t-test is in brackets. *indicates significantly different from chance (33%).

License

The MacArthur Inventario del Desarrollo de Habilidades Comunicativas III: A measure of language development in Spanish-speaking two- to four-year-olds

Donna Jackson-Maldonado
Universidad Autónoma de Querétaro

Margaret Friend
San Diego State University

Virginia A. Marchman
Stanford University

Adriana Weisleder
Northwestern University

Alejandra Auza
Hospital General Dr. Manuel Gea Gonzalez

Barbara Conboy
University of Redlands

Marta Rubio-Codina
Inter-American Development Bank*

Philip S. Dale
University of New Mexico

Abstract: Parent report measures are reliable, valid, and cost-effective means for obtaining information about early child language development. Adaptations of the MacArthur-Bates Communicative Development Inventories are available in multiple languages for children below the age of three but there is a need for such measures for older children. This study introduces the Spanish adaptation of the MacArthur-Bates Communicative Development Inventory-III, the MacArthur Inventario del Desarrollo de Habilidades Comunicativas III (IDHC-III) designed for children 2;6 to 4 years of age. This form complements the MacArthur Inventario Del Desarrollo de Habilidades Comunicativas Palabras y Gestos and Palabras y Enunciados (IDHC:PG and IDHC:PE) for younger children. A total of 571 families of monolingual Spanish-speaking children from a diverse socio-economic sample in Mexico completed the IDHC-III and comprise the norming sample. Data are presented by age and maternal education level showing developmental growth curves for *Lista de Vocabulario* (Vocabulary) and *Tipos de Palabras y Oraciones* (Grammatical Complexity) along with norming tables showing variability by age. For the *Pronunciación* (Pronunciation) and *Conceptos Generales* (General Concepts) sections, only descriptive data are presented. We provide a parent report measure to support language assessment for preschoolers acquiring Spanish in Mexico and possibly in other Latin American countries as well.

Keywords: parent report, norms, Spanish.

Corresponding author(s): Margaret Friend, Department of Psychology, San Diego State University, San Diego, CA, US. Email: mfriend@sdsu.edu.

ORCID ID(s): https://orcid.org/0000-0001-9678-0571
https://orcid.org/0000-0002-6839-5163
https://orcid.org/0000-0001-7183-6743
https://orcid.org/0000-0001-6094-8424
https://orcid.org/0000-0002-7592-4625
https://orcid.org/0000-0001-6506-3692
https://orcid.org/0000-0002-1286-7918

**Note*: The views presented here do not represent the Inter-American Development Bank, its board of directors, or the countries it represents.

Citation: Jackson-Maldonado, D., Friend, M., Marchman, V. A., Weisleder, A., Auza, A., Conboy, B., … Dale, P. (2024). The MacArthur Inventario del Desarrollo de Habilidades Comunicativas III. *Language Development Research, 4*(1), 361—398. http://doi.org/10.34842/12fb-qz93

Introduction

Parent report measures are reliable and valid sources of information about young children's language and have proven useful in both clinical and research settings. Caregivers are an invaluable source of information because they have direct and extensive experience with their child in a variety of settings. Information from caregivers is an essential aspect of the process for identifying children with developmental delays or disabilities, as included in U.S. law (IDEA, 2022) and in guidelines from international organizations (WHO, 2012). In addition to their role in clinical assessment and screening, parent report instruments have also been adopted for estimating population-level metrics about the rate of children who are developmentally on track, a need that has grown in response to the United Nation's Sustainable Development Goals McCray et al., 2023). Beyond clinical and educational applications, parent reports have provided critical insights and expanded our knowledge regarding both the consistency and variability that characterizes early child language development (Skarakis-Doyle et al., 2009; Frank, et al., 2021; Fenson et al., 1994).

Two widely used American English parent report measures of language and communication for children up to 30 months of age are the MacArthur-Bates Communicative Development Inventories (Words & Gestures MBCDI:WG, Words & Sentences MBCDI:WS, and the MBCDI-III; Marchman, et al., 2023). The MacArthur Inventarios del Desarrollo de Habilidades Comunicativas (IDHC) are Spanish language adaptations of these measures (Jackson-Maldonado et al., 2003). At the core of these instruments is a vocabulary checklist asking parents to indicate words their child can "understand" or "understand and say", with other sections focused on early gesture use, morphology, word combinations, and sentence complexity. In a recent large, longitudinal study using in a large longitudinal study in Bogota, Colombia, the IDHC:PE predicted both IQ and school achievement (Rubio-Codina & Grantham-McGregor, 2020). Thus, these instruments offer cost-effective means to provide a comprehensive picture of a range of language and communication milestones in children under three years of age, a period when direct testing can be quite challenging.

The present paper reports the development and psychometric properties of an upward extension of the IDHC for measuring language development in children between the ages of 30 and 48 months acquiring Mexican Spanish. Spanish is spoken by substantial numbers of people in more than 20 countries, and is the third most spoken language in the world (Ethnologue, https://www.ethnologue.com/guides/how-many-languages; Simon-Cereijido, Conboy, & Jackson-Maldonado, 2020). In the United States, more than half the population growth between 2000 and 2010 was Hispanic (Passel, Cohen & Lopez, 2011), and approximately 16-18% of the population is Spanish-speaking (Simon-Cereijido, 2015). There is a growing need for valid Spanish-language instruments that can be used with monolingual Spanish speakers in Latin American countries, where there has been an expansion of research, clinical services, and educational programming focused on young children (Minto-Garcia et al., 2019; Rosemberg et al., 2022; Rubio-Codina et al., 2016; Verdisco et al., 2009). There is a particular need for standardized language assessments that can be used with three-year-old Spanish-speaking children across a broad socioeconomic spectrum for several reasons (Rubio-Codina, et al., 2015). First, the availability of standardized language assessment tools for evaluating Spanish-speaking children, particularly those under the age of 4, is limited. Second, this is an important age as it is often when some important educational management decisions are made. Third, parental input has proven effective in providing indirect assessment of the language of younger Spanish-speaking children and is crucial in understanding language skills even beyond the toddler years since it may reflect parent expectations for acquisition that vary with culture (Auza et al., 2023).

Short Forms and Upward Extensions of the CDI Instruments

The original versions of the MBCDI instruments are quite lengthy, with more than several hundred items, and time consuming for caregivers to complete. To overcome these limitations, short form versions have been developed which typically include a short vocabulary checklist (e.g., 100 items) and only one or two additional questions. While short form versions are less comprehensive than the original long forms, their length is likely to increase their feasibility for use in many contexts, for example, enabling face-to-face oral presentation (Rubio-Codina et al., 2016). Although they do not provide comprehensive information of the type needed for studies of vocabulary composition, short forms have strong correlations with longer versions demonstrating their validity as measures of children's relative status (Fenson et al., 2000; Mokhtari et al., 2022; Urm & Tulviste, 2021). This is especially relevant for the development of measures for somewhat older children, whose full range of language skills is growing rapidly and cannot be assessed comprehensively.

The original long- and short forms of the CDIs were designed and normed for use for children under 30 months (Fenson et al., 2007), thus there remained a need for a form developed and normed for older children. Dale and colleagues developed an upward extension of the American English CDI, the MBCDI-III, for use with children through the age of 37 months (Marchman et al., 2023). The American English MBCDI-III has three Sections: Vocabulary Checklist (100 items), Using Language (12 yes-no questions concerning semantics and pragmatics) and Grammatical Complexity (1 question about word combinations and 12 sentence pairs for assessing morphology and syntax). The Vocabulary Checklist is necessarily brief, given the typical size of children's vocabulary at this age (see Marchman et al., 2023 for norms).

Evidence for the concurrent validity of the MBCDI-III has been reported in several studies and is summarized in Marchman et al. (2023). Vocabulary scores correlate with the Bayley Scales of Infant and Toddler Development-III (Perra et al., 2015), the McCarthy Verbal Scale (Feldman et al., 2005), the Peabody Picture Vocabulary Test (Feldman et al., 2005; Mercure, 1999), and Number of Different Words in language samples (Feldman et al., 2005). Grammatical complexity scores are correlated with MLU in language samples (Feldman et al., 2005).

Moreover, the MBCDI-III has been utilized in diverse research with typically developing children. For example, MBCDI-III scores have been used to estimate genetic influence on vocabulary, grammar, and their relationship (e.g., Dale et al., 2015). In addition, the MBCDI-III has been used in studies of children with language disorders (Feldman et al., 2003; Skaradis-Doyle et al., 2009), otitis media (Feldman et al 2003, 2005), Autism Spectrum Disorder (ASD, Tek et al., 2008), and children born preterm (Perra et al., 2015). The MBCDI-III has adequate discriminant classification validity (Skarakis-Doyle et al., 2009; Ukoumunne et al., 2012) and has been used to help parents identify children with language disorders (Skeat et at., 2010). Nevertheless, a ceiling effect was identified that limited the usefulness of the CDI-III to children at or below 37 months, rather than up to 48 months as originally intended.

Adaptations of the MBCDI-III into Non-English Languages

The success of the MBCDI-III has led to the development of adaptations for several other languages. As always in the adaptation of CDI instruments to new languages, substantial linguistic and cultural adaptation is needed (http://mb-cdi.stanford.edu/documents/ adaptationsnottranslations2015.pdf). Languages differ not only in their vocabulary and syntax, but also in the stages of acquisition of culturally relevant words, morphosyntactic forms and functions. Even the acquisition of translation equivalents may not be developmentally equivalent across languages. Consequently, it is important to take into consideration language specific acquisition data when developing measures (Peña, 2007).

Two main categories of adaptations have been developed (see Brieković & Kraljević, 2023, for an overview of most current adaptations). The first category, including e.g., Basque and Hungarian, have been created based fairly directly on the original MBCDI-III, in that the vocabulary list has been very broad with respect to categories. The emphasis has been to find individual words appropriately difficult for the target age range. The Basque adaptation of the MBCDI-III, the KGNZ-3, for example, was extensively adapted and modified to reflect both the cultural context and structure of this non-Indo-European, ergative language with agglutinative morphology (Ezeizabarrena et al., 2013; Barens & García, 2013). Along with changes in the vocabulary list motivated by linguistic and cultural differences, the sections on grammar were expanded to include nominal case inflections, intransitive and transitive auxiliaries, and inflections to determine subject-object relations. New sections were added to assess pronunciation, pre-reading and school abilities, narrative questions and grammatical markers. Many of these changes were motivated by the goal of developing an instrument appropriate for children up through 50 months. Ezeizabarrena et al (2013) reported steady increments through 42 months for vocabulary production and through 50 months of age for sentence complexity with the KGNZ-3. Thus the ceiling effect for vocabulary was partially resolved. Similarly, the Hungarian adaptation (Kas & Lőrik, 2022) showed a ceiling effect for vocabulary at around 39-42 months.

More recent adaptations of the MBCDI-III have generally followed the Swedish adaptation (Eriksson, 2017), which incorporates a different design for the vocabulary list. Here, four specific semantic categories have been selected for more in-depth assessment based on developmental appropriateness and substantial growth during the target age range: food-related words, body-related words, cognitive words and emotion words. The vocabulary section also contains relatively more verbs, adjectives, and adverbs than the Swedish CDIs for younger children. To evaluate morphology and grammar, a section of 10 items asks about the child’s use of complex phrases and another section of eight items queries the child’s use of grammatical markers. A section of seven items on metalinguistic awareness asks caregivers to comment on their child’s phonological and orthographic awareness as well as their awareness of the existence of other languages. Finally, one question asks about whether children pronounce words more like slightly younger children, their peers, or slightly more advanced children. This version has been normed on a nationally representative Swedish sample for children up to 48 months of age. Eriksson (2017) provided an initial evaluation of developmental validity based on correlations with age. As in most MBCDI studies, vocabulary and syntax were correlated. Internal consistency was high for vocabulary and syntax, and somewhat lower for the other components.

The Estonian MBCDI-III (the ECDI-III; Tulviste & Schultz, 2020) was also based to a considerable extent on the Swedish version. The vocabulary section consists of 101 words, mostly verbs and adjectives, with similar categories to the Swedish forms (body words, food words, mental and emotion words). The grammatical complexity section has 10 sentence pairs focusing on the agglutinative nature of the language. There are also sections which assess metalinguistic and general concepts, and pronunciation. Pilot data on the validity of the ECDI-III have been reported for children at 3 years, a sub-sample of the full longitudinal normative sample ranging from 30-48 months, based on correlations with Reynell Developmental Language Scale (Edwards et al., 2011).

Other adaptations which have generally followed the Swedish model, both with respect to the structure of the vocabulary checklist and the incorporation of scales for aspects of language beyond vocabulary and grammar are those for Norwegian (Holm et al., 2023), Finnish (Stolt, 2023), European Portuguese (Cadime et al., 2021), and Croatian (Brieković & Kraljević, 2023). Although the existing reports differ with respect to design, age range, and validation measures, overall the results are positive and similar to those for Swedish.

Evolution of the MBCDI-III in Spanish

Two preliminary Spanish parent report instruments for three-year-olds similar to the English-language MBCDI-III have been developed: the Pilot Inventario-III (INV-III; Guiberson, 2008 1 & b; Guiberson and Rodriguez, 2010, 2014; Guiberson, et al., 2011; and the Spanish Vocabulary Extension (SVE; Mancilla-Martinez, et al., 2016, Mancilla-Martínez, et al., 2011; Mancilla-Martínez et al., 2013). The INV-III is a direct translation of the English MBCDI-III; it includes a vocabulary checklist, a grammatical complexity section, and a request to provide examples of their child's three longest utterances. Scores on the INV-III correlate with the Ages and Stages Questionnaire (ASQ; Squires & Bricker, 2009) ($r_{(46)} = .69$, $p = .01$) and the Preschool Language Scale-4 (PLS-4; Zimmerman, et al., 2002) ($r_{(46)} = .62$, $p = .01$) and there is good classification accuracy of children with and without language delays (sensitivity = .82 and specificity = .81). However, because participants varied in age, the correlations with other measures are likely somewhat inflated. Also, data on this instrument for monolingual or near-monolingual children are limited.

The Spanish Vocabulary Extension (SVE) (Mancilla-Martínez et al., 2016) consists of a 100-word vocabulary checklist, drawn from the IDHC-PE and spontaneous languages samples. Correlations with the Short Form of the Spanish IDHC-PE (IDHC-IISF) for lower-income Spanish-speaking bilingual children in the U.S (N=48) suggest concurrent and discriminant validity and SVE scores also correlate with the full IDHC-IISF, the Woodcock Language Proficiency Battery–Revised (WLPB-R; Woodcock & Muñoz-

Sandoval, 1995), the Picture Vocabulary subtest, and the Test de Vocabulario en Imágenes Peabody (TVIP; Dunn, et al., 1986). The correlations were strongest for the younger children, and for the WLPB.

Although these measures filled a need in language assessment for Spanish-speaking three-year-olds, both have significant limitations. The INV-III is a direct translation of the English form rather than an adaptation, limiting its validity due to the lack of consideration of the cultural and linguistic relevance of specific words and sentence structures. The SVE was developed for a specific project and consists of a word list only. Norms were not obtained for either measure. Therefore, the present research sought to develop a Spanish MBCDI-III with indicators of both vocabulary and grammar, culturally and linguistically relevant items, and norms derived on a monolingual Spanish-speaking sample. Both the INV-III and the SVE were considered, and their authors contacted, before developing the current Spanish IDHC-III.

The Development of the MacArthur Inventario del Desarrollo de Habilidades Comunicativas III (IDHC-III)

The development of the IDHC-III has drawn on the Basque, Estonian, and Swedish adaptations. The development process followed the process for the original IDHC:PG and IDHC:PE adaptations (Palabras y Gestos and Palabras y Enunciados; Jackson-Maldonado et al, 2003) and the Spanish Short Forms (IDHC-ISF & IDHC-IISF; Jackson-Maldonado et al., 2012), with careful consideration of cultural and linguistic relevance and inspection of Spanish language acquisition data. The process of development of this form consisted of a preliminary norming study and this final version.

Pilot Instrument Development

For both Vocabulary and Grammatical Complexity, item selection began with examination of results at 30 months on the IDHC-PE norming study, to identify items selected by less than 30% of parents. This did not yield enough advanced vocabulary items, so additional words were identified by several other means, including narrative language samples from Mexican children (Jackson-Maldonado & Maldonado, 2015) and two Spanish-language corpora of 3- and 4-year-old children from CHILDES (Diez-Itza Corpus, Diez-Itza et al., 1999). Spanish-language acquisition researchers reviewed the preliminary list and the developers of the English form were consulted.

The *Tipos de Palabras y Oraciones* (Grammatical Complexity) section was expanded to increase the ceiling from the IDHC-II for younger children. On these items, parents are asked to identify which sentence of two examples “sounds most like how your child speaks”. Each example sentence captures the same basic meaning, but one sentence is morphosyntactically more complex. New phrases were constructed from

CHILDES Spanish databases and narrative samples of preschool monolingual Spanish-speakers and Spanish language acquisition studies (Fernández & Aguado, 2007; Jackson-Maldonado & Maldonado, 2015, 2016; Jackson-Maldonado & Conboy, 2007; Morgan, et al., 2009; Perez-Leroux et al., 2012; Sanz-Torrent et al., 2008; Uccelli, 2009; Uccelli & Pavez, 2007). Note that, in most cases, these forms also convey more semantic information, as is common when children begin to use more complex sentence structures.

Preliminary norms were developed using a 100-word list and 26 sentence pairs completed by caregivers for 579 middle and low SES children in Mexico and 640 low SES children in Colombia between the ages of 30 and 47 months. Validity studies compared scores to the INV-III (Guiberson 2008a & b; Guiberson & Rodríguez, 2010; Guiberson et al., 2011) and sub-sections of the Clinical Evaluation of Language Fundamentals-Preschool (CELF-P; Wiig, Secord, Semel, 2004). Significant positive correlations (.54 in both cases) were found between corresponding sections (Vocabulary or Grammatical Complexity) of the IDHC-III and the INV-III. A discriminant analysis of children with varied language disorders yielded moderate sensitivity, 75%, and high specificity, 92%.

However, developmental trends were limited and there were ceiling effects with most participants producing 50 or more words on the 100-item test. Further, several words were identified as extremely low or extremely high frequency, all indicating a need to revise the vocabulary list. In contrast, the complexity section evinced an expected linear increase with age.

Current IDHC-III

Based on these considerations, the current IDHC-III was developed. Following Eriksson (2017), we included a more focused vocabulary list to include more advanced word classes in food related, body related, cognitive and emotional words categories. Further, pilot data on 108 Guatemalan children from low SES backgrounds, half of whom were monolingual Spanish-speaking and half of whom were Spanish-dominant from Kaqchikel-Spanish bilingual homes (Conboy et al., 2017a & 2017b), motivated the inclusion of additional culturally relevant categories (e.g., nature, health, school, abstract nouns–including culture-religion, action specific verbs, and change of state verbs) to allow the inventory to be used with children from a wider range of cultural backgrounds. A new 140-word list was piloted with 45 participants. Extremely high and extremely low frequency words and words with low correlations with age were deleted to obtain the final 100-word list reported here.

The final word categories, number of items and examples are presented in Table 1 and the examples of the grammatical complexity items are shown in Table 2. The full form is presented in Appendix A. As can be seen from Table 1, the vocabulary list can be viewed as intermediate between that of the original MBCDI-III and that of the Swedish model. It is more focused than the original MBCDI-III, but is not as focused as the Swedish model.

Two additional sections are included. *Pronunciación* (Pronunciation), as in the Swedish version, consists of one question about how the child pronounces words. In *Conceptos Generales* (General Concepts), what Eriksson (2017) called metalinguistic awareness, parents are asked about school concepts, specifically, writing letters or numbers, counting, and naming shapes; the wording is based on the preschool academic programs for public schools in Mexico (https://www.gob.mx/sep/acciones-y-programas/educacion-preescolar, SEP 2017-ref) and consultation with preschool teachers and early literacy specialists.

Table 1. ***Word categories for the Vocabulary Checklist***

Category	Number of Items	Example
Abstract Noun	8	Accidente -*accident*
Attributes	13	Envidioso- *envious*
Action	11	Aguantar -*stand it, hold out*
Body	5	Cachete -*cheek*
Change of state	12	Desaparecer -*disappear*
Food	1	Postre- *dessert*
Function wds	7	Desde -*from*
Health	6	Calentura -*fever*
Objects	4	Grúa -*tow truck*
Locatives	8	Ciudad- city
Outside-nature	9	Insecto -*insect*
People	3	Mecánico -*mechanic*
Quantifier	10	Cada -*every*
School	3	Cuadrado -*square*
Total	100	

Table 2. ***Example sentence pairs for Grammatical Complexity***

Sentence Pair	Translation
Como pollo Voy a comer pollo con el tenedor	(I) eat chicken (I) am going to eat chicken with the fork
Ma caí y me duele Cuando me caigo, me duele	I fell and it hurts When I fall, it hurts
Se enfermó No pudo porque se enfermó	They got sick They couldn’t because they got sick
No lo pongo aquí No creo que pueda ponerlo	I don´t put it here I don´t think I can put it

The Current Study

The goals of this study are to: (1) present developmental norms for vocabulary and grammatical complexity on the newly developed Spanish IDHC-III for monolingual, Spanish-speaking children in Mexico; (2) compare vocabulary and complexity development in children from different socioeconomic backgrounds; and (3) determine the relation between vocabulary and complexity on this instrument. Based on previous findings, we expect a strong relationship between vocabulary and complexity. We also expect that there will be developmental change with age and variation in scores as a function of maternal education.

Method

Participants

Data were originally compiled from *n* = 577 caregivers across multiple data collection sites. A total of 6 children were excluded because they were older than the target age range when the forms were completed. The final sample consisted of *n* = 571 caregivers, mostly mothers, who completed the IDHC-III and had children between

30 and 48 months of age (290 M, 281 F). Vocabulary checklist data were available for the full sample, however, not all sites chose to administer the form in its entirety and so data were available for only a subset of the children for *Grammatical Complexity* (*n* = 502) and *Pronunciación* and *Conceptos Generales* (*n* = 542).

Participants were recruited by multiple means to ensure a diverse sociodemographic sample from urban and rural areas of central Mexico. Caregivers were contacted through day care centers, preschools, recreation centers and personal contacts. An additional sample was also obtained as part of the piloting of the first child development module of the 2018-19 Mexican National Health and Nutrition Survey (ENSANUT 2018-19). This subset of caregivers, all beneficiaries of the government Conditional Cash Transfer program *Prospera,* were invited to attend a special session in which the goals of the project were explained, and they were offered a nonobligatory opportunity to fill out the forms with the interviewers. All caregivers completed a consent document fulfilling the first author's university Bio-Ethics committee requirements prior to the study. Caregivers then completed a Basic Information Questionnaire that included questions about the child's gestational age, birth weight, health issues, languages spoken in the home, as well as each caregivers' education and occupation.

For descriptive purposes, participants were divided into six age groups: 30-32 months, 33-35 months, 36-38 months, 39-41 months, 42-44 months, and 45-48 months. Participants were also divided into groups based on maternal education level: Middle School or less (MS), some High School (SHS), Completed High School (HS), and More than High School (MHS). The sample is described by age, child sex, and maternal education in Table 3. The sample is relatively evenly distributed over age, with a balance of females vs. males in each age group. Levels of maternal education were not evenly distributed, as the majority of the sample consisted of caregivers in the lower two groups. Just under 1/3 of the sample had more than a high school education across all age groups. This sample consists of a large and relatively representative sample of the Mexican population, as determined by educational attainment (OECD, 2023).

Procedure

Caregivers completed the forms following two administration formats. Some caregivers filled out in person with the help of linguistics and psychology students and teachers. This method was used most often at day care centers or the government health facility where *Prospera* program activities were carried out. Other caregivers received the forms in person. Parents could complete the forms on site, or if desired, they could take them home, and the forms were picked up no longer than 2 weeks later.

Written instructions appear at the beginning of each section, but to ensure understanding, full instructions with examples were always first explained verbally.

Table 3. ***Number of participants (%) by child age, child sex, and level of maternal education in full sample (n = 571)***

				Level of Maternal Education			
Age Group	Total	Female	Male	Middle School or less	Some High School	High School Graduate	More than High School
30-32 mos	96	46 (47.9)	50 (52.1)	10 (10.4)	46 (47.9)	11 (11.5)	29 (30.2)
33-35 mos	90	46 (51.1)	44 (48.9)	12 (133)	34 (37.8)	10 (11.1)	34 (37.8)
36-38 mos	103	54 (52.4)	49 (47.6)	4 (3.9)	52 (50.5)	11 (10.7)	36 (35.0)
39-41 mos	95	45 (47.4)	50 (52.6)	9 (9.5)	49 (51.6)	9 (9.5)	28 (29.5)
42-44 mos	82	38 (46.3)	44 (53.7)	6 (7.3)	41 (50.0)	12 (14.6)	23 (28.0)
45-48 mos	105	52 (49.5)	53 (50.5)	14 (13.3)	47 (44.9)	16 (15.2)	28 (26.7)
TOTAL	571	281 (49.2)	290 (50.8)	55 (9.6)	269 (47.1)	69 (12.1)	178 (31.2)

Measures

In the Lista de Vocabulario (Vocabulary) section, parents are asked to indicate the words that their child “comprende y dice” ‘understands and says’, yielding a maximum production vocabulary score of 100 words. Caregivers are told that the child should be able to produce the word spontaneously (i.e., repetitions are not allowed), but the words can be pronounced in a “childlike” manner (e.g., *momingo* for “domingo” ‘Sunday’). Similarly, the child may use a different grammatical form that is equivalent to the one listed on the form. For example, if a child says *sabo* for “saber,” an overgeneralization of the regular first-person singular form applied to the irregular verb ‘to know’, or *pesada* for “pesado,” a feminine-marked form for the adjective ‘heavy,’ these would be counted as correct. The caregiver may also substitute synonyms for words that are used in their own family or dialect (e.g., *cavar* instead of *excavar* or *bonito* instead of *hermoso*).

The *Tipos de Palabras y Oraciones* (Word and Sentence Types or, as we refer to it in this paper, Grammatical Complexity) section is used to evaluate emerging morphology and grammar. On each item, the caregiver is asked to indicate which sentence, of each pair, *sounds most* like how their child currently speaks. They are told that the child does not have to produce the same sentence exactly, but rather the caregiver should reflect on which sentence sounds the most like something their child might say. Thus, in the pair, "Me caí y me duele" / "Cuando me caigo, me duele" *'I fell and it hurts'* / *'When I fall, it hurts,'* the child is given a score of 0 if the parent chooses the first phrase and a score of 1 if the parent chooses the second phrase. The maximum score is 15, reflecting the number of times the parent chose the second, more complex, sentence in the pair across 15 items.

For the *Pronunciación* section, caregivers were asked if it was difficult to understand their child´s speech. The score reflects *difficulty of understanding,* that is, when caregivers answered "sí (yes)", a score of 1 was recorded. This was the only item in which a higher score was indicative of less sophisticated development and a lower score was indicative of more advanced development. For *Conceptos Generales,* each question about academic concepts received a score of 1 if the parent answered "sí (yes)." Scores were summed to provide a total score (out of 3 possible responses). Scoring instructions for each section can be found in Appendix B.

Analysis Plan.

We first report descriptive statistics for vocabulary and grammatical complexity for children in each age group. We next report developmental patterns using modeling techniques that allow us to estimate age-related changes for each measure at monthly intervals. We chose to apply generalized additive models in the beta distribution family using non-parametric monotonic P-splines with GAMLSS (Stasinopoulos et al., 2017) in the R statistical package (Version 4.0.3; R_Core_Team, 2020). GAMLSS is a general framework for modeling a range of functions within a regression framework. This approach was recently applied in the 3rd Edition of the American English norms (Marchman et al., 2023) and has advantages over other techniques because it allows fit to a range of possible functions, and provides fit estimates of standard deviation, as well as the central tendency. Based on earlier work (Fenson et al., 2007; Frank et al., 2021), we assumed that the distributions were best captured within the family of Beta distributions, i.e., limited by 0 and 1. We modeled age and child sex, as well as interactions between age and sex as fixed effects. Following Smithson and Verkuilen (2006), scores were first converted to a proportion out of possible responses and extreme scores were imputed as 0.001 and 0.999 so that we could include all observations in the models. Based on our expectations of developmental change, all models used a very high value of lambda (10^4) and set the number of knots at 20. These parameters resulted in estimates of development that approached linear and were smooth over age, while nevertheless being constrained at the higher and lower values. We report unstandardized beta coefficients (B) for all fixed effects. To test the significance of goodness of fit between nested models (e.g., those with and without an interaction term), we applied likelihood ratio tests (LRT, df = 1). Alpha was set at p < .05, two tailed, for all analyses. To generate the normative values, we extracted the percentile ranks for developmental trajectories estimated in the GAMLSS models by 5-percentile intervals from the 5th – 95th percentiles and the 99th percentile over age in months. Plots present the values for the quintiles (10th, 25th, 50th, 75th and 90th). Normative values were also generated separately by child sex.

We next conducted several exploratory analyses examining the intercorrelation between scores on the vocabulary and grammatical complexity sections. We anticipated that the indices would be highly correlated, consistent with earlier reports (Bates & Goodman, 1994). To analyze effects of socioeconomic status, we again modeled age-related changes introducing maternal education level as a potential moderator for each measure. Finally, we present descriptive statistics for the remaining measures of *Pronunciación,* and *Conceptos Generale*s in Table 7; however, no further analyses of those measures are presented here.

Results

Descriptives

Mean scores and standard deviations on the Vocabulary and Grammatical Complexity sections are presented in Table 4 by child age group and sex. For vocabulary, note that even the children in the youngest group were reported to know just under half of the items on the form. For complexity, the children in the youngest group were reported to say the second, more complex, example, only about 1/5th of the time. For both measures, there were substantial increases over age group in children's performance suggesting developmental changes in these critical language abilities over this important period.

Age-related trends.

Vocabulary Size.

To explore these developmental effects more fully, we conducted models that allowed us to capture age-related changes in vocabulary score, as shown in Table 5. As expected, Model 1, the unconditional model, shows a significant main effect of age, reflecting developmental change in vocabulary score from 30 to 48 months. Model 2 adds the factor of child sex. Again, results revealed a significant main effect of age, but the main effect of child sex was not statistically significant, $p = 0.35$. Thus, unlike previous studies of vocabulary development using parent report (Frank et al., 2021; Marchman et al., 2023), the evidence for sex-related differences in vocabulary size was not statistically reliable. Moreover, adding the interaction term in Model 3 did not significantly increase overall model fit, *LRT*(1) = 0.22, *p* = 0.64, suggesting no differences in the magnitude of any sex differences across the age period.

Figure 1 illustrates the developmental effects from the unconditional model in terms of the fitted quantile estimates for all children in the sample and Figures 2 and 3 for boys and girls separately. Full values for percentile levels for both vocabulary and complexity scores, in 5-percentile increments, are presented in Tables 1 – 3 in Appendix C for all children and for girls and boys separately. Even though the main effect of sex and the sex by age interaction terms were not statistically reliable, we nevertheless provide norming tables separately for girls and boys to be consistent with earlier studies and to conform with some requirements for clinical reporting.

Table 4. Means and (SD) of scores for Vocabulary (n = 571) and Grammatical Complexity by age group for all children and by child sex

	Vocabulary[a]			Grammatical Complexity[b]		
Age Group	All	Female	Male	All	Female	Male
30-32 mos	46.2 (23.8)	47.8 (21.8)	44.8 (25.7)	3.2 (3.8)	3.2 (3.7)	3.2 (3.9)
33-35 mos	49.7 (22.6)	47.7 (23.4)	51.8 (21.9)	3.7 (3.6)	4.2 (3.6)	3.3 (3.5)
36-38 mos	56.8 (21.4)	63.1 (21.1)	49.8 (19.7)	5.5 (4.2)	5.7 (4.6)	5.2 (3.7)
39-41 mos	64.7 (21.7)	64.1 (23.1)	65.3 (20.5)	6.2 (4.1)	5.6 (4.0)	6.7 (4.2)
42-44 mos	64.4 (23.3)	62.2 (24.8)	66.3 (22.0)	6.5 (3.8)	6.1 (3.8)	6.9 (3.9)
45-48 mos	67.4 (20.9)	68.2 (21.8)	66.6 (20.4)	6.5 (4.0)	6.2 (4.0)	7.0 (4.0)
All children	58.3 (23.6)	59.1 (23.8)	57.5 (23.4)	5.3 (4.2)	5.2 (4.1)	5.4 (4.2)

Note: [a]Production vocabulary reflects the number of items caregivers selected on the vocabulary checklist (max = 100); [b]Grammatical complexity reflects the number of times the parent chose the more complex answer of two choices (max = 15).

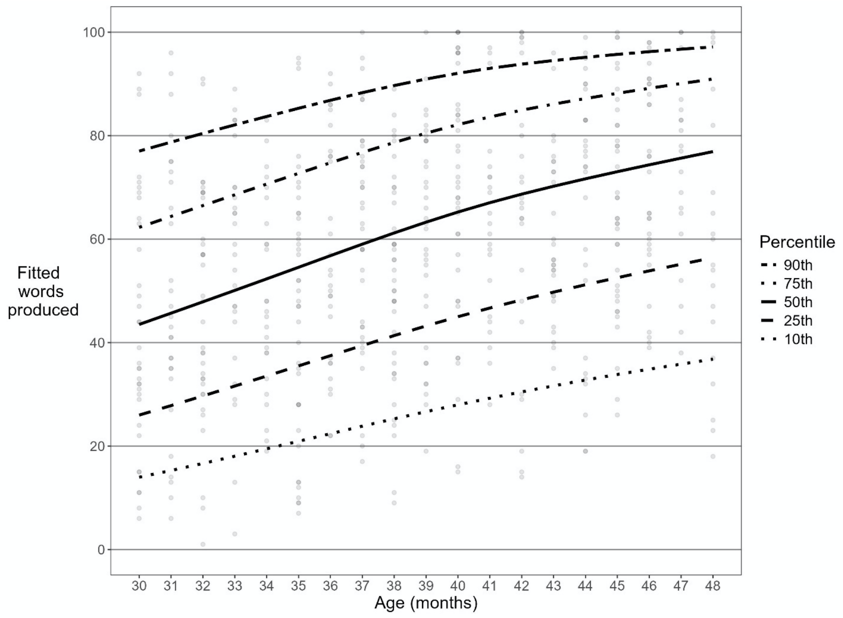


Figure 1. Fitted percentile scores by quintile (10th, 25th, 50th, 75th, and 90th) for Total Words Produced as a function of age group (months), both sexes combined; dots represent individual data points (n = 571).

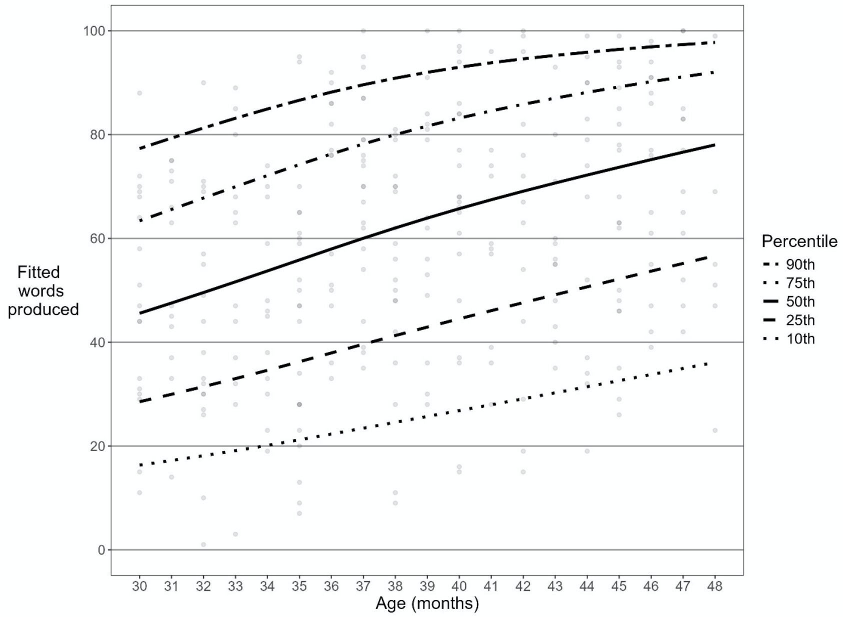


Figure 2. ***Fitted percentile scores by quintile (10th, 25th, 50th, 75th, and 90th) for vocabulary production as a function of age group (months) – girls; dots represent individual data points (n = 281).***

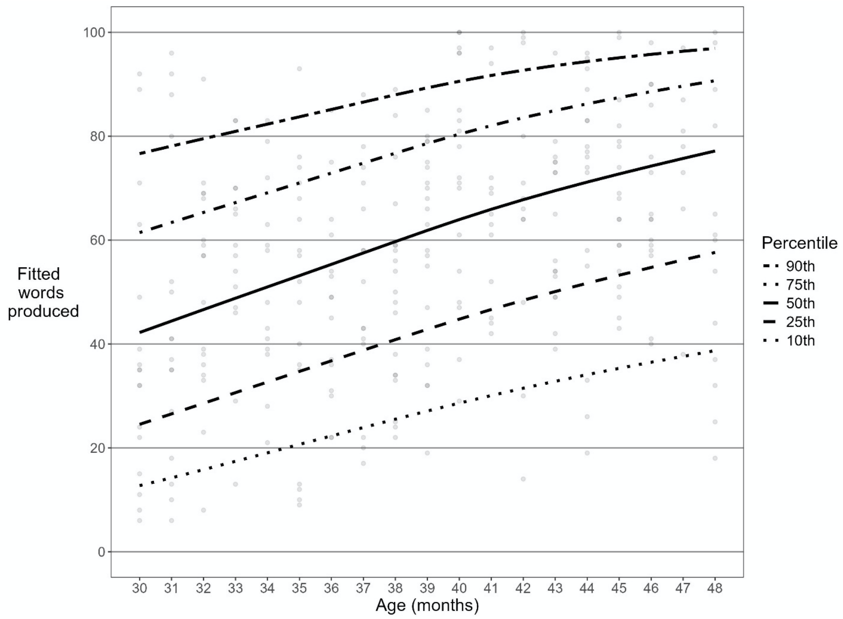


Figure 3. ***Fitted percentile scores by quintile (10th, 25th, 50th, 75th, and 90th) for vocabulary production as a function of age group (months) – boys; dots represent individual data points (n = 290).***

Grammatical Complexity.

Results for models exploring developmental trends in grammatical complexity score are presented in Table 6. Model 4 again revealed a substantial main effect of age on children's scores. Model 3 also revealed a main effect for age, but no main effect of child sex. Adding the interaction term did not increase overall model fit, LRT(1) = 1.5, p = 0.22, again suggesting no advantages for girls over boys in grammatical complexity scores at any developmental level. See Figure 4 for developmental effects from the unconditional model in terms of the fitted quantile estimates for all children in the sample and Figures 5 and 6 for boys and girls separately. Full values for all percentile levels are presented for all children and for girls and boys separately in Tables 4 – 6 in Appendix C.

Interrelation between Vocabulary and Grammatical Complexity.

We next explored the association between scores on the vocabulary and grammatical complexity subsections. Scores on the Vocabulary and the Grammatical Complexity sections were moderately intercorrelated ($r(500) = 0.43$, $p < 0.001$), reflecting that children who scored higher in vocabulary were also scoring higher on the grammatical complexity scale. This correlation remained significant after controlling for age,

$r(499) = 0.37$, $p < 0.001$, suggesting that this association is not due to each measure being individually associated with age.

Maternal Education.

To explore the impact of maternal education on patterns of age-related changes, we added maternal education, as well as the age x maternal education interaction to Models 1 and 3. Looking first at vocabulary production, we again see a significant main effect of age, $B = 0.07$ (0.01), $p < .001$, however, there were no effects of maternal education, such that scores were similar across all 4 groups, as illustrated in Figure 7. Note that children in families with more than high school education had the lowest scores overall when compared to children with less than high school education, $B = -0.29$ (0.16), $p = .07$. This difference did not reach statistical significance and must therefore, be interpreted with caution. Finally, the addition of the interaction term did not increase overall model fit, $LRT(3) = 1.7$, $p = 0.65$, suggesting that the patterns of relations among the caregiver education groups were parallel across age.

To examine whether this pattern was consistent in those sub-samples of families in which the caregivers received additional support in completing the forms, we reanalyzed the effect of maternal education in only those families in which caregivers were not likely to have been given verbal support during administration ($n = 502$). Again, there were no statistically significant group differences on vocabulary scores as a function of maternal education group (all $p > .08$), and adding maternal education to the model did not significantly increase overall model fit, $LRT(1) = 0.96$, $p = .33$.

For Grammatical Complexity, adding maternal education level to Model 4 showed a significant effect of age, $B = 0.07$ (0.01), $p < 0.001$. Importantly, in contrast to the results for vocabulary, children of caregivers with higher education levels were reported to produce more complex sentences than children of caregivers with less than middle school education, $B = 0.66$ (0.19), $p < .001$, as illustrated in Figure 8. Scores for children in all of the other groups were not statistically different from those of children who had caregivers with less than middle school education. Adding the interacttion term did not increase overall model fit, $LRT(4) = 2.83$, $p = 0.63$, suggesting that the advantage for children of caregivers with higher educational levels was consistent across the age period.

Table 5. ***Fitted estimates (unstandardized B (SE)) for vocabulary production by age and child sex, both sexes combined (n = 571)***

	Model 1	Model 2	Model 3
Intercept	-2.13 (0.30)***	-2.10 (0.30)***	-2.01 (0.43)***
Age	0.06 (0.01)***	0.07 (0.01)***	0.06 (0.01)***
Sigma Intercept	-0.46 (0.29)	-0.46 (0.29)	-0.46 (0.29)
Sigma Age	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)
Sex	--	-0.08 (0.08)	-0.22 (0.59)
Age x Sex	--	--	0.01 (0.02)
Number of Observations	571	571	571
R^2	0.12	0.12	0.12
Generalized AIC	-158.22	-157.98	-156.04

Table 6. ***Fitted estimates (unstandardized B (SE)) for grammatical complexity by age and child sex, both sexes combined (n = 502)***

	Model 4	Model 5	Model 6
Intercept	-2.84 (0.43)***	-3.02 (0.43)***	-2.57 (0.57)***
Age	0.06 (0.01)***	0.06 (0.01)***	0.05 (0.01)***
Sigma Intercept	0.59 (0.33)	0.56 (0.33)	0.56 (0.33)
Sigma Age	-0.01 (0.01)	-0.01 (0.01)	0.01 (0.01)
Sex	--	-0.07 (0.10)	-0.87 (0.77)
Age x Sex	--	--	0.02 (0.02)
Number of Observations	502	502	502
R^2	0.10	0.10	0.10
Generalized AIC	-268.02	-266.51	-266.01

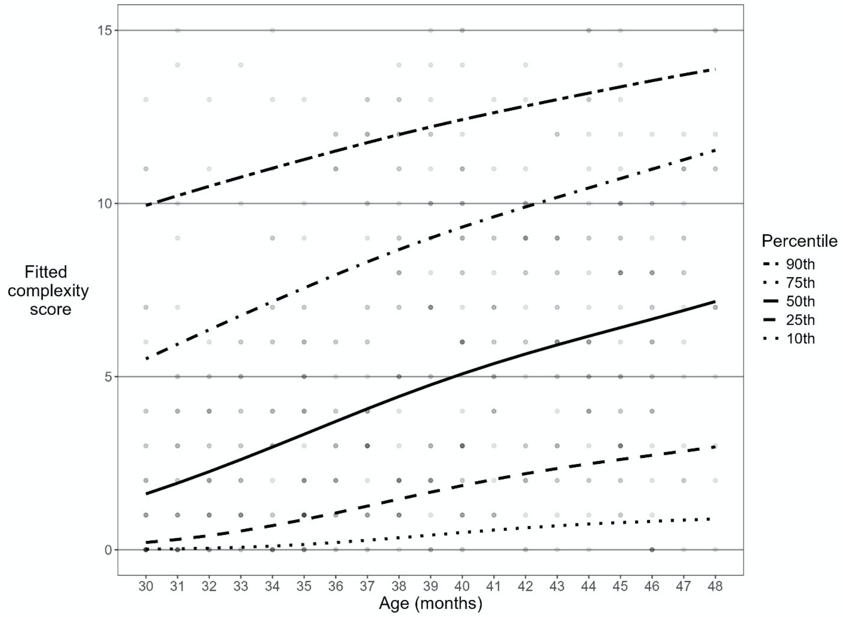


Figure 4. Fitted percentile scores by quintile (10th, 25th, 50th, 75th, and 90th) for grammatical complexity as a function of age group (months) – both sexes combined; dots represent individual data points (n = 502).

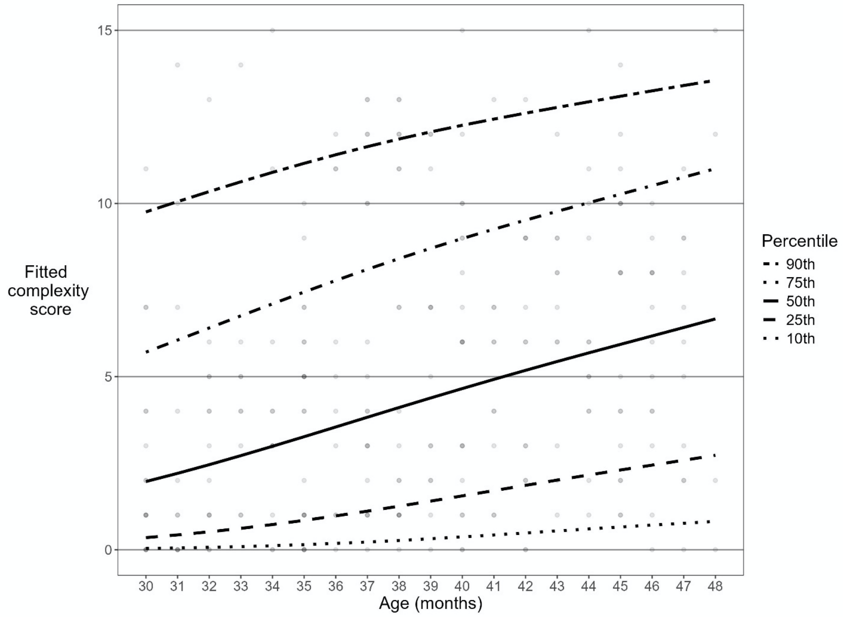


Figure 5. Fitted percentile scores by quintile (10th, 25th, 50th, 75th, and 90th) for grammatical complexity as a function of age group (months) – girls; dots represent individual data points (n = 251).

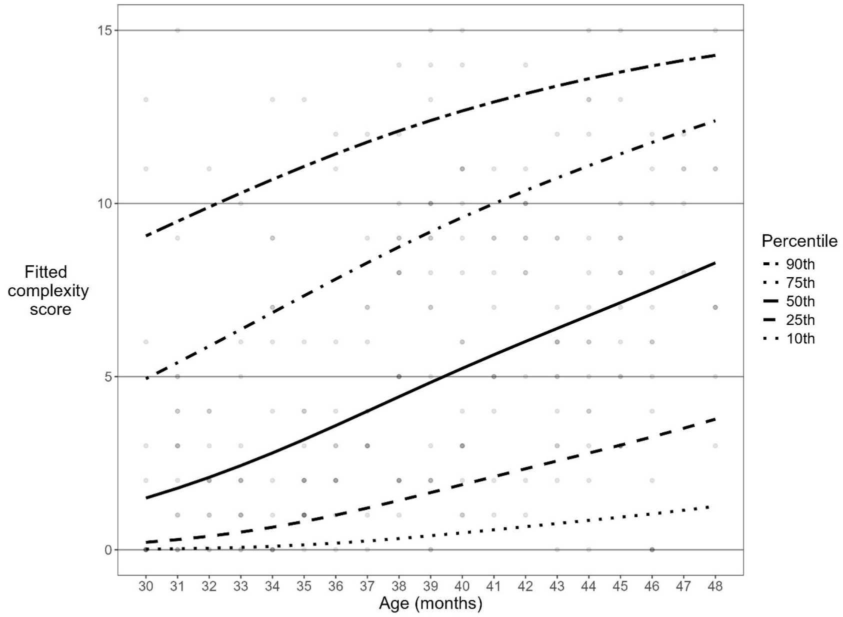


Figure 6. Fitted percentile scores by quintile (10th, 25th, 50th, 75th, and 90th) for grammatical complexity as a function of age group (months) – boys; dots represent individual data points (n = 251).

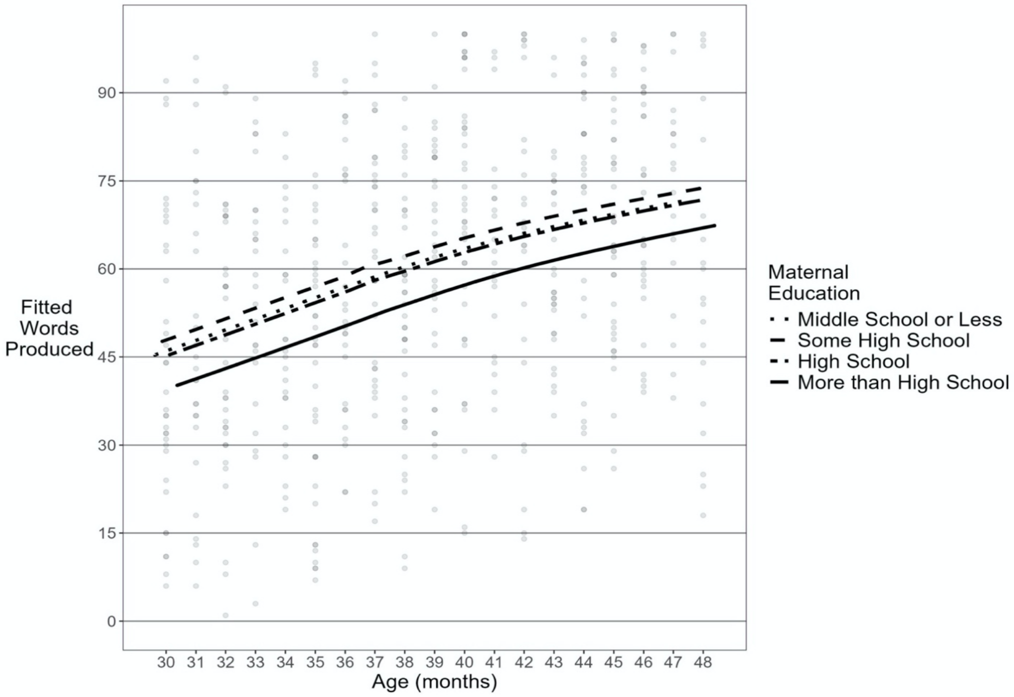


Figure 7. Modeled estimates for words produced as a function of child age and maternal education level; dots represent individual data points (n = 571).

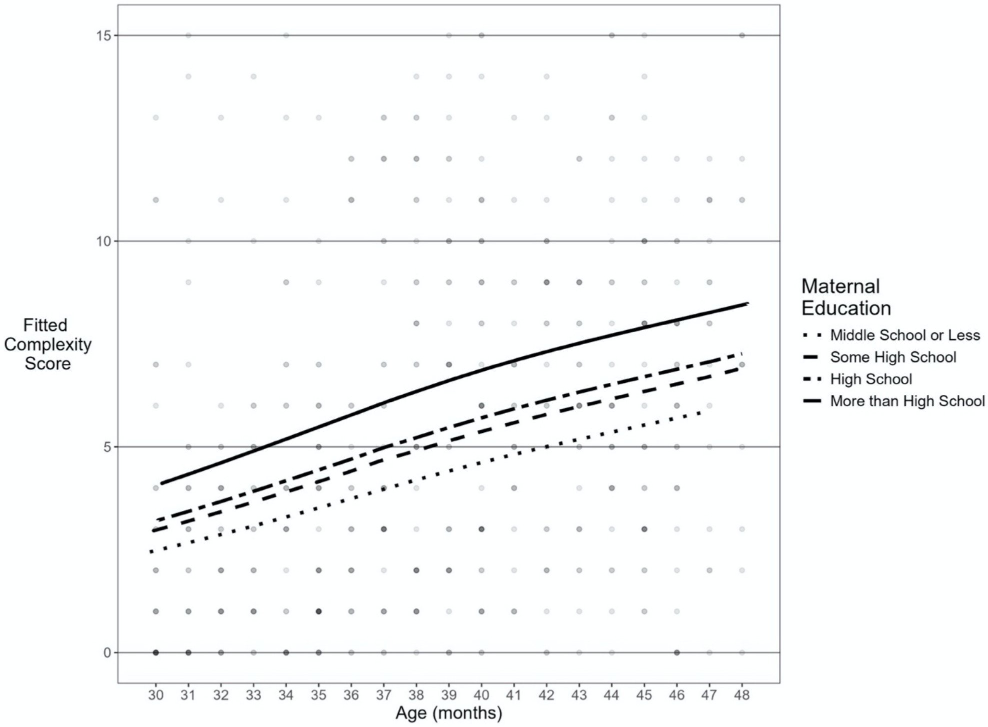


Figure 8. Modeled estimates for grammatical complexity as a function of child age and maternal education level; dots represent individual data points (*n* = 502).

Discussion

This paper presents initial norming data for Spanish adaptation of the upward extension of the MBCDIs (the IDHC-III), for children 2.5 through 4 years of age. This adaptation consists of a 100-word checklist for word production and a complexity section consisting of 15 sentence pairs to identify word-level and sentence-level grammatical complexity. The other sections of the IDHC-III are not analyzed in depth here. For the Vocabulary and Grammatical Complexity sections, the results included developmental trends by age groups and present differences by maternal education. Importantly, the data set includes a large and relatively representative sample of the population of Spanish-speaking, Mexican population (OECD, 2023).

Our analytic models for vocabulary production revealed developmental changes across age groups, but also substantial individual variation from 30 months through 4 years of age. This checklist, like the IDHC Short Forms, was only 100 words long and yet captured a wide range of levels of vocabulary knowledge. The results for the Grammatical Complexity section also revealed steep age-related changes as well as substantial variation, consistent with findings from the Basque form (Ezeizabarrena et al., 2013; Barnes & Garcia, 2013). The strength of the relation between Vocabulary and Grammatical Complexity was similar to what has been reported for younger children (e.g., Frank et al., 2021), which may suggest that these two abilities are driven by a common set of learning mechanisms.

In addition, we observed no reliable effects of maternal education on vocabulary production scores. Prior studies with English-speaking children have found that caregivers with lower levels of education report higher vocabulary comprehension scores in very young children (8-12 months) but not older children (13-18 months, e.g., Dollaghan et al., 1999). In older English-speaking (21-30 months; Fenson et al., 2007), and Spanish-speaking children (26-30 months; Jackson-Maldonado et al., 2003) a positive relation between SES and vocabulary production has been reported and several studies show a positive correlation between maternal education and child vocabulary through the preschool years (e.g., Hoff, 2006; 2013). At the same time, other studies report no relation between maternal education and vocabulary in monolingual and bilingual (DeAnda et al., 2015; Friend et al., 2022; Montanari et al., 2020) Spanish-speaking children. Still others report that the relation is mediated by parent literacy behaviors (e.g., Gonzalez et al., 2017).

One possibility is that differences in administration methods for caregivers with varying education levels may have masked our ability to detect any effects of maternal education on vocabulary production in the current study. In accordance with recommendations for low literacy contexts (Alcock et al., 2015), some caregivers with lower maternal education received assistance from a researcher when completing the IDHC-III forms, while caregivers with higher maternal education typically completed the forms independently. This aligns with previous pilot research on the IDHC-III (Conboy et al., 2017b; Conboy, 2019) conducted in Guatemala and Mexico. This format may draw parent attention to how children use words in a way that written checklists do not, facilitating attention to the content of the vocabulary checklist among caregivers with lower levels of education. Indeed, at the ages covered by the IDHC-III, vocabulary can be quite large, and the great majority of words will have low frequency. Based on the enormous growth of vocabulary during the previous year or two in preschoolers, deciding if a given word has been produced by a child is a challenging task. There are two possible unintended consequences of providing additional support to caregivers with lower education levels. Caregivers completing the inventory on their own may not have attended to the checklist in the same way and underreported the words their children knew or, on the other hand, mothers who received additional support may have overreported their children's vocabulary. In an effort to further understand this finding, we reanalyzed the subset of the data in which no systematic assistance in completing the forms was offered to caregivers with low education. Patterns mirrored the effects for the full dataset; administration differences within sub-groups do not appear to have masked effects of maternal education on vocabulary production in this study.

In contrast to the results with Vocabulary, Grammatical Complexity scores showed the expected tendency: caregivers with higher education levels reported that their child produced more complex forms than those with lower education levels. These results are generally consistent with those reported in the literature. Based on a series of studies including older children, Hoff (2013) reported that higher SES children outperform lower SES peers on most tests of grammatical development and produce more complex sentences with a larger variety of structures (Dollaghan et al., 1999; Huttenlocher et al., 2010; Vasilyeva, Waterfall, & Huttenlocher, 2008). In the norming studies for the original English and Spanish long form instruments, caregivers of older children (21-30 months) from mid-SES families report longer mean length of utterance and higher grammatical complexity scores than caregivers from lower-SES families (Fenson et al., 2007; Jackson-Maldonado et al., 2003).

It is important to speak to the appropriateness of parent report in low SES samples. Without question, adapting methods for socio-cultural context is paramount to acquiring valid and reliable indicators of language development. For example, parents with lower levels of education and literacy may need additional support to complete the forms (e.g., Rubio-Codina et al., 2016) or may have different values and motivations than middle class families (Roseberry-McKibbin, 2013; Gonzalez et al., 2018) that may be reflected in their responses. Nevertheless, parental report measures have been used successfully in diverse socio-cultural contexts. For example, Weber et al. (2018) has shown the validity of parental report measures in Wolof communities in Africa, and Alcock et al. (2015) have noted that traditional written formats may need to be modified for face-to-face interviews in Kenya (Alcock et al., 2015). Most studies have shown that low SES parents are valid reporters (Alcock et al 2015, Dar et al, 2015; DeAnda et al., 2015; Hamadani et al 2010; Prado et al, 2016). Indeed, comparisons of parent reports with child speech samples (Dollaghan et al., 1999; Feldman et al, 2000) and with a behavioral comprehension measure (DeAnda et al., 2015) have revealed comparable accuracy across SES in reporting on vocabulary production on the MBCDI. Others, assessing vocabulary by means of parent report, language samples, and language tests, also showed comparable accuracy across SES (Furey, 2011; Sachse & Suchodoletz, 2008) in reporting vocabulary.

Recent efforts in developing vocabulary assessments for young children have used statistical techniques such as Item Response Theory (IRT; Embretson & Reise, 2013) to select a set of items that are most efficient for discriminating children with different ability levels (Bohn et al., 2023; Kachergis et al., 2022; Chai et al., 2020). However, data-driven methods such as IRT require a large set of data to be able to calculate each item's difficulty and discrimination power. Thus, these methods are most useful when one is selecting from a larger population of items with substantial data on each, but do not offer help in identifying items for a new instrument. Given that the goal of the present paper was to develop these initial set of items, we focused on manual iterations to arrive at the set of items in the current S-CDI III. Once sufficient data has been collected using this instrument, future studies could use IRT to further refine it and improve its psychometric properties.

Limitations

This study had several limitations. First, there is some evidence of a ceiling effect for vocabulary in high-performing children as they approach 48 months of age. This suggests that the instrument may not reflect the full range of variability for these children. Nevertheless, this is quite modest for the IDHC-III compared to the pilot version and even the original MBCDI-III and other language adaptations. Further, the instrument does well at differentiating the lowest performing children from those in the mid and upper ranges. We also observed a floor effect on the sentence complexity scale for younger children. This likely reflects the more complex mature grammatical forms chosen for this instrument relative to other adaptations of the MBCDI-III and may have implications for assessment of grammatical development in children up to about 36 months of age with shorter MLUs.

Second, whereas the maternal education levels of the present sample approximate the larger Mexican population, this led to unique challenges and findings: some caregivers with lower education levels required assistance to complete the form. Our analyses suggest that this difference in administration did not alter the reported effects. Nevertheless, we found no effect of maternal education on vocabulary in contrast to our expectation (e.g., Hoff, 2006; 2013). We note however that other recent studies report similar null effects in monolingual and bilingual Spanish-speaking samples (DeAnda et al., 2015; Friend et al., 2022; Montanari et al., 2020) and encourage further research on the influence of administration practices and on the role of maternal education in early Spanish vocabulary.

Finally, early validity studies (Guiberson 2008a & b; Guiberson & Rodríguez, 2010; Guiberson et al., 201) suggest that the IDHC-III may be useful in clinical settings. In particular, in our pilot research, both Vocabulary and Grammatical Complexity correlated positively with scores on the Clinical Evaluation of Language Fundamentals-Preschool (CELF-P; Wiig, Secord, Semel, 2004) and an assessment of discriminant validity yielded moderate specificity and high sensitivity in the detection of language disorders. These findings require confirmation by additional studies.

Conclusion

Our results suggest that parent report is a useful means to obtain language development information in preschoolers in this language setting, just as it is for toddlers.

Thus, it can contribute to meeting the need for cost-effective, valid language assessment instruments as part of the global effort to bring assessment to scale, that is, making it broadly accessible to diverse populations. There are few measures available for Spanish-speaking children and we find that parent report may be a viable means to fill this need, obtaining information about language development for preschoolers acquiring Spanish in Mexico and possibly in other Latin American countries as well. As expected, we found age-related changes in vocabulary and grammatical complexity scores and a strong relationship between the two measures. We also observed expected differences in grammatical complexity scores with maternal education. Together, these metrics provide preliminary indicators of validity: expected relations of age with both vocabulary and grammatical complexity, between vocabulary and grammatical complexity, and between maternal education and grammatical complexity. Given reported variability in parent reports of Spanish vocabulary and our observations, the relation between maternal education and vocabulary in preschool children merits further research.

This study provides a new parental report instrument, the IDHC-III for Spanish-speaking children, normed on a representative sample in a monolingual setting in Mexico. As data from other research has shown (Mancilla-Martínez et al., 2016, 2011, 2013; Marchman & Martínez-Sussman, 2002) it may also be useful in the assessment of bilingual children, but specific studies need to be developed to analyze its use in this case. Norms are now available for this measure that are appropriate for assessing language development in preschool Spanish-speakers from diverse socio-economic backgrounds.

References

Alcock, K.J., Rimba, K., Holding, P., Kitsao-Wekulo, P., Abubakar, A. & Newton, C.R.J.C. (2015). Developmental inventories using illiterate parents as informants: Communicative Development Inventory (CDI) adaptation for two Kenyan languages. *Journal of Child Language*, *42*(4), 763-785.

Auza, A., Murata, C. & Peñaloza, C. (2023) Predictive validity of a parental questionnaire for identifying children with Developmental Language Disorders. *Frontiers in Psychology 14*: 1110449. doi:10.3389/fpsyg.2023.1110449

Barnes, J., & Garcia, I. (2013). Vocabulary growth and composition in monolingual and bilingual Basque infants and toddlers. *International Journal of Bilingualism, 17,* 357-374. https://doi.org/10.1177/1367006912438992

Brieković, L. Š., & Kraljević, J. K. (2023)._Parental reports on language development in toddlers and preschoolers based on the Croatian version of Communicative

Development Inventories III. *Frontiers in Psychology. 14.* https/doi.org/10.3389/fpsyg.2023.1188550

Cadime I., Santos A.L., Ribeiro I., Viana F.L. (2021). Parental reports of preschoolers' lexical and syntactic development: Validation of the CDI-III for European Portuguese. *Frontiers in Psychology, 12*. doi: 10.3389/fpsyg.2021.677575.

Carrow-Woolfolk, E. (1999). Test for Auditory Comprehension of Language (3rd Ed.). Austin, TX: ProEd.

Conboy, B.T. (2019). Spanish language parent report language inventory data from low- and middle-SES families in Guatemala and Mexico. Paper presented as part of the symposium, The Effect of Socioeconomic Level on Language Development in Diverse Populations, at the IX Congress of the Asociación Para el Estudio de la Adquisición del Lenguaje [Spanish Association for the Study of Language Acquisition], UNED, Madrid, September 4 – 6, 2019.

Conboy, B.T., Stansbury, A., Sanchez, E. & Jackson-Maldonado, C. (2017a). Language acquisition and development in bilingual (Kaqchikel-Spanish) Preschool children in Guatemala. Poster presentation. *American Speech, Language, and Hearing Association (ASHA) Convention,* Los Angeles, CA.

Conboy, B.T., Stansbury, A. & Jackson-Maldonado, D. (2017b). Language Development in Bilingual Kaqchikel-Spanish Speaking Contexts. Poster presentation. *International Symposium on Bilingualism,* Ireland.

Dale, P.S., Grazia Tosto, M., Hayiou-Thomas, M.E., Plomin, R. (2015). Why does parental language input style predict child language development? A twin study of gene–environment correlation. *Journal of Communication Disorders, 57,* 106-117, https://doi.org/10.1016/j.jcomdis.2015.07.004.

Dar, M., Anwaar, H., Vihman, M., & Keren-Portnoy, T. (2015). Developing an Urdu CDI for early language acquisition. *York Papers in Linguistics, 14,* 1-14.

DeAnda, S., Arias-Trejo, N., Poulin-Dubois, D., Zesiger, P., & Friend, M. (2015). Effects of minimal L2 exposure and SES on early word comprehension in English and Spanish: new evidence from a direct assessment. *Journal of Bilingualism: Language and Cognition.*

Díez Itza, E., Snow, C. E., & MacWhinney, B. (1999). La metodología RETAMHE y el proyecto CHILDES: breviario para la codificación y análisis del lenguaje infantil. *Psicothema, 11(3), 517-530.*

Dollaghan, C. A., Campbell, T. F., Paradise, J. L., Feldman, H. M., Janosky, J. E., Pitcairn, D. N., & Kurs-Lasky, M. (1999). Maternal education and measures of early speech and language. *Journal of Speech, Language, and Hearing Research, 42*(6), 1432-1443.

Dunn, L. M., & Dunn, L. M. (1997). Peabody Picture Vocabulary Test—III. Circle Pines, MN: American Guidance Service.

Edwards, S., Letts, C., & Sinka, I. (2011). *The New Reynell Developmental Language Scales*. London: GL Assessment Ltd.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.
Eriksson, M. (2017). The Swedish Communicative Development Inventory III. Parent reports on language in preschool children. *International Journal of Behavioral Development,* 41(5), 647-654.

Ezeizabarrena, M.J., Barnes, J.; García, I., Barreña, A. & Almgren, M. (2013). Using Parental Report Assessment for Bilingual Preschoolers: the Basque experience. In V. C. Mueller Gathercole (ed.), *Solutions for the assessment of bilinguals* (57-80). Clevedon: Multilingual Matters.

Feldman, H. M., Dale, P. S., Campbell, T. F., Colborn, D. K., Kurs-Lasky, M., Rockette, H. E., & Paradise, J. L. (2005). Concurrent and predictive validity of parent reports of child language at ages 2 and 3 years. *Child Development, 76,* 856-868.

Feldman, H. M., Dollaghan, C. A., Campbell, T. F., Kurs-Lasky, M., Janosky, J. E., & Paradise, J. L. (2000). Measurement properties of the MacArthur Communicative Development Inventories at ages one and two years. *Child Development, 71,* 310-322.

Feldman, H. M., Dollaghan C. A., Campbell T. F., Colborn D. K., Janosky J., Kurs-Lasky, M., Rockette, H. E., Dale, P. S. & Paradise, J.L. (2003). Parent-reported language skills in relation to otitis media during the first 3 years of life. *Journal of Speech, Language & Hearing Research, 46.*

Fenson, L., Bates, E., Dale, P., Goodman, J., Reznick, J. S., & Thal, D. (2000). Reply: Measuring variability in early child language: Don't shoot the messenger. *Child Development, 71*(2), 323-328.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J. & Pethick, S. J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development,* 59 (5, Serial 242).

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, S., & Bates, E. (2007). *MacArthur-Bates Communicative Development Inventories (2nd Ed.).* Baltimore, MD: Brookes Publishing.

Fenson, L., Pethick, S., Renda, C., Cox, J. L., Dale, P. S., & Reznick, J. S. (2000). Short-form versions of the MacArthur communicative development inventories. *Applied Psycholinguistics, 21,* 95-116.

Fernández Vázquez, F. & Aguado, G. A. (2007). Medidas del desarrollo típico de la morfosintaxis para la evaluación del lenguaje espontáneo de niños hispanohablantes. *Revista de Logopedia, Foniatría y Audiología, 27,* 140-152.

Frank, M. C., Marchman, V., Yurovsky, D., & Braginsky, M. (2021). The Wordbank Project: Variability and Consistency in Children's Language Learning Across Languages. Cambridge, MA: MIT Press.

Friend, M., Lopez, O., De Anda, S., Abreu-Mendoza, R. A., & Arias-Trejo, N. (2022). Maternal education revisited: Vocabulary growth in English and Spanish from 16 to 30 months of age. *Infant Behavior and Development, 66,* https://doi.org/10.1016/j.infbeh.2021.101685

Furey, J. E. (2011). Production and maternal report of 16-and 18-month-olds' vocabulary in low-and middle-income families. *American Journal of Speech-Language Pathology,* 20, 38-46.

Gonzalez, J.E., Bengochea, A., Justice, L., Yeomans-Maldonado, G., & McCormick, A. (2019). Native Mexican Parents' Beliefs About Children's Literacy and Language Development: A Mixed-Methods Study, *Early Education and Development, 30,* 259-279, 10.1080/10409289.2018.1542889

Guiberson, M. (2008a). Concurrent validity of a parent survey measuring communication skills of Spanish-speaking preschoolers with and without delayed language. *Perspectives on Communication Disorders in Culturally and Linguistically Diverse Populations, 15,* 73-81.

Guiberson, M. (2008b). Validity of a parent vocabulary checklist for young Spanish-speaking children of Mexican immigrants. *International Journal of Speech-Language Pathology, 10,* 279–285.

Guiberson, M., & Rodríguez, B. L. (2010). Measurement properties and classification accuracy of two Spanish parent surveys of language development for preschool-age children. *American Journal of Speech-Language Pathology, 19,* 225-237.

Guiberson, M., Rodriguez, B. L., & Dale, P. S. (2011). Classification accuracy of brief parent report measures of language development in Spanish-speaking toddlers. *Language, Speech, and Hearing Services in Schools, 42,* 536-549.

Hamadani, J. D., Baker-Henningham, H., Tofail, F., Mehrin, F., Huda, S. N., & Grantham-McGregor, S. M. (2010). Validity and reliability of mothers' reports of language development in 1-year-old children in a large-scale survey in Bangladesh. *Food and Nutrition Bulletin, 31*(2_suppl2), S198-S206.

Hoff, E. (2006). How social contexts support and shape language development. *Developmental Review, 26,* 55-88.

Hoff, E. (2013). Interpreting the early language trajectories of children from low-SES and language minority homes: implications for closing achievement gaps. *Developmental Psychology, 49,* 4-14.

Hoff, E., Burridge, A., Ribot, K. M., & Giguere, D. (2018). Language specificity in the relation of maternal education to bilingual children's vocabulary growth. *Developmental Psychology, 54,* 1011–1019. https://doi.org/10.1037/dev0000492

Holm, E., Hansen, P. B., Romøren, A. S. H., & Garmann, N. G. (2023). The Norwegian CDI-III as an assessment tool for lexical and grammatical development in preschoolers. *Frontiers in Psychology, 14,* 1175658.

Individuals with Disabilities Act (2022) Annual Report to Congress. https://sites.ed.gov/idea/2022-individuals-with-disabilities-education-act-annual-report-to-congress/

Huttenlocher, J., Waterfall, H., Vasilyeva, M., Vevea, J., & Hedges, L. V. (2010). Sources of variability in children's language growth. *Cognitive Psychology, 61,* 343-365.

Jackson-Maldonado, D. (2012). Verb morphology and vocabulary in monolinguals, emerging bilinguals, and monolingual children with Primary Language Impairment. In B. Goldstein (Ed.), *Bilingual Language Development and Disorders in Spanish-English Speakers. 2nd edition.* Baltimore: Brookes, pp 153-173.

Jackson-Maldonado, D. & Conboy, B. T. (2007). Utterance length measures for Spanish-speaking toddlers: the morpheme vs word issue revisited. En J.G. Centeno, L.K. Obler y R. Anderson (Eds) *Studying Communication Disorders in Spanish Speakers:*

Theoretical, research & clinical aspects. Multilingual Matters: North Somerset, England.

Jackson-Maldonado, D. & Maldonado, R. (2017). Grammaticality differences between Spanish-speaking children with Specific/Primary Language Impairment (SLI/PLI) and their typically developing peers. *International Journal of Language and Communication Disorders 52,* 750-765.

Jackson-Maldonado, D. & Maldonado, R. (2015). La complejidad sintáctica en niños con y sin Trastorno Primario de Lenguaje. In I. Rodríguez Sánchez y E.Vázquez (Eds.) *Lingüística Funcional.* Querétaro: Universidad Autónoma de Querétaro, pp. 253-301.

Jackson-Maldonado, D. & Maldonado, R. (2016). El uso de conectores en niños con y sin Trastorno del Lenguaje. *Lingüística Mexicana, 8,* 33-50.

Jackson-Maldonado, D., Marchman, V. A., & Fernald, L. C. (2013). Short-form versions of the Spanish MacArthur–Bates Communicative Development Inventories. *Applied Psycholinguistics, 34,* 837-868.

Kas, B., Jakab, Z., & Lőrik, J. (2022). Development and norming of the Hungarian CDI-III: A screening tool for language delay. *International Journal of Language & Communication Disorders*. https://doi-org.libproxy.unm.edu/10.1111/1460-6984.12686

McCray, G., McCoy, D., Kariger, P., Janus, M., Black, M. M., Chang, S. M., Tofail, F., Eekhout, I., Waldman, M., Buuren, S. van, Khanam, R., Sazawal, S., Nizar, A., Schönbeck, Y., Zongo, A., Brentani, A., Zhang, Y., Dua, T., Cavallera, V., ... Gladstone, M. (2023). The creation of the Global Scales for Early Development (GSED) for children aged 0–3 years: Combining subject matter expert judgements with big data. BMJ Global Health, *8,* e009827. https://doi.org/10.1136/bmjgh-2022-009827

Mancilla-Martinez, J., Gámez, P. B., Vagh, S. B., & Lesaux, N. K. (2016). Parent reports of young Spanish–English bilingual children's productive vocabulary: A development and validation study. *Language, Speech, and Hearing Services in Schools, 47,* 1-15.

Mancilla-Martinez, J., Pan, B. A., & Vagh, S. B. (2011). Assessing the productive vocabulary of Spanish–English bilingual toddlers from low-income families. *Applied Psycholinguistics, 32*(2), 333-357.

Mancilla-Martinez, J., & Vagh, S. B. (2013). Growth in toddlers’ Spanish, English, and conceptual vocabulary knowledge. *Early Childhood Research Quarterly, 28,* 555-567.

Marchman, V. A., Dale, P. S., & Fenson, L. (2023). MacArthur-Bates Communicative Development Inventories: User's Guide and Technical Manual, 3rd Ed., Brookes: Baltimore.

Marchman, V. A., & Martínez-Sussmann, C. (2002). Concurrent validity of caregiver/parent report measures of language for children who are learning both English and Spanish. *Journal of Speech, Language, and Hearing Research, 45,* 983-997.

Minto-García, A., Canto, E. A. A., & Arias-Trejo, N. (2020). Mothers' Use of Gestures and their Relationship to Children's Lexical Production. *Psychology of Language and Communication, 24,* 175–200. https://doi.org/10.2478/plc-2020-0010

Mokhtari, F., Kazemi, Y., Feizi, A., & Dale, P. (2022). Psychometric Properties of the MacArthur-Bates Communicative Development Inventories-III (CDI-III) in 30 to 37 Months Old Persian-Speaking Children. *Archives of Rehabilitation, 23,* 372-391.

Morgan, G., Restrepo, M. A., & Auza, A. (2009). Variability in the grammatical profiles of Spanish-speaking children with specific language impairment. *Hispanic child languages: Typical and impaired development, 50,* 283-303.

Montanari, S., Mayr, R., & Subrahmanyam, K. (2020). Speech and language outcomes in Spanish-English preschoolers: the role of maternal education. *International Journal of Bilingual Education and Biilngualism,* 1–19.
https://doi.org/10.1080/13670050.2020.1781780

Muggeo, V. (2013). quantregGrowth: growth charts via regression quantiles. *R package version 0.1, 1.*

Organization for Economic Cooperation and Development (2023), Education at a Glance 2023: OECD Indicators, OECD Publishing, Paris,
https://doi.org/10.1787/e13bef63-en.

Passel, J. S., Cohn, C., & Lopez, M. H. (2011). Hispanics Account for more than Half of Nation's Growth in Past Decade. Census 2010: 50 Million Latinos. Washington, DC: Pew Hispanic Center.

Peña, E. D. (2007). Lost in translation: Methodological considerations in cross-cultural research. *Child Development, 78,* 1255-1264.

Pérez-Leroux, A. T., Castilla-Earls, A. P., & Brunner, J. (2012). General and specific effects of lexicon in grammar: Determiner and object pronoun omissions in child Spanish. *Journal of Speech, Language, and Hearing Research, 55,* 313–327.

Perra, O., McGowan, J. E., Grunau, R. E., Doran, J. B., Craig, S., Johnston, L., Jenkins, J., Holmes, V. & Alderdice, F. A. (2015). Parent ratings of child cognition and language compared with Bayley-III in preterm 3-year-olds. *Early Human Development, 91,* 211-216.

Prado, E. L., Adu-Afarwuah, S., Lartey, A., Ocansey, M., Ashorn, P., Vosti, S. A., & Dewey, K. G. (2016). Effects of pre-and post-natal lipid-based nutrient supplements on infant development in a randomized trial in Ghana. *Early Human Development, 99,* 43-51.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Roseberry-McKibbin, C. (2013). Increasing language skills of students from low-income backgrounds: Practical strategies for professionals. Plural Publishing.

Rosemberg, C. R., & Barreiro, A. (2022). *Interacción social, interaccionales y discursivos en el uso infantil de lenguaje mentalista.* XVIII REUNIÓN NACIONAL - VII ENCUENTRO INTERNACIONAL Asociación Argentina de Ciencias del Comportamiento, 6-7. https://dialnet.unirioja.es/descarga/articulo/9066708.pdf
.

Rubio-Codina, M., Attanasio, O., Meghir, C., Varela, N., & Grantham-McGregor, S. (2015). The socioeconomic gradient of child development: Cross-sectional evidence from children 6–42 months in Bogota. *Journal of Human Resources, 50,* 464-483.

Rubio-Codina M., Araujo M. C., Attanasio O., Muñoz P., Grantham-McGregor S. (2016). Concurrent validity and feasibility of short tests currently used to measure early childhood development in large scale studies. *PLoS ONE,* 11: e0160962. https://doi.org/10.1371/journal.pone.0160962

Rubio-Codina M., Grantham-McGregor S. (2020). Predictive validity in middle childhood of short tests of early childhood development used in large scale studies compared to the Bayley-III, the Family Care Indicators, height-for-age, and stunting: A longitudinal study in Bogota, Colombia. *PLoS One.* doi: 10.1371/journal.pone.0231317.

Sachse, S., & Von Suchodoletz, W. (2008). Early identification of language delay by direct language assessment or parent report? *Journal of Developmental & Behavioral Pediatrics, 29*(1), 34-41.

Sanz-Torrent, M., Serrat, E., Andreu, L., & Serra, M. (2008). Verb morphology in Catalan and Spanish in children with specific language impairment: a developmental study. *Clinical Linguistics & Phonetics, 22*, 459-474.

Secretaría de Educación Pública. (2017). *Aprendizajes Clave para la Educación Integral. Educación Preescolar*. México: Secretaría de Educación Pública.

Simon-Cereijido, G. (2015). Preschool language interventions for Latino dual language learners with language disorders: What, in what language, and how. *Seminars in Speech and* Language, *36,* 154-164.

Simon-Cereijido, G., Conboy, B.T., & Jackson-Maldonado, D. (2020). El derecho humano de ser multilingüe: recomendaciones para logopedas [The human right of being multilingual: recommendations for speech-language therapists]. *Revista de Logopedia, Foniatría y Audiología, 40,* 178-186.

Skarakis-Doyle, E., Campbell, W., & Dempsey, L. (2009). Identification of children with language impairment: Investigating the classification accuracy of the MacArthur–Bates Communicative Development Inventories, Level III. *American Journal of Speech-Language Pathology, 18,* 277-288.

Skeat, J., Eadie, P., Ukoumunne, O., & Reilly, S. (2010). Predictors of parents seeking help or advice about children's communication development in the early years. *Child: Care, health and development, 36,* 878-887.

Smithson, Michael and Verkuilen, Jay, (2006). A Better Lemon Squeezer? Maximum-Likelihood Regression with Beta-Distributed Dependent Variables. Psychological Methods, *11,* 54-71.

Squires, J. & Bricker, D. (2009). *The Ages and Stages Questionnaire (ASQ) in Spanish.* Baltimore: Brookes Publishing Co.

Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V., & De Bastiani, F. (2017). *Flexible regression and smoothing: using GAMLSS in R*. CRC Press.

Stolt, S. (2023). Internal consistency and concurrent validity of the parental report instrument on language in pre-school-aged children–The Finnish Communicative Development Inventory III. *First Language, 43,* 492-515.

Tek, S., Jaffery, G., Fein, D., & Naigles, L. R. (2008). Do children with Autism Spectrum Disorders show a shape bias in word learning? *Autism Research, 1,* 208-222.

Tulviste, T., & Schultz, A. (2020). Parental reports of communicative development at the age of 36 months: The Estonian CDI-III. *First Language, 40,* 64-83.

Uccelli, P. (2009). Emerging temporality: past tense and temporal/aspectual markers in Spanish-speaking children's intra-conversational narratives. *Journal of Child Language, 36,* 929-966.

Uccelli, P., & Páez, M. M. (2007). Narrative and vocabulary development of bilingual children from kindergarten to first grade: Developmental changes and associations among English and Spanish skills. *Language, Speech, and Hearing Services in Schools, 38,* 225–236.

Ukoumunne, O. C., Wake, M., Carlin, J., Bavin, E. L., Lum, J., Skeat, J. & Reilly, S. (2012). Profiles of language development in pre-school children: a longitudinal latent class analysis of data from the Early Language in Victoria Study. *Child: Care, Health and Development, 38,* 341-349.

Urm, A., & Tulviste, T. (2021). Toddlers' Early Communicative Skills as Assessed by the Short Form Version of the Estonian MacArthur-Bates Communicative Development Inventory II. *Journal of Speech, Language, and Hearing Research, 64,* 1303-1315.
Vasilyeva, M., Waterfall, H., & Huttenlocher, J. (2008). Emergence of syntax: Commonalities and differences across children. *Developmental Science, 11,* 84-97.

Verdisco, A., Cueto, S., Thompson, J., Engle, P., Neuschmidt, O., Meyer, S., González, E., Oré, B., Hepworth, K., & Miranda, A. (2009). Urgency and possibility results of PRIDI A first initiative to create regionally comparative data on child development in four Latin American countries technical annex. Technical Annex. Inter-American Development Bank, Washington DC.

Weber, A. M., Marchman, V. A., Yatma, D. I. O. P., & Fernald, A. (2018). Validity of caregiver-report measures of language skill for Wolof-learning infants and toddlers living in rural African villages. *Journal of Child Language, 45*, 939-958.

Wiig, E. H., Secord, W. A., & Semel, E. (2004). *Clinical Evaluation of Language Fundamentals—Preschool (2nd ed.).* (CELF Preschool-2). Toronto, Canada: The Psychological Corporation/A Harcourt Assessment Company.

World Health Organization & UNICEF. (2012). Early childhood development and disability: A discussion paper. https://apps.who.int/iris/handle/10665/75355Zimmerman, I. L., Steiner, V. G., & Pond, R. E. (2002). *Preschool Language Scale, Fourth Edition: Spanish Edition.* San Antonio, TX: Harcourt Assessment.

Data, Code and Materials Availability statement

The IDHC-III form is non-commercially distributed by the MacArthur-Bates CDI Advisory Board, Larry Fenson, Chair, and is available free-of-charge from https://mb-cdi.stanford.edu. The form is also available, along with all data, code, and materials (including Appendices) on OSF at https://osf.io/r6fep/, DOI 10.17605/OSF.IO/R6FEP

Ethics statement

Ethics approval was obtained from the Universidad Autónoma de Querétaro by the first author. All participants gave informed written consent before taking part in the study.

Authorship and Contributorship Statement

DJM developed the measure, data collection and analysis, and prepared the first draft of this manuscript, MF contributed refinements to the measure, data interpretation, organization, and writing the new manuscript, VM contributed refinements to the measure, data analyses, interpretation, and writing of the new manuscript, AW contributed refinements to the measure, data interpretation, and writing of the new manuscript, AA contributed refinements to the measure, acquisition of original data files, data interpretation and writing of the new manuscript, BC contributed to the interpretation of the data and writing of the new manuscript, MRC contributed to the acquisition of original data files and interpretation of data, and PD contributed refinements to the measure, data interpretation, and writing the new manuscript. All authors reviewed and approved the final manuscript.

Declaration of conflict of interests

Authors Friend, Weisleder, Marchman, and Dale are members of the MacArthur-Bates CDI Advisory Board.

Dedication and acknowledgement

This paper is dedicated to the memory of our colleague and good friend, Donna Jackson-Maldonado, whose untimely passing occurred on November 29, 2021. We miss her infectious good spirit, the stories, jokes, and laughter that made collaboration with her such a pleasure. Underpinning that friendship was her intense commitment to the study of language development and disorders, to training new clinicians and researchers, and to helping young children grow. We will miss her tremendously.

We gratefully acknowledge Marta Rubio-Codina, Ph.D., at Inter-American Development Bank, for her assistance in locating many of the files that facilitated our work on this manuscript.

License

The Development of color terms in Shipibo-Konibo children

Martin Fortier*
Paris Sciences et Lettres University, FR

Danielle J. Kellier*
University of Pennsylvania, USA

María Fernández-Flecha
Pontificia Universidad Católica del Perú, PE

Michael C. Frank
Stanford University, USA

* these authors contributed equally.

Abstract: Color word learning is an important case study for the relationship between language and perception. While English color word learning is well-documented, there is relatively limited evidence on the developmental trajectory for color words, especially in languages from non-Western populations. We study color words and their acquisition in the Shipibo-Konibo (SK), an indigenous group within the Peruvian Amazon. In Study 1, we measure the color vocabulary in SK adults, updating findings from the World Color Survey. We then study receptive and productive knowledge of color words in children, conducted in both SK (Study 2) and Spanish (Study 3). Children learning the SK system show a protracted developmental trajectory towards adult-like color term knowledge compared to contemporary studies of English-speaking children. Further, when SK children lack precise color term knowledge, they appeared to follow different strategies for SK and Spanish, using Spanish vocabulary in SK and overgeneralizing in Spanish. For both children and adults, bilingual vocabulary is used adaptively to facilitate task performance, broadly supporting communicative views of color vocabulary.

Keywords: Shipibo-Konibo; Vocabulary development; Color; Bilingualism

Corresponding author(s): Michael C. Frank, Department of Psychology, Stanford University, Stanford, 450 Serra Mall, Jordan Hall, Building 420, Stanford, CA 94301, USA. Email: mcfrank@stanford.edu.

ORCID ID(s): https://orcid.org/0000-0003-3282-8120 (Fortier, M.); https://orcid.org/0000-0001-7811-3468 (Kellier, D.J.); https://orcid.org/0000-0002-2699-2509 (Fernández-Flecha, M.), https://orcid.org/0000-0002-7551-4378 (Frank, M. C.)

Citation: Fortier, M., Kellier, D. J., Fernández-Flecha, M., Frank, M. C. (2024). The development of color terms in Shipibo-Konibo children. *Language Development Research, 4*(1), 399–430. http://doi.org/10.34842/7zf8-gk25

Introduction

Color is where language and perception meet. Words such as “blue” and “red” draw boundary lines across a perceptually continuous space of hues and shades. In English, there are 11 high frequency color terms that together span the color space, but this categorization system is not universal. For instance, Russian speakers use two distinct words to describe the colors light blue (“goluboy”) and dark blue (“siniy”); other languages have as few as two words (e.g., the Jalé people only have terms for “light” and “dark”; Berlin & Kay, 1969). Why do languages vary in their color systems? One emerging consensus is that languages categorize the color spectrum in different ways in part due to functional demands (Gibson et al., 2017): both smaller and larger color systems are relatively optimal for different communicative needs (Regier et al., 2007; Zaslavsky et al., 2018).

Learnability is hypothesized to be one contributor to this cross-linguistic diversity (Chater & Christiansen, 2010; Culbertson et al., 2012). Some color systems may be easier for children to learn than others, or children may show inductive biases that shape the color vocabulary. But the actual acquisition of color terms – while relatively well-studied in English (e.g., Forbes & Plunkett, 2019; Saji et al., 2015; Sandhofer & Smith, 1999; Wagner et al., 2013, 2018) – is relatively under-studied across other populations (cf. Forbes & Plunkett, 2020).

In the current project, our goals were (1) to characterize color term knowledge in an indigenous population, the Shipibo-Konibo (SK), and then (2) to build on this foundation to characterize the developmental trajectory of color language acquisition in a group of children raised learning Shipibo-Konibo, a departure from the WEIRD (Western Educated Industrialized Rich Democratic) populations that are over-represented in behavioral science (Henrich et al., 2010; Nielsen et al., 2017). This work provides a developmental comparison to understand both consistencies and variabilities in the trajectory of color word learning for children who are growing up in environments with far fewer manufactured, multi-colored plastic toys (Gibson et al., 2017) but probably exposed to many more hues of color on plants, fruits, etc. in their natural setting.

In the remainder of the introduction, we review color vocabulary development in children, and then we turn to what is currently known about color terms in Latin American varieties of Spanish, such as Mexican, Colombian, and Bolivian Spanish, and in some Amazonian languages, such as Candoshi, Pirahã, and Shipibo-Konibo. These two literatures set the stage for our own study.

The Development of Color Vocabulary

To adults, colors are extremely salient attributes of the perceptual world; even when color is seemingly task-irrelevant, we mention it (e.g., Sedivy, 2003). It is quite surprising then that children sometimes struggle to master color vocabulary. Early observations by Darwin, Bateman, Nagel, and others attest to individual children's delays in the correct use of color terms well into middle childhood; several diarists report 5- to 8-year-olds with limited mastery of basic level color terms (reviewed in Bornstein, 1985). Importantly, different tasks may give different answers about whether a child has "learned" a color word – recognizing that a word names the dimension of hue (even if you don't know which hue) suggests some partial learning. A child might be able to recognize that a word matches one color chip better than another even if they could not spontaneously recall that color word.

Yet the early observations about difficulties in school-aged children's color naming are still surprising in light of the body of infant research that suggests that infants' color discrimination abilities are relatively well-developed by the end of the first year of life (for review see e.g., Bornstein, 2015).

Indeed, the age at which color words are learned has been shifting over the past hundred years, at least for English-speaking children. Bornstein (1985) documents substantial decreases in the age at which many children master their colors, citing four years as an age at which most children are proficient. In fact, this age may have even decreased further in the last thirty years, judging from recent studies (Forbes & Plunkett, 2019; Wagner et al., 2013, 2018). What makes color words hard to learn, and why are they getting easier?

One prominent account of what makes color word learning difficult is that children may not recognize that color words pick out the perceptual dimension of hue at all (Bartlett, 1977; Sandhofer & Smith, 1999), and that once they do, children rapidly map colors correctly onto the appropriate range of hues in color space. This account nicely explains the observation that there is often a period during which children will produce an inappropriate color word when asked "what color is this?" – they know that color words go together and answer a particular question, they just don't know which color is which. A further point of parsimony for this account is that infants' color boundaries are not all that different in their placement from those of adults; thus, presumably the mapping task they face – from words to hues – is not all that difficult, once they recognize the dimension that they are attempting to map (Bornstein et al., 1976; Franklin et al., 2005).

On the other hand, when children's mapping errors are examined in detail, they show more systematicity than would be predicted by this account. Wagner et al. (2013) show that children who have not yet fully mastered the color lexicon nevertheless use colors in ways that are more consistent with overextension than with ignorance of the dimensional mapping – for example, using “blue” to refer to blue and green hues (which are close together in color space). These overextensions are reminiscent of noun overextensions that have been documented in early word learning, for example calling a horse “dog” (Clark, 1973). Further, the order of acquisition for color word meanings in Wagner et al. (2013) was well-predicted by the frequency and perceptual salience of color categories (Yurovsky et al., 2015), supporting the view that color categories are learned gradually from perceptual experiences rather than all at once. Finally, both behavioral and eye-tracking evidence suggest that children show earlier comprehension than production for color words (Forbes & Plunkett, 2019; Sandhofer & Smith, 1999; Wagner et al., 2018), a phenomenon that is seen throughout early word learning. In eye-tracking tasks, comprehension also shows evidence of perceptual overextensions, such that children fixate perceptually close distractor colors more than far distractors (Wagner et al., 2018). In sum, although attention to the dimension of hue may be one difficult component of color word learning, systematic mapping of words to particular regions of perceptual space is likely another.

Why is color word learning occurring earlier in development, at least for English-learning children (Bornstein, 1985)? There are at least two obvious, plausible reasons. The first is the increasing prevalence of manufactured toys for children that vary exclusively in color (e.g., sets of plastic blocks of different colors; Gibson et al., 2017). Such objects provide perfect contrastive input for mapping: if one is called “blue” and the other is not, such input implicates pragmatically that “blue” is an informative term (Clark, 1987; Frank & Goodman, 2014). The second is a cultural landscape for parents and early educators that presupposes color words are an important part of early childhood education practices, and as such should be taught explicitly (perhaps using toys specifically made for this purpose). One further explanation comes from Scott et al. (2023), who noted that Japanese children appear to learn color boundaries more slowly than US and German children. They speculated that these differences might be due to differences in the organization of Japanese, German, and English color naming systems. At present, none of these theories has been tested quantitatively, so all are relatively speculative at this time.

In the current paper, we ask about the trajectory of color word learning in an environment where the first two factors are less prevalent: that is, manufactured toys are less frequent, and parents are (at least anecdotally) less motivated to provide color labels to their children. Here we are inspired by the work of Piantadosi et al. (2014), who studied the learning of number word meanings in children in an Amazonian culture. They found that, despite differences in developmental timing, the patterns of generalization of number meaning were generally similar to those documented in WEIRD populations. We are interested in whether we observe similar dynamics in color word learning. In the next section, we turn to the question of adults' color vocabulary in Spanish and Amazonian language, setting the stage for our studies of acquisition.

Color in Latin American varieties of Spanish and Amazonian languages

Since the color systems local to the SK provide a backdrop for our work, in this section, we provide a brief overview of descriptive work on Latin American Spanish and some Amazonian languages. In brief, our conclusion is that ad hoc color terms – descriptors of objects or properties that are adopted for the description of hue (e.g., the use of terms like "blood" or "bloody" to refer to red objects) – are quite common, presaging some of our findings. They are likely present in several Latin American Spanish dialects and are well-attested in Amazonian color systems.

An initial framework for the cross-linguistic study of color came from the World Color Survey or WCS (Berlin & Kay, 1969; Kay et al., 2009). WCS presented adult speakers of over 100 languages with differently colored chips and asked them to produce a label, characterizing the space of color vocabulary in a range of written and unwritten languages. The WCS focused on basic level color terms, the color words that are highest frequency and most consistently used.

The WCS framework has been revised and questioned in subsequent work, however (e.g., Levinson, 2000). In particular, there has been significant controversy about the applicability of the framework to Amazonian languages, centered around the status of ad hoc color terms. Such ad hoc terms are a common way that languages supplement color vocabulary (e.g., Kristol, 1980). Historical case studies suggest that ad hoc terms can often become conventionalized basic level color terms (e.g., the English color "orange" derives from an ad hoc term based on the fruit; St. Clair, 2016).

Since the WCS, however, later research has suggested that ad hoc terms are present in some South American dialects of Spanish and that they play a central role in Amazonian color systems. With respect to Spanish, the WCS identified the following basic level terms in the Mexican dialect: "blanco" (white), "negro" (black), "rojo" (red), "verde" (green), "amarillo" (yellow), "azul" (blue), "café" (brown or coffee-colored), "morado" (purple), "rosa" (pink or rose), "anaranjado" (orange, strictly referring to the color) and "gris" (gray). However, Aragón (2016) offers an ethnolinguistic study of color terms in Mexican Spanish and concludes that the local natural and cultural referents constitute a point of consensus among Mexicans when defining terms of color, even though these colors still follow the general schema of basic level terms. Further, Monroy and Custodio (1989) suggest that Colombian Spanish may include ad hoc color terms referring to colors through objects prototypically instantiating these colors (e.g., vegetables, animals, food, metals, precious stones, fire and its derivatives, and atmospheric phenomena). Lillo et al. (2018) generally confirm these observations, finding an additional basic level term in Uruguayan Spanish, “celeste” (sky blue), which may be a conventionalized ad hoc term (“celeste” is etymologically related to “sky”). This observation is also confirmed by Gibson et al. (2017) for Bolivian Spanish.

Turning now to Amazonian languages, SK color terms were studied in the original WCS. In this data collection effort, they list 21 distinct terms (though this could be categorized as 20 as “huiso” and “wiso” are alternative spellings of the same color term).[1] As their protocol has the field experimenters ask only for basic level color terms, it is assumed that all recorded terms are basic, but only six terms appear in >5% of WCS trials; 10 terms appear in <1% of trials (see Figure 1A for a representation of this data). Immediately the issue of ad hoc terms rears its head, since it is likely that many of these other words are ad hoc color terms (Levinson, 2000).[2]

Several other indigenous Amazonian color systems were studied in the WCS and one of them, Candoshi, has been examined more recently (Surrallés, 2016). Contrary to the WCS, Surrallés argues that no proper color terms exist in this language. If the fieldworkers of the WCS found otherwise, he claims, it is only because they misidentified the elicited terms as basic level color terms when they are nothing more than a series of ad hoc terms referring to objects or animals of the surrounding environment. For example, in Candoshi, the word for yellow is “ptsiyaromashi” (“like the feathers of a milvago bird”), the word for red is “chobiapi” (“ripe fruit”), the word for green is “kamachpa” (“unripe fruit”), etc. These findings lead Surrallés to argue that the Candoshi do not have a proper color system. When they use “color terms” they are not trying to subsume objects of the world under abstract color categories, but they are rather establishing horizontal and ad hoc comparisons between similar objects of the world.

A similar criticism of the WCS approach was given by Everett (2005) based on his study of Pirahã, another Amazonian language. Everett also rejected the idea that there are basic level terms, arguing that the four color terms identified as basic in the WCS are not such. For example, the word identified as the basic level for *red/yellow* in Pirahã (“bi i sai”) was argued to be simply a property descriptor meaning “blood-like.” The argument here is that Pirahã color terms might be ad hoc comparisons rather than

[1] Two anthropological studies (Morin, 1973; Tournon, 2002) have also investigated the color terms used in SK. However, these two studies contain some serious methodological pitfalls: a very limited number of color chips were tested with only a few participants. As a result, we will not further discuss these studies in the remainder of this article and will only focus in our study on a comparison with the WCS data.

[2] In fact, a greater diversity of color terms beyond the basic level is used in the data for the majority of WCS languages (Gibson et al., 2017; Figure S1), suggesting that the effort to elicit only basic level color terms in WCS may not have been successful.

proper basic terms, though there was no quantitative evaluation of this claim such as analysis of the variability in term use.

Finally, Gibson et al. (2017) compared their Bolivian Spanish data with Tsimane, a language of the Amazonian piedmont. Out of a total of 80 color chips, the Tsimane system exhibited 8 apparently basic color terms. However, in their free-choice paradigm, Tsimane speakers showed high variability in nearly all the color terms used for all color chips presented in their study. Thus, Tsimane speakers appear to show substantial ad hoc term usage as well.

The Current Study

The Shipibo-Konibo people are an indigenous group located within the Peruvian Amazon. They are mainly horticulturalists, fishermen, occasionally hunters but are noted for their strong display of tradition (e.g., via traditional art) despite increasingly regular interactions with the western world. They are also skilled traditional artists or artisans, resorting to these activities as a way to earn an income for their household. Their children receive formal schooling for 4 hours a day, both in SK[3] and Spanish. The proportion of input in Spanish they receive at school increases towards adolescence when they enter secondary education. There can be variation in how both languages coexist in the school setting from one village to another. Most SK adults are considered SK-Spanish bilinguals to different degrees although the elders may have only a minor grasp on Spanish.

The SK are an interesting group to examine from the perspective of color word learning. Although their cultural experience is quite different from the English-speaking WEIRD populations who have been the focus of color word acquisition studies, they are not an isolated hunter-gatherer group. Because of their location on the Ucayali River, one of the main tributaries of the Amazon, the SK culture has always been enmeshed in rich trading networks involving other indigenous groups of the Andes and the Lowlands (in pre-conquest times) as well as Mestizos and Westerners (in post-conquest times; Lathrap, 1970). It would thus be mistaken to think of this culture as an "isolated" or "preserved" one. On the contrary, having been extensively exposed to numerous influences, the SK culture has been constantly reworking and reshaping itself through the centuries. The first deep transformation in Shipibo-Konibo culture can be traced to the 18th century, when Shipibos, Konibos and Shetebos were forced to live together by Franciscan evangelization (Myers, 1974). Later, the second half of the 20th century was characterized by intense contact with the Spanish-speaking Mestizo populations established along the Ucayali River. As a result, today's SK culture

[3] The phonemic inventory of SK language has 4 vowels (/i/, /ɨ/, /a/ and /o/) and 15 consonants: 3 plosives (/p/, /t/ and /k/), 2 affricates (/ts/ and /ʧ/), 2 nasals (/m/ and /n/), 5 fricatives (/β/, /s/, /ʃ/, /ʂ/, /ɦ/), and 3 approximants (/w/, /ɹ/ and /j/).

straddles two worlds: children grow up in a traditional culture but with some exposure to formal education (where both Spanish and SK are used by teachers in a formal setting) and – critically – some of the manufactured, colored plastic goods that have been argued to create a context for easily disambiguating between rich color vocabulary terms (Gibson et al., 2017).[4]

Regarding formal education, SK children start attending school when they are 6 or 7 years old, although some children may enroll later on. Boys are more likely to complete the 11 years of basic education than girls. Education is intercultural and bilingual, at least in theory, with some variation occurring in practice, and most classrooms tend to be multi-grade, so children from different school grades may gather together. How they spend time outside school is influenced by gender too: girls usually help with chores around the house, as well as taking care of younger siblings and also working in the "chacra" (small farmland plot where vegetables are grown for the family), while boys also help in the "chacra" but are generally free to move around.
Further, the SK are heavily bilingual. To our knowledge, relatively little work has looked at effects of bilingualism on color word learning. Yet, with much of the world's population growing up multilingual (Shin, 2017), it is important to characterize how learners navigate a conceptual space where they may have words that appropriately name a target concept, but in a different language.

In Study 1, we examine the color vocabulary of current SK adults, comparing their vocabulary to results from the World Color Survey, which was done more than 50 years prior (Berlin & Kay, 1969). Next, we examine SK children's color vocabulary, focusing on their knowledge and their generalization of color terms across both SK (Study 2) and Spanish (Study 3). Through these three studies, we attempted to answer four primary research questions:

1. What is the color vocabulary of SK and how has it changed since the WCS data collection effort?
2. What is the developmental timeline of color term acquisition in a population that has fewer industrial products (toys) and where parents are apparently less inclined to provide color labels to children?
3. Is the developmental course – especially with respect to generalization and the dynamics of comprehension and production – similar to that which has been documented in studies of English color term learning?
4. How is color term learning development affected by bilingual exposure in this group?

[4] Access to manufactured goods varies across SK villages based in part on how close they are to Pucallpa, the regional capital.

To presage our conclusions, we find that SK color vocabulary has remained relatively consistent, with the exception of some intrusions from Spanish in semantic or vocabulary areas of low coverage by the SK color system. Children learning the SK system show a protracted developmental trajectory towards adult-like knowledge compared with modern descriptive studies in WEIRD contexts. Further, when children lack precise color term knowledge, they appear to follow different strategies for SK and Spanish: for SK, children fell back on Spanish knowledge, while for Spanish, we observed substantial over-generalization of terms (Wagner et al., 2013, 2018). Finally, we find that children draw on their Spanish knowledge especially for colors where there is high uncertainty among adult speakers, suggesting that they are adaptively using their bilingual knowledge to facilitate accurate naming.

Study 1

Before we could assess the developmental trajectory of color term knowledge in SK children, our goal was to replicate and update the characterization of the adult SK color system given by the World Color Survey. As the WCS study took place generations prior, we could not assume the SK color term mappings had remained static, especially through years of industrialization and exposure to the Spanish language and its own color term system. As such, Study 1 used a modified version of the original WCS protocol, with an identical color chip set (subsampled to decrease task length). The goals were to characterize the current SK vocabulary and to generate a standard of adult knowledge against which subsequent child participants could be scored.

Methods

Participants

We recruited 39 adult participants (7 men) from multiple sites ranging from more urban and industrialized (from Yarinacocha or San Francisco; 16, 3 men) to more traditional and in closer proximity to the surrounding rainforest (Nueva Betania, Paoyhan, or Puerto Belén; 23, 4 men). All villages were integrated in Peruvian economy and society. Given our relatively small sample size, we did not find apparent differences between participants from different sites.

We experienced difficulty recruiting male participants as many of the men were away from the village during the day, resulting in a sample that was predominantly female. Most participants (31, 4 men) were from SK villages of the Middle Ucayali region (Yarinacocha, San Francisco, and Nueva Betania), with a subset from communities of the Lower (Paoyhan) and Upper (Puerto Belén) Ucayali regions. Within the small town of Yarinacocha (in the vicinity of Pucallpa), we recruited participants (9, 2 men) from Bena Jema, a predominantly SK neighborhood. The remaining recruitment sites (8, 3 men) were native community villages with exclusively SK residents but a strong relationship with those outside their community.

The median age for participants was 38 years (*IQR* = 26-48) ranging from 20 to 64 years. Regarding occupations, 13 out of 32 (41%) female participants were home makers or housewives (33% of the overall sample) and 13 female participants (41%) were artisans (33% overall). Three of the 7 male participants (43%, 8% overall) were horticulturalists. Across both sexes, 5 women (16%, 13% overall) and 3 men (43%, 8% overall) identified as students, comprising a total of 21% of the population. Although all adult participants were required to be native SK speakers, all were introduced to the Spanish language prior to adolescence (median age = 8yo, *IQR* = 5-10).[5]

Materials and procedure

Similar to the original WCS, we used a set of 330 Munsell color chips and asked participants to name them (Berlin & Kay, 1969). We made a number of changes to the procedure, however. In the WCS, every participant provided terms for all 330 chips. Due to fear of participant fatigue, we split up color chips based on their ID numbers (even or odd) and participants were randomly assigned to work with either even- or odd-numbered color chips. As a result, each participant was presented with only 165 chips. All 330 hues within the set are visualized in Appendix 1. Dimensions of the chips were 2 cm × 2.5 cm.

One researcher, MF, traveled to each site to conduct testing aided by a trained research assistant fluent in both SK and Spanish to facilitate communication. In terms of compensation, adult participants were paid for their time and children were offered sweets during testing. All sessions were video recorded. First, the experimenter explained the general procedure and goals of the study to the participant. The experimenter would then present a single color chip to the participant and ask in SK: “What is the color of this chip?” The study was conducted solely in SK language with the assistance of a bilingual SK- and Spanish-speaking research assistant. It should be noted that although the experiment was conducted in SK, the SK word for color used is identical to the Spanish word “color” (an example of SK speakers adopting Spanish words into their lexicon), which might have encouraged Spanish language use.

Besides the reduction in set size, our procedure also differed from that of WCS (see Kay et al., 2009, pp. 585–591) in other aspects. Participants sat in front of the experi-

[5] In this and subsequent studies, we did not perform any checks for colorblindness. Various forms of color vision deficiency are estimated to affect around 8% of men and .5% of women (Simunovic, 2010), and so we expect that a small number of individuals with impaired color vision might have participated in our studies, potentially adding noise to our findings.

menter. To manage changes in natural light intensity between participants, the experiment took place indoors near a window or door instead of outdoors. Another difference between our study and the WCS procedure is in our approach for encouraging participants to describe chips using basic level color terms. In the WCS, the experimenter would instruct participants to only provide basic level color terms during the task (e.g., describing a chip as "blue" as opposed to "navy blue" or "sky-like"). However, we had difficulties concisely explaining the concept of a basic level term compared to other terms.[6] We decided to allow participants to describe a chip with any term they wished, and to ask further questions to elicit a basic level term when they did not do so on their first try. For example, when presented with a red color chip, the participant might use the term "blood-like" (a non-basic term). The experimenter would then ask: "Do you know of any other word to refer to the color of this chip?" Should the participant subsequently respond with "dark red" (another non-basic level term), the experimenter would further ask: "How would you refer to this color with only one word?" Eventually, the participant might use the term "red" (a basic term). For some chips, participants provided a basic level term as their first description. For others, a basic level term might be preceded by 1 or 2 non-basic level terms. When participants failed to provide a basic level term after 3 attempts (i.e., two follow-up questions), no further questions were asked, and the experimenter moved on to the next chip. All responses, basic level or not, were recorded in the order produced by the participant.

Results and Discussion

Figure 1 compares the original WCS data (Panel A) to a summary of results (Panel B) along with the prevalence of Spanish-language responses (Panel C) for Study 1. All participants used the following set of color terms to describe a color chip at least once during their session: "joxo" (light/white), "wiso" (dark/black)[7], "panshin" (yellow), "joshin" (red), and "yankon" (green/blue). Given the widespread use of this term set and their interpretations, we will refer to these five terms as SK-language basic level color terms.

Most (79%) participants also described at least 1 chip as "manxan" (faded), referring to a chip's saturation. In addition, fifty-one percent of participants used the color term "naranja" (or "naransha") to describe at least one chip. "Naranja" may be known as a Spanish-language color term used to describe both the orange fruit and its associated color – as opposed to "anaranjado," a term strictly for the orange color.

[6] Indeed, as Kay et al. (2009, pp. 587–589) acknowledge, there is no straightforward necessary and sufficient criteria for the basicness of a color term (cf. Levinson, 2000).

[7] For the purposes of our studies "huiso" and "wiso" were considered one term.

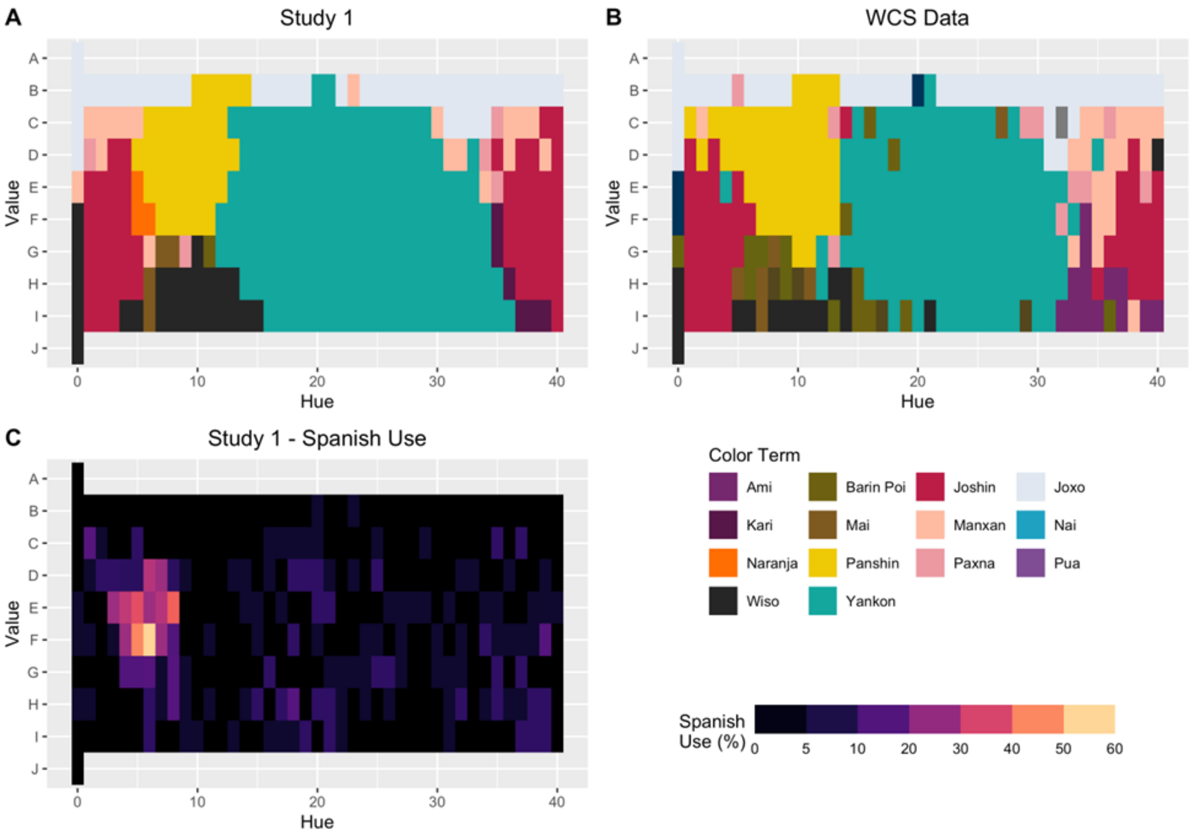


Figure 1. ***(A and B) Plots of the modal term given for a particular chip. Color coordinates were represented in 2-D Munsell space, with Munsell hue represented on the x-axis and Munsell value or lightness represented on the Y-axis. Modal responses were given by SK adults during (A) our Study 1 and during (B) the original World Color Survey. (C) Heat map of prevalence of Spanish-language responses during Study 1. Legends for all three subplots located in the bottom-right quadrant.***

In terms of overall popularity, participants described a median of 32% of chips as "yankon" (*IQR* = 26-39%) followed by "joshin" (*Mdn* = 10%, *IQR* = 7-16%), "panshin" (10%, 6-12%), "joxo" (9%, 6-15%), "manxan" (6%, 1-10%), and "wiso" (5%, 3-8%). We failed to find any significant sex differences in the overall spread of color term usage across chip set ($t(59) = 0.00$, $p > .999$) or in the proportion of subjects who used a term at least once during their session ($t(59) = -1.41$, $p = .164$).[8]

Participants used an SK-language basic level term (i.e., "yankon") to describe a median of 68% of chips (*IQR* = 56-90%). Besides basic level terms, 59% of participants

[8] This failure to find sex differences suggests that color vision impairments (which differ by sex) likely did not have a major effect on our data – perhaps because relatively few men participated in the study.

used SK-language ad hoc color terms (i.e., “nai” or sky for blue chips) for an overall median of 6% of chips (*IQR* = 0-19%). SK-language terms referring to saturation or luminosity of a chip, such as “manxan” (faded) were used for an overall median of 13% of chips (*IQR* = 6-20%). Most instances (86%) of Spanish use involved a Spanish basic color term (BCT) such as “rojo” although only 10% of participants used a Spanish BCT at least once (other than “naranja”).

Compared to the WCS dataset which only reported SK language terms, Spanish use was prevalent throughout Study 1. Fifty-nine percent of our participants used a Spanish-language color term to describe at least 1 chip, which accounted for 4% of all responses. Across chips, the most common Spanish-language color term was “naranja” (51% of participants), followed by “rosa” (10%) and “morado” (8%). Spanish use peaked at 55% when participants were asked to label chips that English speakers would consider to be orange or “anaranjado”/“naranja” by Spanish speakers. Indeed, the relatively common use of “naranja” by these adult SK speakers despite being prompted entirely in SK brings the possibility that “naranja” has been adopted into the SK color lexicon. If we allow “naranja” to be counted as an SK rather than Spanish-language term, then only 15% of participants used a Spanish-language term other than “naranja” at least once throughout the study, accounting for 2% of all responses. One participant responded in Spanish 68% of the time despite being prompted solely in SK.

In sum, our data show similar variability to the WCS data, but with Spanish terms (as described above) mixed in with ad hoc terms. Notably, we observed the modal term for a few chips to be loanwords from Spanish, in some cases already established as part of the SK vocabulary (the last seems to be the case of “naranja,” “orange” in English), suggesting some fairly extensive borrowing of Spanish words due to the fact that both languages are commonly used in these studied communities.

Study 2

After generating an updated SK color term map using the responses from adult participants in Study 1, we designed Study 2 to assess child participants’ production and comprehension of SK color terms. Because we did not think that we could feasibly ask children across a range of ages about more than 100 color chips, we selected a subset of chips representing the prototypical instances for prominent SK terms from Study 1. We considered prototypical chips to be those that presented a hue towards the center or middle of the color category in question and so were supposed to be more salient or identifiable as exemplars of said category.

Table 1. ***Demographics of participants in Studies 2 and 3***

Age Group	Study 2		Study 3	
	n	Boys	n	Boys
5	3 (5% of overall sample)	1	2 (4% overall)	1
6	8 (14%)	3	2 (4%)	0
7	12 (21%)	4	11 (24%)	4
8	15 (26%)	5	9 (20%)	1
9	10 (18%)	5	11 (24%)	4
10	4 (7%)	2	8 (17%)	3

Methods

Participants

Fifty-seven children (23 boys) ages 5-11 were recruited in predominantly SK neighborhoods in Yarinacocha (Nueva Era and Bena Jema) and in Bawanisho, a native community settled along the Ucayali River, more than 500 kilometers southeast of Pucallpa. Recruitment occurred either through direct contact with interested parents or through their local school. If recruited via school, consent for participation had to be given by both teacher and parent. Outside of the school environment, consent was given by the parent.

Materials and procedure

Based on the findings of Study 1, we chose 8 color chips from our original set of 330 to serve as prototypical instances of major SK color terms. These color chips were blue (WCS n°1), green (n°234), red (n°245), white (n°274), yellow (n°297), black (n°312), greenish-yellow (WCS n°320), and purple (WCS n°325). Study 2 was conducted entirely in SK and participants were explicitly instructed to give responses in SK as opposed to Spanish. In the production and comprehension tasks, children sat at a table across from the experimenter with color chips arranged between them. The production task was always performed before the comprehension task. All administrations were video recorded.

Production task

Similar to Study 1, the experimenter introduced the participant to the general procedure and the goals of the study. The experimenter would then ask: “What is the color of this chip?” As in Study 1, we used follow-up questions to elicit a basic level term when the child’s initial response was not one. In a departure from Study 1, we were more explicit in soliciting an SK-language response. When a participant provided a Spanish-language term, the experimenter would record their response but further ask: “What is the name of this color in SK?” If a participant could not respond with an SK term, the experimenter would not ask further questions and would move forward to the next chip. As a result, some children only produced SK non-basic level terms or Spanish-language terms for particular chips.

Comprehension task

The comprehension task had a notably different procedure compared to the preceding production task. We tested the comprehension of 9 SK color terms. The choice of these terms was based on common responses given by adult participants in Study 1. The color term prompts included basic level terms: “yankon” (green/blue), “joshin” (red), “panshin” (yellow), “joxo” (white/light), “wiso” (black/dark). We also included non-basic but prominent terms as prompts which were “nai” (sky or sky blue), and “barin poi” (greenish-yellow, meaning the Sun´s excrement, also used to refer to an alga) and two dyads of non-basic terms: “pei” (leaf) and “xo” (unripe) to represent the color green, along with “ami” (a type of tree used to dye fabrics) and “pua” (sachapapa, a tuber) to represent purple. Children sat at a table across from the experimenter with the 8 color chips of the production task displayed between them. The experimenter asked: “Can you give me the [color term] chip?” Participants chose one of the 8 chips and their response was recorded.

Our findings from Study 1 suggested that color terms varied in their degrees of specificity. For example, “wiso” best describes a narrow range of very dark to black. By contrast, “yankon” could encompass blue, green, greenish-yellow, and purple; “joshin” could describe red, purple, and orange; “pei” or “xo” could describe green or greenish-yellow. In cases where a term could apply to multiple chips (i.e., “yankon”), the chip selected first would be removed from the table, leaving 7 remaining chips. The experimenter would then ask: “Can you give me another [color term] chip?” The participant would then pick another one of the 7 chips, have their response recorded, and so on. We prompted participants 4 times for “yankon” and 2 times each for “joshin” and “pei”/“xo”; every other term only received a single prompt. Due to the inherent ambiguity in term-hue pairings, accuracy for a child participant was coded based on adult responses given during Study 1. If at least 15% of adult participants in Study 1 associated a chip with a particular term, we coded a similar term-chip pairing from a child participant as correct. Some trials could have multiple pairings; in those cases, accuracy was scored as an average, rather than as dichotomous. For instance, if a child correctly chose 3 out of 4 chips for the “yankon” trial, instead of 1 (correct) or 0 (incorrect) they would receive a score of 0.75.

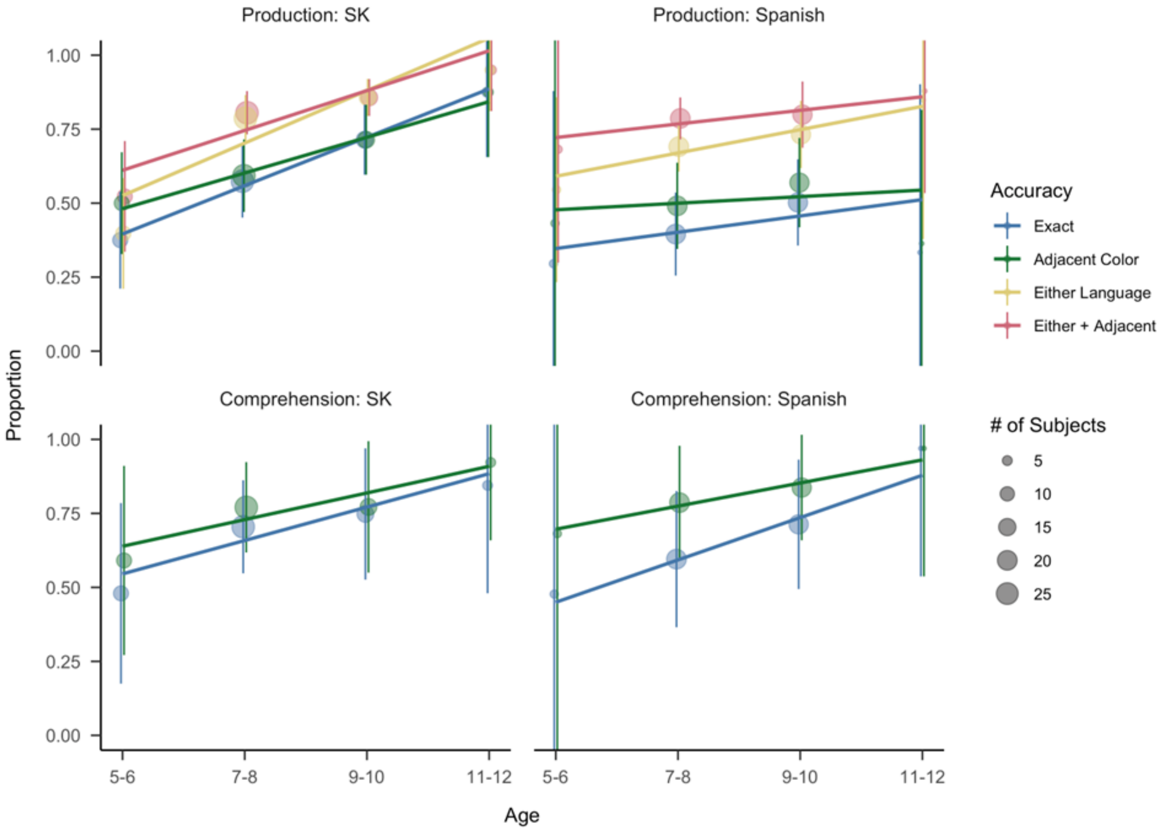


Figure 2. ***Proportion of accurate responses when applying different accuracy criteria, by age and study. Points show the mean for a 2-year age group (chosen arbitrarily for visualization) with 95% confidence intervals. Lines show a linear fit, weighted by the number of datapoints in each age group.***

Results and Discussion

We begin by presenting general results from both the production and comprehension tasks, and then turn to specific analyses of overextensions. Figure 2 shows general trends across measures. For Study 2, we saw robust developmental changes in both production and comprehension towards more adult-like performance. Because we had limited expectations regarding the amount of data that would be gathered during visits to the SK, we did not preregister our analyses. Thus, all reported inferential statistics should be interpreted with some caution, and we do not adopt a specific cutoff of $\alpha = .05$ for interpretation.

To quantify these trends, we fit generalized linear mixed-effects models (GLMMs) with a binomial link function using the *lme4* R package (Bates et al., 2015). These models predicted accuracy for both production and comprehension tasks with fixed effects of age in years (centered), random slopes of age for each color, and random intercepts for each color and participant. When these models failed to converge, we removed random slopes. We found highly significant age effects for both production ($\hat{\beta} = 0.85$, 95% CI [0.46,1.24], $z = 4.26$, $p < .001$) and comprehension ($\hat{\beta} = 0.36$, 95% CI [0.18,0.54], $z = 3.85$, $p < .001$).[9] Most children in our study knew some SK color words, but few except some of the oldest children knew all of them.

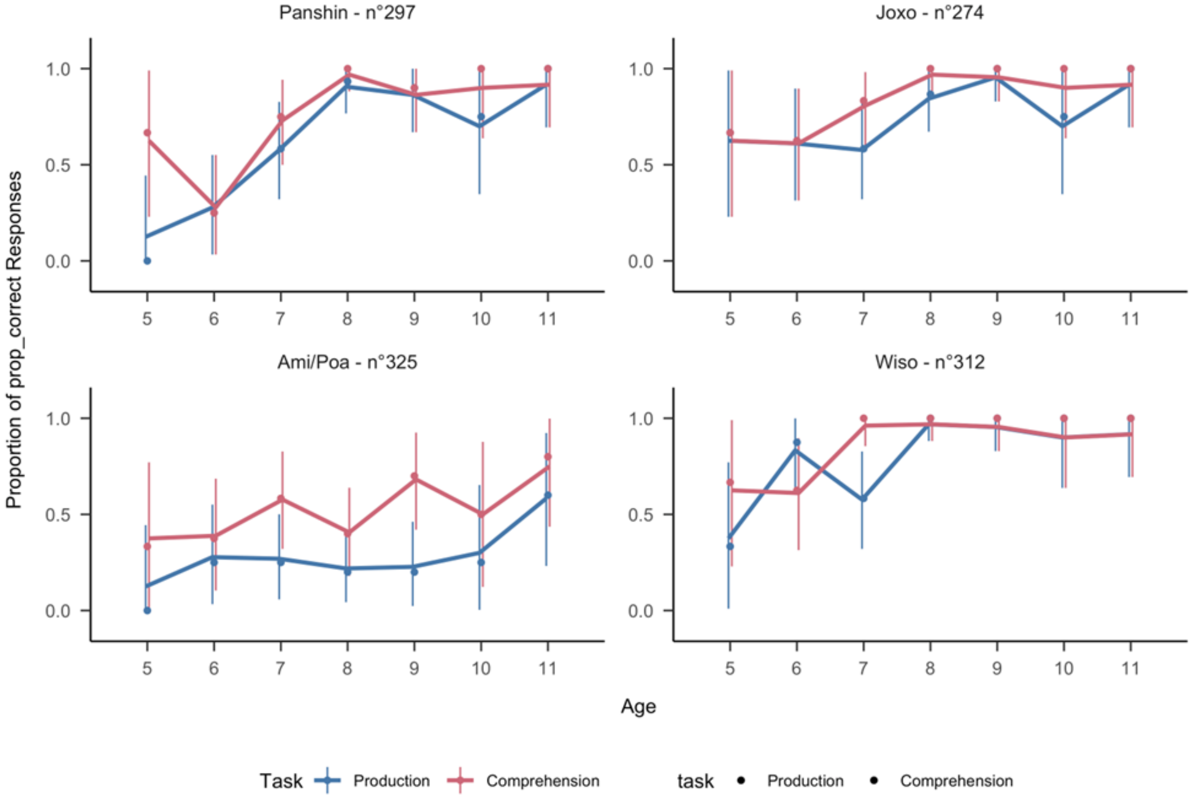


Figure 3. ***Production and comprehension data for selected color chips, plotted by age group. Points show the raw average for a 2-year age group. Lines and error bars represent the 95% confidence intervals.***

Production vs. comprehension

[9] Coefficients β denote log odds change associated with a predictor; z-values are derived from the coefficients and the estimated standard error.

While, overall, production and comprehension accuracies were quite close, there were exceptions. For some term-chip pairings such as “ami/pua” and “pei/xo,” children failed to produce the correct term in the production task but performed substantially better during the comprehension task (Figure 3). While there was a consistent ordering of tasks (production always first), there was no feedback on the production task, thus we think it is unlikely that children learned (or remembered) these labels as a function of task order. More likely is that these labels are relatively lower frequency and some children recognized them despite being unable to produce them.

Age of Acquisition

Following Frank et al. (2021), we used the dichotomous responses given during the production task to predict the “age of acquisition” when at least half of SK children are predicted to properly label a particular chip. First, we split responses by the prompted chip for which each participant had a single entry. For each chip, we attempted to fit a generalized linear model using the robust Mallows' quasi-likelihood estimator within the *robustbase* R package with the structure `accuracy of response [0 or 1] ~ age` (Cantoni & Ronchetti, 2001; Maechler et al., 2020). The choice of robust regression follows the earlier work, which showed that the resulting estimates were less sensitive to outliers. The coefficients for age ranged from 0.33 (odds of success multiplied by $\exp(\hat{\beta})$ = 1.40 with every added year of age) to 1.35 (odds multiplied by $\exp(\hat{\beta})$ = 3.80). To find age of acquisition, we then predicted the probability of success for the range of participant ages–5 to 12 years at increments of 0.05 years–and selected the earliest age at which the accuracy crossed 0.5.

Using this method, we predict that half of SK children first learn to label the “joxo” chip (white) at 5.4 years of age. This is followed by the “wiso” chip (black) at 5.5, the “hoshin” chip (red) at 6.2, the “panshin” chip (yellow) at 7.2, the “yankon” chip (green) at 7.8, the “nai” chip (sky-blue; “yankon” also accepted) at 9.4, and the “yankon”/“panshin” chip (greenish-yellow) at 9.5. The model for one chip (“purple”) did not predict that age of acquisition would have been met within our age range, with an estimated probability of 46% of children successfully labelling at 11.5 years of age.[10]

Our predictions suggest that SK children obtain color term knowledge at notably older ages compared to monolingual English-speaking children in the United States at the current time (Wagner et al., 2013). Although this comparison is of course confounded in many ways, English data suggest acquisition in early childhood (ages 2-3) while the current study suggests that SK children’s learning extends well into middle childhood. However, as there are large cross-cultural differences between the two groups, it is

[10] It is worth noting that, in Study 1, adult participants used 7 different labels for this chip (*ambi, ami, jimi, joshin, kari, morado,* and *yankon*), none of which were used more than 25% of the time.

possible that SK children's age of term acquisition differs more broadly between term comprehension and production compared to previous estimates with monolingual English-speaking children. Further, the ordering of acquisition is substantially different from that attested in previous studies (where the English colors "white" and "black" are learned later than "blue," "green," and "red"). It is an interesting question what properties of children's input or the color terms themselves lead to this order of acquisition. Following Yurovsky et al. (2015), we might speculate about the potential that "joxo" is substantially higher frequency in SK than "white" is in English.

Language switching

Over a quarter (27%) of all responses were given in Spanish, despite children being prompted solely in SK (i.e., labeling a *panshin*-colored chip as "amarillo"). The distribution of Spanish responses was non-random, with median use in 2/8 trials (IQR = 0-5 trials). We did not find a significant correlation between age and number of trials with Spanish-language responses throughout the production task ($t(55) = -0.97, p = .335$).

Table 2. *Naming entropy by color chip and whether the chip was used in Study 2 and Study 3*

Chip ID	Entropy	Study 2	Study 3	Shipibo term	Spanish term
1	0.71	×		Nai	Celeste
46	1.72		×	-	Gris
65	1.21		×	-	Rosa
121	1.49		×	-	Naranja
234	0.00	×	×	Pei/Xo	Verde
245	0.21	×	×	Joshin	Rojo
266	0.82		×	-	Marron
274	0.33	×	×	Joxo	Blanco
291	0.90		×	-	Azul
297	0.21	×	×	Panshin	Amarillo
312	0.80	×	×	Wiso	Negro
320	1.34	×		Barin Poi	Mierda sol
325	1.94	×	×	Ami/Pua	Morado

As a further exploratory analysis, we attempted to assess whether low naming consensus amongst adult SK speakers may be linked to children's naming strategies by quantifying naming entropy (following Gibson et al., 2017). Entropy is a measure of uncertainty: higher entropy corresponds to a broader distribution of labels, while lower entropy refers to a more focused distribution. We computed the naming entropy for each chip by computing the probabilities for each chip c to be named with a particular label l ($p(l \mid c)$) and then taking $H(c) = -\sum_l p(l \mid c)\log[p(l \mid c)]$ (see entropy values by chip in Table 2). For example, chip #1 had 14 adult subjects describe its color as "yankon" (blue/green), "nai" (sky), or "azul" (blue in Spanish), resulting in an entropy score of 0.71. However, chip #325 had much less consensus on its label, resulting in 7 different color terms, leading to an entropy score of 1.94. To assess the hypothesis that naming entropy in adults was related to Spanish use in children, we fit a GLMM as above to predict likelihood of switching languages from SK to Spanish (a binary variable) as a function of child age, entropy of the chip's naming distribution for adults in Study 1, and their interaction (as well as random effects of subject). Despite age not being very correlated with overall frequency of Spanish responses, within this model, we found a trending but ultimately non-significant trend of older children being less likely to respond in Spanish ($\hat{\beta} = -0.44$, 95% CI [−0.96,0.07], $z = -1.69$, $p = .092$). Children were significantly more likely to respond in Spanish when presented with a chip with greater entropy (low naming consensus) among adult participants in Study 1 ($\hat{\beta} = 1.70$, 95% CI [1.15,2.24], $z = 6.10$, $p < .001$). We found a marginal, but non-significant positive interaction between age and entropy ($\hat{\beta} = 0.30$, 95% CI [−0.03,0.62], $z = 1.78$, $p = .074$), suggesting greater Spanish use from older children for chips with low adult agreement. Together these findings suggest that children rely on language-switching to describe chips which lack consensus among adults.

Overextensions

One reason to use Spanish would be if children fail to recall the proper SK color term but do know the proper mapping in Spanish. But another possibility is that children may have more imprecise representations and choose to respond with a same-language but adjacent color term (i.e., labeling a *panshin*-colored chip as "joshin"). Following Wagner et al. (2013), we aggregated across color chips and examined the pattern of children's first responses, categorizing them as same-language, adjacent, and different-language. We used a GLMM to assess whether calculated word entropy and age were associated with frequency of adjacent responses. We predicted the outcome using fixed effects of age in years (centered) and entropy, with random intercepts by participant. We found that younger children were more likely to respond with SK-language adjacent terms ($\hat{\beta} = -0.96$, 95% CI [−1.58, −0.34], $z = -3.02$, $p = .002$) but chip entropy did not predict this strategy ($\hat{\beta} = -1.38$, 95% CI [−3.06,0.29], $z = -1.62$, $p = .106$). Further, coefficients in this model were almost identical to the coefficient for strict scoring, confirming the impression that these overextensions were relatively rare compared to the use of Spanish terms.

Study 3

Noting the apparent strategy of language switching from SK to Spanish seen in Study 2, we designed Study 3 as its complement. Here, we tested children's production and comprehension of Spanish color terms with a similar protocol to Study 2 but with a different set of chips meant to represent the prototypical basic colors within the Spanish color system. Our goal was to more directly probe SK children's knowledge of the Spanish language and its color term lexicon as well as to observe whether children would employ language-switching as a strategy similar to what was seen in Study 2.

Methods

Participants

We recruited a separate sample of 46 children (16 boys) ages 5-11 from the neighborhood of Bena Jema in Yarinacocha and from Bawanisho. Recruitment occurred either through interested parents or a local school. As in Study 2, we received consent from parents and, if in a school environment, teachers as well.

Materials and procedure

Based on Study 1, we selected 11 color chips to serve as prototypical instances of prominent Peruvian Spanish color terms. These color chips included 6 also used during Study 2: green (n°234), red (n°245), white (n°274), yellow (n°297), black (n°312) , and purple (n°325). Five additional chips were selected: gray (n°46), pink (n°65), orange (n°121), brown (n°266), and blue (n°291; see Appendix 1). The blue chips differed between Studies 2 and 3 as we decided that the prototypical hues for *yankon* and *azul* differed enough to warrant the use of a different chip.

As we found that many SK children in our sample were not very fluent in Spanish – despite receiving some school instruction in Spanish – the production and comprehension tasks were both conducted in SK, and Spanish was only used for color terms (i.e., Spanish color terms were embedded within otherwise SK sentences). In both tasks, a participant would sit at a table across from the experimenter with 11 color chips in front. As in Study 2, the production task was always performed prior to the comprehension task.

Production task

The procedure was similar to that of both Studies 1 and 2. The experimenter would introduce a participant to the general procedure and aims of the study. Despite much of the study being conducted in SK, the experimenter would specify that participants would be expected to provide color terms in Spanish. The experimenter would then ask: “What is the color of this chip?” If the participant responded in SK, the experimenter would record their response but further ask: “What is the name of this color in Spanish?” If a participant responded with “I don’t know” to this prompt, the experimenter would not prompt any further and would move forward to the next chip. As a result, some responses lack Spanish-language basic level terms and only consist of non-basic and/or SK color terms. In total, we collected production data for 11 color chips. For each chip, the data include either one response (when children provided a Spanish basic color term in the first trial) or two to three responses (when children’s initial responses were either non-basic and/or in SK).

Comprehension task

The procedure was similar to that of Study 2. The experimenter would ask: Can you give me the [*color term*] chip? (for 11 Spanish color terms) The choice of these terms was based on both previous studies examining Spanish color terms as well as responses given by adult participants in Study 1 (as some adult participants used Spanish color terms to label particular color chips). The 11 terms used as prompts were “blanco” (white), “verde” (green), “rojo” (red), “amarillo” (yellow), “azul” (blue), “negro” (“black”), “naranja” (orange), “gris” (gray), “morado” (purple), “marrón” (brown), and “rosa” (pink). Since each color term was best instantiated by a single color chip and lacked the ambiguity seen with certain SK color terms, we defined a correct response as choosing the single color chip that matched the word, in contrast to Study 2.

Results and Discussion

As in Study 2, we observed age-related changes in color term accuracy for both production and comprehension. Aggregate results are visualized in Figure 2. To assess these, we again fit GLMMs for both production and comprehension tasks with an identical structure to Study 2. Age was a significant predictor of accuracy in the comprehension task ($\hat{\beta} = 0.63$, 95% CI [0.21,1.06], $z = 2.90$, $p = .004$), but the age effect weakened in the production task ($\hat{\beta} = 0.33$, 95% CI [-0.06,0.72], $z = 1.65$, $p = .098$, see Figure 2).

As in Study 2, over a quarter (30%) of all responses were given in SK. While participants were reminded to give responses in Spanish, their initial consent/assent was conducted in SK which could have predisposed some participants to respond in SK. There was significant variation in language-switching with some children responding solely in Spanish while others responded in SK for upwards of 9/11 trials (*Mdn* = 5 trials, *IQR* = 1.25-6). We found only a marginal correlation between age and accuracy ($t(44) = 1.91$, $p = .063$) and no significant correlation between age and language-switching ($t(44) = 0.44$, $p = .663$).

To assess our hypothesis that older children would have more Spanish-language exposure and color term knowledge, we included age as a predictor in our GLMM assessing the effect of adult color naming entropy on likelihood to switch languages from Spanish to SK, similar to the one we fit for Study 2. This model did not show a significant interaction between age and adult color naming entropy ($\hat{\beta} = -0.27$, 95% CI $[-0.63, 0.09]$, $z = -1.49$, $p = .137$), however one without the interaction term did show an entropy effect ($\hat{\beta} = -1.49$, 95% CI $[-2.07, -0.92]$, $z = -5.10$, $p < .001$), tending to respond in SK for low-entropy items (those that were presumably more prototypical for the SK words). There was no significant effect of age ($\hat{\beta} = -0.02$, 95% CI $[-0.49, 0.45]$, $z = -0.08$, $p = .939$). Across studies, it appears that children preferred to respond in SK when presented with a chip for which adults had high consensus about the SK label, and in Spanish for low-consensus chips.

Similar to Study 2, we adopted alternative scoring to accommodate language-switching from Spanish to SK (different-language) and adjacent same-language responses. We used a GLMM identical to that of Study 2 in order to assess whether changes in scoring criteria were associated with significant changes in task performance for production. Age was again a weaker predictor for production accuracy even with this more lenient scoring ($\hat{\beta} = 0.25$, 95% CI $[-0.07, 0.58]$, $z = 1.53$, $p = .126$), in concordance with earlier analyses. However, we did find that participants had greater accuracy when we included SK responses ($\hat{\beta} = 1.76$, 95% CI $[1.43, 2.08]$, $z = 10.46$, $p < .001$) or adjacent same-language responses ($\hat{\beta} = 0.51$, 95% CI $[0.20, 0.81]$, $z = 3.27$, $p = .001$). In sum, we find frequent use of language switching in both Studies 2 and 3, but only Study 3 exhibits significant use of same-language but adjacent terms.

We speculate that early, informal Spanish language exposure can explain the discrepancies seen in Studies 2 and 3. With limited knowledge of Spanish color terms, children may spontaneously refer to their Spanish color term knowledge during SK-language Study 2 but struggle to succeed in a more systematic evaluation in Study 3. More generally, we see children relying on a mixture of strategies to communicate colors even in the absence of mastery in either language.

General Discussion

In three studies, we mapped the color vocabulary of the Shipibo-Konibo (SK) language and used these data to study the development of color vocabulary in SK children growing up in a bilingual environment. This effort contributes to filling the gap in studies of color word development in non-WEIRD cultures by analyzing the case of a bilingual population involving an Amazonian language and Spanish, and more generally parallels other efforts to use methods from language development to study populations that are under-represented in developmental science (Fortier et al., 2023; e.g., Piantadosi et al., 2014).

With respect to the adult data, we found that the SK color vocabulary was relatively unchanged over the generations since the original WCS. Several interesting observations emerged, however. First, consistent with our review of the prior literature, there was substantial use of non-basic color terms (including both ad hoc and luminance-based terms). These terms were used more often in SK than in Spanish, supporting the idea that Amazonian languages have been suggested to make greater use of ad hoc color terms (at least in naming tasks) than Spanish (Everett, 2005). Our data do not speak to whether this use is due to a desire to succeed on specific experimental tasks or whether it is comparable to use in naturalistic contexts. Nevertheless, our findings are reminiscent of a suggestion by Levinson (2000) that even purported basic level terms in Yélî Dnye did not fully span hue space and were often supplemented creatively with ad hoc terms.

Second, we saw substantial use of Spanish terms by adults, even though the task was conducted in SK. We speculate that this is because the adults were recognizing focal colors for Spanish basic level terms that have no parallel in SK (e.g., “morado” for purple). On the other hand, “naranja” could in fact be a loan word that has been assimilated into the SK vocabulary by some speakers. Either way, this finding suggests an adaptive use of color vocabulary from both languages to succeed on the labeling task; future work will be required to understand whether such strategies are used in naturalistic communication as well.

When we turned to the children’s data, we observed a much longer developmental trajectory for color word learning than is observed in contemporary English-learning children within the United States. As noted by Bornstein (1985), however, it is a very recent development that color terms are mastered as early as they are – one hundred years ago, English-speaking US children’s timeline of acquisition looked broadly similar to that observed in our study for SK children. We can only speculate as to the drivers of this historical change, but the industrialization hypothesis propounded by Gibson et al. (2017) appears to be a reasonable starting point. That is, industrialization allows for the production of identical objects that are usefully distinguished by color labeling. This communicative pressure can then lead to differentiation of color terms on a historical timescale and – relevant to our study here – is a likely driver of faster acquisition of color words by children.[11] SK children have some access to such artifacts, but according to anthropological accounts it is substantially sparser. Although we have not found previous studies on access to industrialized toys specifically in the SK population, research on other Amazonian groups such as the community dwelling in Rio Araraiana (Estado do Pará, Brasil) points to children making their own toys with seeds and wood (Castro dos Reis et al., 2012). Further, we note that SK children are

[11] These speculations are informed by personal experience; the children of one author both learned their color terms in their second years of life through repeated practice with sets of manufactured plastic artifacts that varied only in hue, providing ideal teaching examples.

bilingual and so likely receive less color word input in either language. Although bilingual vocabulary acquisition is typically similar in trajectory to monolingual acquisition, there may be asymmetries between languages based on exposure (Thordardottir, 2011).

We did not find strong evidence for overextension in children's SK production or comprehension (with one or two exceptions), though there was somewhat more evidence for overextension in Spanish. This asymmetry might be due to less systematic or consistent exposure to Spanish vocabulary, as Beekhuizen and Stevenson (2018) suggested that color term frequency may influence developmental errors in discrimination. We did, however, observe robust evidence for mixing and competition between the SK and Spanish color systems. Children differentially used Spanish terms in Study 2 when there was high uncertainty about the SK label for a particular color chip among adults in Study 1. Similarly, they reached into their SK vocabulary in Study 3 when there was high consistency in SK labels among adults. These findings suggest that children were using their bilingual vocabulary adaptively to choose terms that are more likely to be interpreted correctly. Further, they suggest a potential route for functionally-driven language change, such that Spanish terms are borrowed – and perhaps eventually conventionalized – by children in cases where adult input indicates uncertainty about the appropriate SK label.

Comprehension is thought to proceed production in language development generally (Clark, 2009; Frank et al., 2021) and in color word learning specifically (Wagner et al., 2018). In our data we did not observe large asymmetries between comprehension and production, a surprising finding given prior literature.[12] Production and comprehension may be especially divergent for the youngest children, those who have the most difficulty with phonological encoding and the motoric aspects of production (Frank et al., 2021); there is less evidence for production-comprehension divides in middle childhood. One natural question is whether comprehension and production dissociated in earlier times when US English-learners similarly acquired colors late; unfortunately we do not know of data that could be used to evaluate this question.

Our data here are consistent with models of color word meaning in which color word use is driven by functional need and languages adapt by developing vocabularies that appropriately allow for communication about those needs (Gibson et al., 2017; Zaslavsky et al., 2018). These models have not yet been generalized to either the bilingual setting or the acquisition setting, however. Our data suggest that functional language

[12] One caveat is that comprehension and production tasks are by their nature different and have different demands (Sandhofer & Smith, 1999); this was especially true in our case given that the two tasks were sequenced and scored somewhat differently. Thus, we have chosen not to make quantitative comparisons between accuracies across these two tasks.

use can cross language boundaries, inviting models that consider code switching and borrowing as part of the process of change (e.g., Myslin & Levy, 2015).

Studying SK children's learning provides a descriptive comparison to studies of color naming in children learning English in the US (the focus of the majority of developmental work). Nonetheless, it has a number of limitations, some shared with this previous literature and some due to the specifics of our study and context. First, we regrettably do not have access to the kind of deep ethnographic observations that would allow us to hazard generalizations about how color terms are used in daily life (and whether they are primarily used in Spanish or SK) among the SK communities we studied. Second, our study of development is cross-sectional and does not afford precision regarding the specific knowledge state of individual children due to the limited length of the task; further, due to data collection issues we could not sample children younger than age five. Third, the limited number of color chips that we investigated means that our ability to generalize about the precision of particular color generalizations is much more limited for the children than the adults (limiting our entropy analyses). Finally, and perhaps most prominently, the kinds of tasks that we used are likely more unfamiliar to all of our participants and especially our child participants than they are to the populations being tested in investigations of WEIRD cultures (e.g., US English-learning children). While the performance of the oldest children in our studies was close to ceiling, the lower performance observed with younger children could be in part a product of task unfamiliarity or other factors.

Going beyond convenience populations in experimental research with children is a new frontier for developmental science (Nielsen et al., 2017). Our work here suggests some of the benefits and challenges of this approach. On the positive, we can compare and generalize models of acquisition that are largely based on a single language and population (US English-acquiring children). At the same time, there is a paucity of resources describing language use, home environment, and cultural practices once we venture outside of WEIRD contexts. To best understand acquisition across cultures, we must document both children's knowledge and the structure of their environments.

References

Aragón, K. (2016). *Color language and color categorization* (G. Paulsen, M. Uusküla, & J. Brindle, Eds.). Cambridge Scholars.

Bartlett, E. J. (1977). Semantic organization and reference: Acquisition of two aspects of the meaning of color terms. Biennial Meeting of the Society for Research in Child Development.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beekhuizen, B., & Stevenson, S. (2018). More than the eye can see: A computational model of color term acquisition and color discrimination. *Cognitive Science, 42*(8), 2699–2734. https://doi.org/10.1111/cogs.12665

Berlin, B., & Kay, P. (1969). *Basic color terms: Their universality and evolution*. University of California Press.

Bornstein, M. H. (1985). On the development of color naming in young children: Data and theory. *Brain and Language, 26*(1), 72–93.

Bornstein, M. H. (2015). Emergence and early development of color vision and color perception. In *Handbook of Color Psychology*. Cambridge University Press.

Bornstein, M. H., Kessen, W., & Weiskopf, S. (1976). Color vision and hue categorization in young human infants. *Journal of Experimental Psychology: Human Perception and Performance, 2*(1), 115–129.

Cantoni, E., Ronchetti, E. (2001). Robust Inference for Generalized Linear Models. *Journal of the American Statistical Association, 96*(455), 1022-1030.

Castro dos Reis, D., Freire Monteiro, E., Ramos Pontes, F. A., & Souza da Costa Silva, S. (2012). *Brincadeiras em uma comunidade ribeirinha amazônica* [Play in an Amazonian riverine community]. *Psicologia: Teoria e Pratica, 14*(3), 48–61.

Chater, N., & Christiansen, M. H. (2010). Language acquisition meets language evolution. *Cognitive Science, 34*(7), 1131–1157.

Clark, E. V. (1973). *Cognitive development and acquisition of language* (pp. 65–110). Academic Press.

Clark, E. V. (1987). *Mechanisms of language acquisition* (B. MacWhinney, Ed.; pp. 1–33). Psychology Press.

Clark, E. V. (2009). *First language acquisition*. Cambridge University Press.

Culbertson, J., Smolensky, P., & Legendre, G. (2012). Learning biases predict a word order universal. *Cognition, 122*(3), 306–329.

Everett, D. L. (2005). Cultural constraints on grammar and cognition in pirahã another look at the design features of human language. *Current Anthropology, 46*(4), 621–646.

Forbes, S. H., & Plunkett, K. (2019). Infants show early comprehension of basic color words. *Developmental Psychology, 55*(2), 240.

Forbes, S. H., & Plunkett, K. (2020). Linguistic and cultural variation in early color word learning. *Child Development, 91*(1), 28–42.

Fortier, M., Kellier, D., Fernández Flecha, M., & Frank, M. C. (2023). Ad-hoc pragmatic implicatures among Shipibo-Konibo children in the Peruvian Amazon. Language Development Research, *3*(1), 105—120. https://doi.org/10.34842/2023.76

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability and consistency in early language learning: The wordbank project*. MIT Press.

Frank, M. C., & Goodman, N. D. (2014). Inferring word meanings by assuming that speakers are informative. *Cognitive Psychology, 75*, 80–96.

Franklin, A., Pilling, M., & Davies, I. (2005). The nature of infant color categorization: Evidence from eye movements on a target detection task. *Journal of Experimental Child Psychology, 91*(3), 227–248.

Gibson, E., Futrell, R., Jara-Ettinger, J., Mahowald, K., Bergen, L., Ratnasingam, S., Gibson, M., Piantadosi, S. T., & Conway, B. (2017). Color naming across languages reflects color use. *Proceedings of the National Academy of Sciences, 114*(40), 10785–10790.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2-3), 61–83.

Kay, P., Berlin, B., Maffin, L., Merrifield, W. R., & Cook, R. (2009). *The world color survey*. Center for the Study of Language; Information.

Kristol, A. M. (1980). Color systems in southern italy: A case of regression. *Language, 56*(1), 137–145.

Lathrap, D. W. (1970). *The Upper Amazon*. Thames; Hudson.

Levinson, S. C. (2000). Yélî dnye and the theory of basic color terms. *Journal of Linguistic Anthropology, 10*(1), 3–55.

Lillo, J., González-Perilli, F., Prado-León, L., Melnikova, A., Álvaro, L., Collado, J. A.,

& Moreira, H. (2018). Basic color terms (BCTs) and categories (BCCs) in three dialects of the Spanish language: Interaction between cultural and universal factors. *Frontiers in Psychology, 9*.

Maechler, M., Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T., Koller, M., Conceicao, E. L. T., & Anna di Palma, M. (2020). *Robustbase: Basic robust statistics* (R package version 0.93-6).

Monroy, M., & Custodio, S. (1989). Algunos usos de los términos del color en el español de Colombia [some uses of color terms in Colombian Spanish]. *Thesaurus: Boletín Del Instituto Caro y Cuervo, 44*(2), 441–450.

Morin, E. (1973). Le paradigme perdu: La nature humaine [the lost paradigm: The human nature]. Éditions du Seuil.

Myers, T. P. (1974). Spanish contacts and social change on the Ucayali River, Peru. *Ethnohistory, 21*(2), 135–137.

Myslin, M., & Levy, R. (2015). Code-switching and predictability of meaning in discourse. *Language, 91*(4), 871–905.

Nielsen, M., Haun, D., Kärtner, J., & Legare, C. H. (2017). The persistent sampling bias in developmental psychology: A call to action. *Journal of Experimental Child Psychology, 162,* 31–38.

Piantadosi, S. T., Jara-Ettinger, J., & Gibson, E. (2014). Children's learning of number words in an indigenous farming-foraging group. *Developmental Science, 17*(4), 553–563.

Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences, 104*(4), 1436–1441.

Saji, N., Asano, M., Oishi, M., & Imai, M. (2015). How do children construct the color lexicon?: Restructuring the domain as a connected system. *Proceedings of the 37th Annual Conference of the Cognitive Science Society.*

Sandhofer, C. M., & Smith, L. B. (1999). Learning color words involves learning a system of mappings. *Developmental Psychology, 35*(3), 668–679.

Scott, M. E., Kanero, J., Saji, N., Chen, Y., Imai, M., Golinkoff, R. M., & Hirsh-Pasek, K. (2023). From green to turquoise: Exploring age and socioeconomic status in the acquisition of color terms. *First Language, 43*(1), 3–21.

Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of Psycholinguistic Research, 32*(1), 3–23.

Shin, S. J. (2017). *Bilingualism in schools and society: Language, identity, and policy* (2nd ed.). Routledge. https://doi.org/10.4324/9781315535579

Simunovic, M. P. (2010). Colour vision deficiency. *Eye, 24*(5), 747–755.

St. Clair, K. (2016). *The secret lives of colour*. John Murray.

Surrallés, A. (2016). On contrastive perception and ineffability: Assessing sensory experience without colour terms in an Amazonian society. *Journal of the Royal Anthropological Institute, 22*(4), 962–979.

Thordardottir, E. (2011). The relationship between bilingual exposure and vocabulary development. *International Journal of Bilingualism, 15*(4), 426–445.

Tournon, J. (2002). La merma mágica: Vida e historia de los Shipibo-Conibo del Ucayali [the magic reduction: Life and history of the Ucayali Shipibo-Conibo]. Centro Amazónico de Antropología y Aplicación.

Wagner, K., Dobkins, K., & Barner, D. (2013). Slow mapping: Color word learning as a gradual inductive process. *Cognition, 127,* 307–317.

Wagner, K., Jergens, J., & Barner, D. (2018). Partial color word comprehension precedes production. *Language Learning and Development, 14*(4), 241–261.

Yurovsky, D., Wagner, K., Barner, D., & Frank, M. C. (2015). Signatures of domain-general categorization mechanisms in color word learning. *Proceedings of the 37th Annual Conference of the Cognitive Science Society*.

Zaslavsky, N., Kemp, C., Tishby, N., & Regier, T. (2018). *Color naming reflects both perceptual structure and communicative need* (Vols. 1250–1255). Proceedings of the 40th Annual Conference of the Cognitive Science Society.

Data, code, and materials availability statement

De-identified data and all analytic code are available on GitHub at https://github.com/langcog/amazon_color.

Ethics statement

Our protocol for Studies 1, 2, and 3 received ethical approval from the Pontificia Universidad Católica del Perú's institutional review board. We chose to use a short consent form based on advice that many SK participants would be unfamiliar with the consent process. Before recruitment, we received approval from the community authorities for each site which was contingent on their conversations with other community members during weekly meetings. MF recruited participants based on both recommendations from community authorities and also through directly approaching community members. He then orally informed potential participants of the overall study tasks and duration, compensation, and that participation was entirely voluntary and could stop at any time. If community members were still interested, they scheduled a later time to participate. At the beginning of each session, we received consent from adult participants in Study 1, and parental consent and participant assent for Studies 2 and 3. For child participants recruited within a school, we received additional consent from the supervising teacher. As all sessions were video recorded, participant consent varied based on the participant's literacy and comfort. Adults who were relatively literate were asked to give written consent, which involved giving a signature. In cases where the participant had difficulty understanding the consent form or felt more comfortable with not having to write, participants gave oral consent which was documented by video. Child assent was always obtained verbally.

Authorship and Contributorship Statement

All authors approved the final version of the manuscript and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

- Conceptualization: Martin Fortier, Danielle Kellier, María Fernández-Flecha, and Michael C. Frank.
- Data Curation: Martin Fortier, Danielle Kellier, and Michael C. Frank.
- Formal Analysis: Danielle Kellier and Michael C. Frank.
- Funding Acquisition: Michael C. Frank.
- Investigation: Martin Fortier, Danielle Kellier, María Fernández-Flecha, and Michael C. Frank.
- Methodology: Martin Fortier, Danielle Kellier and Michael C. Frank.
- Project Administration: Martin Fortier, Danielle Kellier, María Fernández-Flecha, and Michael C. Frank.
- Software: Danielle Kellier and Michael C. Frank.
- Supervision: Martin Fortier, María Fernández-Flecha, and Michael C. Frank.
- Visualization: Danielle Kellier and Michael C. Frank.
- Writing: Martin Fortier, Danielle Kellier, María Fernández-Flecha, and Michael C. Frank.

Author's Note

This paper is dedicated to the memory of Martin Fortier.

Acknowledgements

We are especially grateful to Paul Kay for his assistance both conceptually and materially throughout the study. We further thank Roberto Zariquiey, Katie Wagner, and David Barner for their insightful feedback during the writing process. DK was supported by an NIH research grant (3U24NS115691-01S2). MCF was supported by a Jacobs Foundation Fellowship.

License

No evidence that age affects different bilingual learner groups differently: Rebuttal to van der Slik, Schepens, Bongaerts, and van Hout (2021)

Joshua K. Hartshorne
Department of Communication Sciences & Disorders
MGH Institute of Health Professions, USA

Abstract: Hartshorne, Tenenbaum, and Pinker (2018, A critical period for second language acquisition: Evidence from 2/3 million English speakers. *Cognition*, 117, 263-277) presented the first direct estimate of how the ability to learn the morphosyntax of a second language changes with age, showing a sharp decline in late adolescence. Recently, van der Slik, Schepens, Bongaerts, and van Hout (2021, Critical period claim revisited: Reanalysis of Hartshorne, Tenenbaum, and Pinker (2018) suggests steady decline and learner-type differences. *Language Learning*) purport to show that in fact Hartshorne et al's (2018) data are better explained by a gradual decline in learning with age, at least for some types of learners. However, these conclusions are based a misunderstanding of their own analyses, which in fact do not test whether the decline in learning is sharp or gradual but whether it is asymmetric, slowing with time. After correcting conceptual and mathematical errors in their analyses, the results strongly confirm the original conclusions of Hartshorne and colleagues: every type of bilingual investigated shows a sharp drop in learning rate in late adolescence.

Keywords: critical periods; sensitive periods; L2

Corresponding author: Joshua K. Hartshorne, Department of Communication Sciences & Disorders, MGH Institute of Health Professions, Boston, MA 02129, USA. Email: joshua.hartshorne@hey.com.

ORCID ID: https://orcid.org/0000-0003-1240-3598

Citation: Hartshorne, Joshua K. (2024). No evidence that age affects different bilingual learner groups differently: Rebuttal to van der Slik, Schepens, Bongaerts, and van Hout (2021). *Language Development Research, 4*(1), 431–454. http://doi.org/10.34842/y8sz-4q83

Introduction

One hardly needs to conduct an experiment to show that people who begin learning a second language (L2) as an adult rarely if ever reach the same level of proficiency as those who start in early childhood, though plenty of experimental data do exist (Birdsong, 2018; Flege, 2019; Hartshorne, 2020b). What remains highly controversial is *why*: is poor achievement by later learners due to differences in neural plasticity or motivation, differences in the input, interference from a first language, or something else?

One way of constraining these possibilities is to measure the age at which learning rate starts to decline: if learning success begins to decline at age X, it is probably due to something that happened at age X, not at age Y.[1] Unfortunately, until recently there were no effective estimates of how learning rate changes with age. Measuring learning in the laboratory proved fruitless: During the initial stages of learning, older learners actually learn second languages faster (Asher & Price, 1967; Chan & Hartshorne, in press; Ferman & Karni, 2010; Krashen, Long, & Scarcella, 1979; Snedeker, Geren, & Shafto, 2012; Snow & Hoefnagel-Höhle, 1978). Decades-long longitudinal studies would work better but have not been done. Another approach, developed in the 1960s, is to find the oldest age at which someone can start learning a language and still reach native-like proficiency (Asher & García, 1969; Johnson & Newport, 1989). This method is limited by its inherent ambiguity: finding that (for instance) people who started learning a language at age 8 do better than those who started at 10 does not tell you much about when that difference appeared: the latter group might start off more slowly, or they may start off just as fast but fall behind after 3 years. Or 5. Or 10. In fact, it can be shown that such data do not constrain theory much if at all (for a more thorough exposition, see Appendix B).

Hartshorne, Tenenbaum, and Pinker (2018) (henceforth HTP) addressed the limitations reviewed above by applying a novel analytic model to a massive dataset of English morphosyntactic knowledge of 669,498 native and non-native English speakers, including monolinguals, simultaneous bilinguals, and second-language learners who either learned in an English-speaking country (“immersion learners”) or not in an English speaking country (“non-immersion learners”). Morphosyntactic knowledge of each subject was assessed by a comprehensive 132-item self-paced written grammaticaly judgment and usage test (a complete list of stimuli are available in HTP’s supplementary materials). Critically, the model (described below) disentangles how learning ability changes with age from other factors, including ceiling effects and years of exposure. The results indicated that the rate at which learners acquire English mor-

[1] Arguably, the rapidity with which learning rate declines is also informative. However, such arguments rest primarily on intuition, and intuitions can vary. For instance, while it is generally argued that a rapid decline in learning rate is most consistent with biological causes and that a slower decline would suggest social causes, a reviewer of this article made the opposite argument.

phosyntax declines substantially at around 17-18 years old, followed by an increasingly gradual decline into old age (Fig. 3).

Recently, van der Slik, Schepens, Bongaerts, and Hout (2021) (henceforth, *SSBH*) have challenged this conclusion. They compare HTP's model to an alternative model that they assert models age-related declines as gradual, not sharp. They report that their model "had a better fit when applied separately to monolinguals, bilinguals, and early immersion learners. Only for nonimmersion learners and later immersion learners did [HTP's] model have a better fit" (p. 87). They conclude that HTP's "overall conclusion of one sharply defined critical age at 17.4 for all language learners is based on artificial results" (p. 87).

In this paper, I show that these conclusions are based on conceptual confusions and mathematical errors. Critically, SSBH's alterrnative model does not require age-related declines in learning to be gradual. In fact, in every one of SSBH's analyses, their model either finds qualitatively the same result as HTP's original model or fits less well. Ironically, it is only for the late immersion learners and non-immersion learners that it finds a decline more gradual than what is inferred by HTP's model, but those are the cases in which HTP's model fits substantially better. Below, I show this with graphs and explain mathematically why this would be the case.

Regardless, the difference in results across different learner groups is mostly artifactual. Both HTP's and SSBH's models are designed to predict variability in learning outcomes based on age at which one starts learning the language. When applied to groups where everyone started learning at the same age (e.g., monolinguals or simultaneous bilinguals), they have difficulty distinguishing effects of experience from effects of age, complicating any conclusions.

Below, I first lay out the logic of HTP's model in more detail than can be found in the original paper. I then use this foundation to explain how SSBH's model works and how it differs from HTP's. It emerges that the difference between the models is not in whether one is "continuous" or "discontinuous" (the terms used by SSBH) but rather in terms of symmetry (HTP's model can detet a decline that is initially rapid and then slows down; SSBH's cannot). In this context, I re-present and re-evaluate SSBH's results, finding they strongly support HTP's original conclusions. However, it is conceivable that this finding is an artifact in limitations of SSBH's method. Thus, I test SSBH's hypotheses using a more precise model and a larger dataset. If anything, the results only more strongly support HTP's conclusions and militate against those of SSBH. These analyses also address the question that SSBH inadvertently tested, showing that the decline is indeed asymmetric: initially sharp and then more gradual.

A close look at HTP's model

HTP start by analyzing learning curves: knowledge as a function of years spent learning. Intuitively, if native speakers are more successful at learning a language than are late-L2 learners, their learning curves should be different. The fact that native speakers ultimately learn more means that their learning curves reach a higher asymptote. It is possible that the learning curves are steeper as well (they learn faster). By comparing learning curves for learners who began at different ages, it is possible to mathematically infer how learning changes with age, type of exposure, etc.

We describe the mathematical inference process below. First, we note that one important innovation of HTP was to empirically measure these learning curves. That is, they had enough monolinguals who had been speaking English for different lengths of time to empirically plot the monolingual learning curves. The same was true for simultaneous bilinguals, immersion learners who started at ages 1-3, and so on (Fig. 3A).

They note that, particularly for native speakers and early-L2 learners, the learning curve follows a very clear exponential decay (Fig. 1). Exponential decay is extremely common in natural processes. In this case, what it means is that how much a language learner learns at any given time depends on how much is left to learn. Formally, they model grammatical knowledge g at time t given that one started learning at time t_e (time of exposure) as:

$$g(t) = 1 - e^{\int_{t_e}^{t} -r dt} \tag{1}$$

where r is the learning rate. Note that if r is constant, the integral reduces to $(t - t_e)r$. Note that dt indicates that we are integrating over t.

This formula fits monolingual data almost perfectly by assuming that in each year, monolinguals learn a constant 13% of what is left to learn (Fig. 1, left). That is, they learn 13% in the first year, 11% in the second year [(100% - 13%) * 13%], 10% in the third year [(100% - 13% - (100% - 13%)*13%) * 13%], and so on. This is an asymptotic process and never quite ends, though after a certain amount of time the changes become negligible.

However, the curve reaches very different asymptotes for different learners: highest for monolinguals (Fig. 1, left), somewhat lower for simultaneous bilinguals (Fig. 1, center), and even lower for some groups of L2 learners (e.g., Fig. 1, right). More generally, HTP show that while there are main effects of whether the learner is monolingual or bilingual and whether they learned in an immersion or non-immersion setting, the learning curves are otherwise indistinguishable regardless of the age at which learning began, up to about 10 years of age: not only are the asymptotes the same, but the curve is just as steep (see also Fig. 3). For learners who begin later, however, the later they started learning English, the shallower the decay rate and the lower the asymptote.

One might suspect this could be accounted for by assuming different types of learners learn at different rates. Perhaps monolinguals are simply faster learners than are late L2 learners. HTP considered a model where r was fit separately for monolinguals, simultaneous bilinguals, immersion L2 learners, and non-immersion L2 learners. Formally, HTP introduced a parameter E, which was set to 1 for monolinguals and allowed to vary between 0 and 1 for each of the other three groups:

$$g(t) = 1 - e^{\int_{t_e}^{t} -Erdt} \tag{2}$$

HTP dubbed this E the "Experience discount factor", reflecting the intuition different learner groups get different amounts of exposure to English. However, this is an interpretation: mathematically, it simply means the learning rate is different across groups.

HTP report that this is insufficient, something we reproduce here with slightly different math: forcing the learning curves for different learner groups to share the same asymptote results in very poor fits (Fig. 1, dashed red lines). In intuitive terms, it is not just that simultaneous bilinguals or later learners learn more slowly than monolinguals, but that they stop learning before reaching monolingual proficiency.

HTP showed that this could be well-explained by assuming the learning rate (formally, variable r in the exponential decay curve) decreases as the learner ages. Specifically, they modified the exponential decay model above to one where the learning rate starts at a relatively high value. Specifically, they assumed that r is initially constant through some "critical" age t_c, after which it declines according to a sigmoid (an s-shaped curve). Thus, learning rate r at time t is given by:

$$r(t) = \begin{cases} r_0 & t \leq t_c \\ r_0(1 - \frac{1}{1+e^{-\alpha*(t-t_c-\delta)}}) & t > t_c \end{cases} \tag{3}$$

where r_0 is the initial learning rate, t_c is the critical age, and α and δ are parameters governing the steepness of the sigmoid and the location of its decline, respectively.

Critically, assuming that the decline in learning rate is sigmoidal was not so much a theoretical assumption as a lack of one. Sigmoids can decline slowly and gradually or

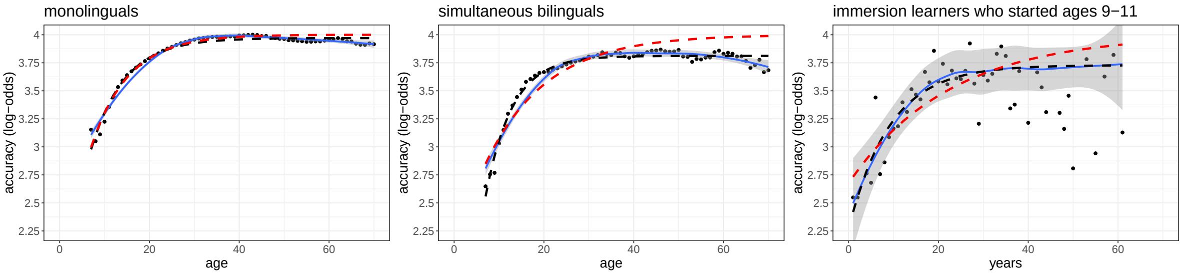


Figure 1. ***Y-axis shows accuracy on HTP's test using the empirical logit scale. Data is from HTP (Hartshorne, 2020a). Data for monolinguals (N=244,840; left), simultaneous bilinguals (N=30,347; center), and learners who began at ages 9-11 and learned in an immersion/immigration setting (N=1,373; right) are each shown. Performance is averaged by year (dots), with a LOESS curve in blue and a 95% confidence interval shown in gray. Results were fit to an exponential decay model with free parameters for intercept, asymptote, and rate (dashed black line) or with the parameter for rate fixed to be the same across all three learner groups (dashed red line). It is clear that the latter provides a poor fit.***

in a sharp step function (Fig. 2, Top) or not at all (simulating no age-related change). The decline can start at any point from birth to not at all.[2]

Sigmoids are mathematically convenient, having few parameters and being integrable. The latter is critical, since we must be able to perform integrals over r (see Equation (1)). Moreover, the parameters α and δ allow sigmoids to vary in how sharply they decline and where on the x-axis the decline occurs (Fig. 2, top row).

The is one crucial limitation to a sigmoid: It is symmetric. That is, if the decline begins quickly, it must reach floor quickly; if it reaches the floor slowly, the decline must begin slowly. By requiring r to initially be a constant (r_0) up until some age t_c, HTP circumvented this issue. This allows for declines that start rapidly but then level off (Fig. 2, second and third rows). Note that if $t_c = 0$, then the formula reduces to a standard, symmetric sigmoid. That is, HTP's model does not assume that learning declines quickly and then the decline levels off; it merely includes that as one of the hypotheses being tested.

Thus, by fitting the parameters to the data, HTP inferred the shape of age-related decline. As shown in Fig. 3, the model finds that the data were best explained by a sharp drop in learning rate at 17-18 years of age, followed by a more gradual decline. This fits the data quite well (compare Fig. 3 A&B with C&D). Not surprisingly, the model

[2] Strictly speaking, sigmoids are declining at every point along the x-axis, but for most of that span it is so mild as to be negligible. The exception is the degenerate case where $\alpha = 0$ and there is no decline at all.

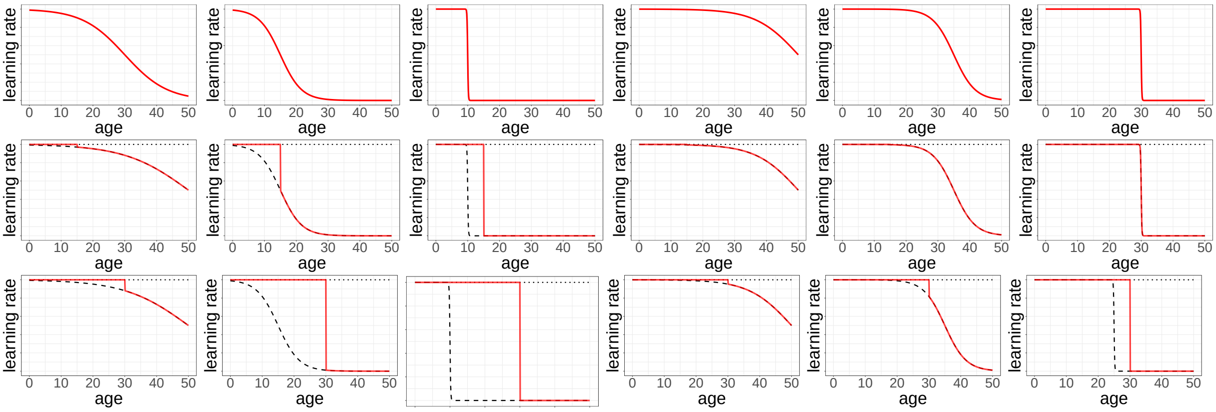


Figure 2. ***Top row: Sigmoids can decline more or less sharply (compare within the three left panels or within the three right panels) and can decline earlier or later (the three right panels are shifted versions of the three left panels). Second and third rows: HTP augmented sigmoids by composing them with a straight line, joined at age t_c. In these panels, the straight line is shown as a dotted line and the sigmoid is shown as a dashed line. The composed curved is outlined in red. Columns: Each column involves the same underlying line and sigmoid, but with different values of t_c, which is either 0 (top row), 15 (second row), or 30 (bottom row). Note that this figure was designed to illustrate the differences between HTP and SSBH and does not cover the full range of possible curves; for more examples, see HTP or Chen and Hartshorne (2021).***

fit much less well if learning rate (r) was forced to be constant across ages ($R^2 = 66\%$ vs. $R^2 = 89\%$; see Fig. 3, F, G, and H).

The fact that learning rate drops substantially with age explains two key empirical findings. Most importantly, it means that learning that happens in adulthood is quite slow, changing the effective asymptote. Although the model fit indicates that simultaneous bilinguals learn almost as fast as monolinguals (E is slightly less than 1), they are just a little behind. As a result, when the age-related decline kicks in at around 18 years old, they lose the ability to catch up. Later-learners are doubly-affected: their inferred learning rate (the product Er) is substantially lower than monolinguals', and they have fewer years of learning before speed slows and the asymptote lowers.

The second key empirical finding fit by the model is that learning curves learning curves continue to become progressively more shallow the later the learner began, even for learners who began learning after the critical age t_c (comp. Fig. 3 I & C). This cannot be explained if learning rate falls to a floor at t_c. Indeed, HTP report the results of such a model (a "step-function model"), and found that this fit less well, precisely because it fails to capture differences among learners who began learning at different intervals after t_c (Fig. 3 I, J, and K).

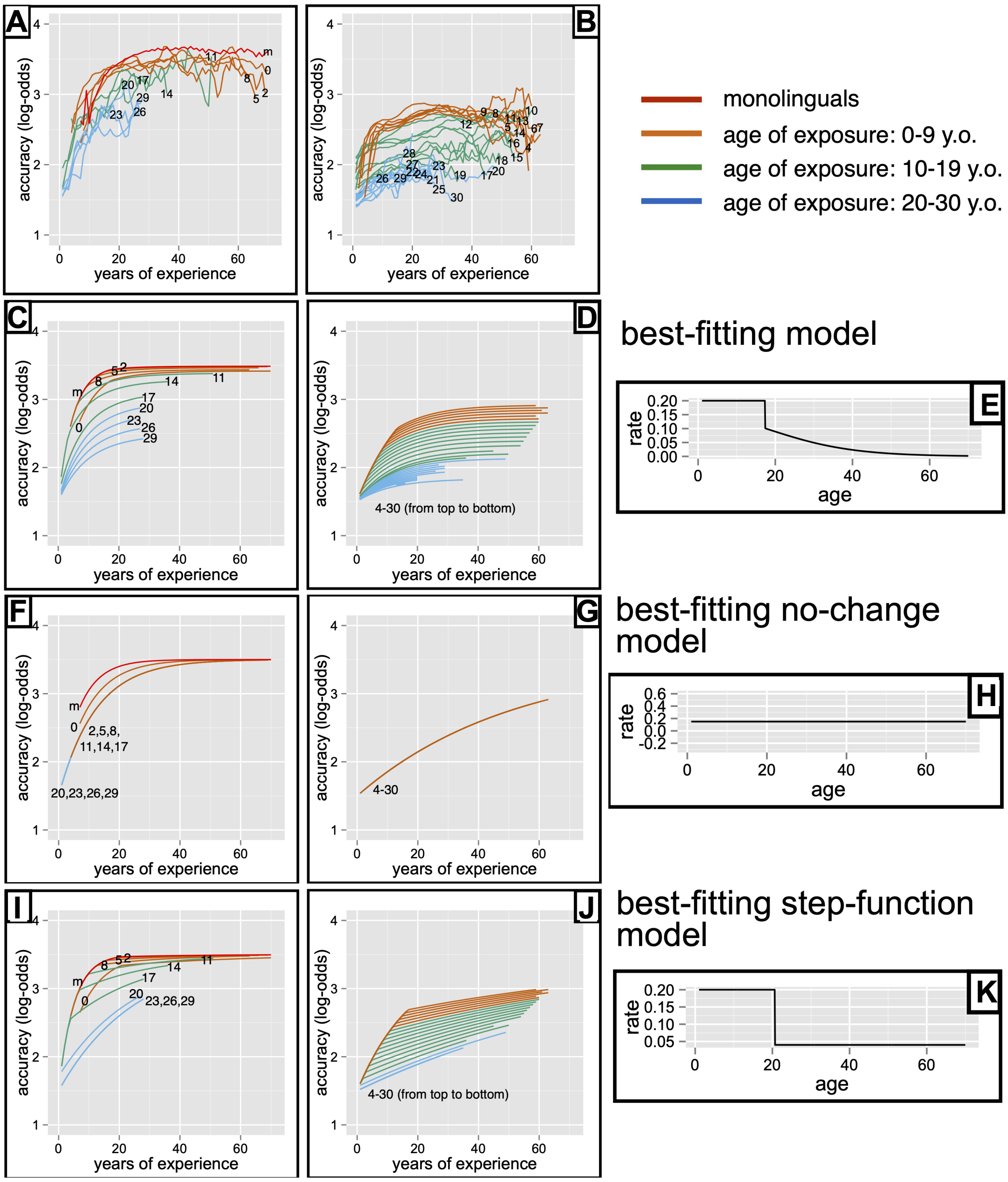


Figure 3. ***Figures from HTP, used with permission. Panels show the empirical results (top row), and the best fits for HTP's model (row 2nd from top) and two alternative models (bottom two rows). Monolinguals and immersion learners are plotted in the left panels (*A, C, F, & I*). Non-immersion learners are shown in the middle column (*B, D, G, & J*). In both the left and center columns, data/fits are plotted in terms of years of experience with the language, which makes the contrasts between models easier to see. Finally, panels in the right column show the models' estimated learning rate (r) as a function of age.***

Understanding SSBH's alternative analyses

As previewed in the introduction, van der Slik et al. (2021) (henceforth, *SSBH*) claim to show that HTP's findings are based on "artificial results" (p. 87). Specifically, they report that a sharply-defined drop in learning rate is true only for non-immersion learners and immersion learners who began learning at 10+ years of age ("late immersion learners"). They support this claim by comparing the fit of HTP's model to that of a "continuous" model independently for each of five learner groups: monolinguals, simultaneous bilinguals, early immersion learners (age of acquisition < 10), late immersion learners (age of acquisition >= 10), and non-immersion learners, finding that the continuous model fits the first three groups better while HTP's model fits the latter two better. They write, "The early immersion learners now share their continuous model with the monolingual and [simultaneous] bilingual learners. The later immersion learners share their discontinuous model with the nonimmersion learners, with a similar age boundary of 19.0 years" (p. 101).

In fact, their analyses do not support their claims and actually strongly support HTP's. There are a number of issues, but at the heart are a conceptual confusion and an empirical error.

The first conceptual confusion is that their "continuous" model does not assume gradual, continuous change in learning rate, nor does HTP's model assume a sharp discontinuity. As a result, which model fits better does not, by itself, speak to SSBH's question. (I return to what question it does speak to below.) Specifically, SSBH's "continuous" model is HTP's model where t_c is fixed at 0 (or sometimes fixed at 1; this varies across analyses for unexplained reasons; see Appendix C). By fixing t_c to 0, SSBH are forcing age-related decline to follow a sigmoidal shape.[3] Critically, as shown in the previous section, this does not constrain the decline to be early or late, or gradual or sharp. In fact, it does not even constrain there to be any decline at all (a degenerate sigmoid is flat horizontal). As a reminder, $t_c = 0$ in the first row of Fig. 2, and yet the third and sixth panels show fairly sharp declines in learning rate. Moreover, it is not the case that $t_c > 1$ means the decline is sharp. In the second row of Fig. 2 $t_c = 15$, yet the first and fourth panels show mild, gradual declines.

Thus, the model comparisons reported by SSBH do not address the question of when the age-related decline begins or how sharp it is, but rather whether the data are fit better if $t_c = 0$ or $t_c > 0$. Recall that HTP introduced the free parameter t_c to expand the range of possible age-related decline curves beyond strictly sigmoidal shapes. In particular, HTP wanted to include the possibility of a sharp decline followed by a more gradual decline — something that is impossible with a sigmoid only. Indeed, this is

[3] It is a little more complicated when $t_c = 1$, which could allow for a sharp drop in learning at the age of 1. However, one does not really require a study to know that this is not the case, and indeed it is not what any model finds.

exactly what HTP report when fitting their model to the full dataset: a sharp decline at 17.4 years old, followed by a more gradual decline. Note that SSBH criticize HTP's use of model-fitting rather than model-comparison, since the latter method takes into account number of parameters and HTP's model has an additional free parameter (t_c). However, in this case, they get the same result: HTP's model if preferred by several orders of magnitude, even correcting for the additional parameter (see SSBH Table 1 and surrounding text).

To recap: SSBH's analyses amount to asking whether including the t_c parameter provides a sufficiently better fit to justify the extra parameter. (It almost always fits the data better.) However, this does not by itself say anything about the sharpness of the decline. Making that determination requires looking at the actual inferred age-related change curves. Doing so paints a picture diametrically opposed to SSBH's conclusions (this is the empirical error).

SSBH do not include figures of the age-related change curves, so they are plotted them in Fig. 4. (Note that SSBH's paper contains a number of calculation errors. While these do not change the qualitative results, I use the corrected numbers throughout; see Appendix C for details.) While the full model (with t_c as a free parameter) provides at least as good a fit in all cases, this fit is not sufficiently better to justify the additional parameter for monolinguals (log-likelihood = 45.90 vs. 45.80; $AIC_{diff} = 1.90$; corrected relative log-likelihood: 2.44:1, favoring SSBH), simultaneous bilinguals (log-likelihood = 71.30 vs. 71.30; $AIC_{diff} = 2$; corrected relative log-likelihood: 2.73:1, favoring SSBH), and early immersion learners (log-likelihood = 112.92 vs. 112.86; $AIC_{diff} = 1.88$; corrected relative log-likelihood: 2.55:1, favoring SSBH).[4] However for simultaneous bilinguals and early immersion learners, this is a distinction without much of a difference: in both cases, there is a fairly sharp drop at around the same age regardless of whether one looks at HTP's free-t_c or SSBH's fixed-t_c model. Indeed, given that the two models fit the data roughly equally well, this is exactly what one should expect. SSBH's fixed-t_c model is very slightly smoother, but this is an artifact of the fact that during curve-fitting, SSBH forced the sigmoid sharpness parameter to be no greater than 1.0; had they relaxed that restriction, the decline would have been sharper.[5] In any case, the practical differences here are trivial (see right-hand side of Fig. 4)

The situation for monolinguals departs even further from SSBH's assertion of a "con-

[4] Akaike's Information Criterion [AIC; Akaike (1974)] is commonly used to compare models with different numbers of parameters. Formally, $AIC = 2k - 2 * log(likelihood)$, where k is the number of free parameters in the model. A reasonable rule of thumb is to choose the model with fewer parameters unless the more complex model improves the AIC by at least 4 (Burnham & Anderson, 1998).

[5] HTP's model achieves a sharper drop by use of the t_c parameter. Without a variable t_c parameter, the only way SSBH's model can achieve a sharp drop is through adjusting the sharpness of the sigmoid. This again illustrates that the relationship between the parameters and the theoretically-relevant curves is non-trivial.

tinuous" decline, in that SSBH's fixed-t_c model finds a sharp drop at around 50 years old, whereas HTP's free-t_c model finds no age-related change at all. This difference is an artifact of the fitting procedures used for SSBH: although in theory a sigmoid can be a straight line, SSBH restricted the parameter values such that this is unachievable (for details, see footnote).[6] This finding recapitulates what I showed in the previous section: monolingual learning curves are well-fit by exponential decay.

Does the finding with monolinguals nonetheless support SSBH's contention that there is no "sharply delimited critical period for normally developing monolinguals" (p. 102)? Not really. The models we are are discussing here try to estimate changes in learning rate due to *age* deconfounded from changes due to *years of experience*. Recall that we expect the total amount learned each year to decline as there is less and less left to learn (this is formally implemented as exponential decay). The problem is that since monolinguals all started learning English at the same age (0), age and amount of experience are full confounded and cannot be disentangled. By analogy, suppose a researcher wanted to investigate how children's height changes with age, and so measured the heights of a large number of children on their 5th birthdays. The researcher would find that age was completely unrelated to height, but only because the dataset was constructed that way. The models gamely try to disentangle age and experience anyway, but they have little to go on.

In principle, though, a decline in learning rate should still be detectable as a deviation from perfect exponential decay. In practice, this is very difficult. As shown in Fig. 1, by the mid-20s, both monolinguals simultaneous bilinguals are very close to ceiling. Thus whether or not there is a learning-rate decline in late adolescence will have at best subtle effects (see Fig. A1), making it difficult to detect muich less time exactly. We return to this issue below and show that more precise analyses with a larger dataset in fact suggest a decline starting in late adolescence, same as for other bilingual groups.

As alluded to above, HTP's free-t_c model fits better than SSBH's fixed-t_c model, even after correcting for number of parameters, for both the late immersion learners (log-likelihood = 127.53 vs. 72.39; AIC_{diff} = 108.29; corrected relative log-likelihood: $10^{23.52}$:1, favoring HTP's free-t_c model) and for non-immersion learners (log-likelihood

[6] As described in Fig. 2, the sigmoid shape parameters allow the decline to move left or right. This is governed by the variable δ in Eq. (3), which is the midpoint in the decline (the sharpness of the decline — and, hence, its effective starting point — is governed by α). Critically, the 0 point is t_c. Thus, if $t_c = 40$ and $\delta = 50$ the midpoint of the decline is at 90. In order to speed up computation, HTP limited δ to run from -50 to 50 and t_c to run from 0 to 40. This means that the decline could happen anywhere from the age of -50 to 90. Declines starting after the age of 70 do not affect our analysis, since we exclude subjects over the age of 70. (The reason we allow negative numbers for δ is that if t_c is 40, values for δ between -50 and 0 are meaningful; numbers below -50 are not. This can be easily demonstrated by experimenting with the function $plotr$ included in the reproducible manuscript.) When SSBH set t_c to 1, this restricted the ages in which the decline could happen from around -50 to around 50. Ideally, they would have expanded the available range for δ.

= 526.50 vs. 262.70; AIC_{diff} = -261.80; corrected relative log-likelihood: $10^{114.14}$:1, favoring HTP's free-t_c model) (see Fig. 4).

In summary, for every one of the five learner groups SSBH considers, their own analyses point to a sharply defined critical period. In no case do they provide statistical evidence for a "continuous" decline.

Their results do differ from HTP's in one important way, which is that the timing of the decline varies across the learner groups, with the critical period appearing earliest for late immersion and non-immersion learners, later for early immersion learners, even later for simultaneous bilinguals, and latest for monolinguals (though, as we explain above, this last fact is likely an artifact of how they fit their model).

The significance of this result is unclear. As just explained, age and experience are confounded for the latter three groups, making it difficult to accurately assess the effect of age. Nonetheless, with greater power and precision, we might be able to detect deviations from a constant learning rate even in these groups. I turn to this possibility in the next section.

Differences between Learner Groups, Redux

In order to more precisely compare age-related changes in learning rate within each learner groups, I made several improvements on SSBH's analyses. First, I used the more flexible model from Chen and Hartshorne (2021). Chen and Hartshorne (2021) provided an alternative formulation of HTP's model that allows for a wider range of age-related changes curves. While they found that this did not qualitatively change HTP's results, the model is more precise.

Second, I used the larger dataset reported by Chen and Hartshorne (2021), which has nearly half a million additional subjects, resulting in 319,565 monolinguals, 41,534 simultaneous bilinguals, 21,174 immersion learners, and 543,407 non-immersion learners.

Finally, inspired by Frank (2018), I reexamined how HTP and SSBH had defined the asymptote for exponential decay (that is, the asymptote that would be reached if learning rate remained steady). Prior analyses had assumed fixed the asymptote to 3.5, which HTP had determined by visual inspection. With the larger dataset, it was clear that this was a little too low, and that actual maximum hit by subjects is 3.66. (Frank (2018) suggests finding an analytic method that does not require specifying an asymptote. This seems like a very good idea but so far nobody has found one.)

I then fit the expanded dataset twice, either requiring the age-related changes in learn-

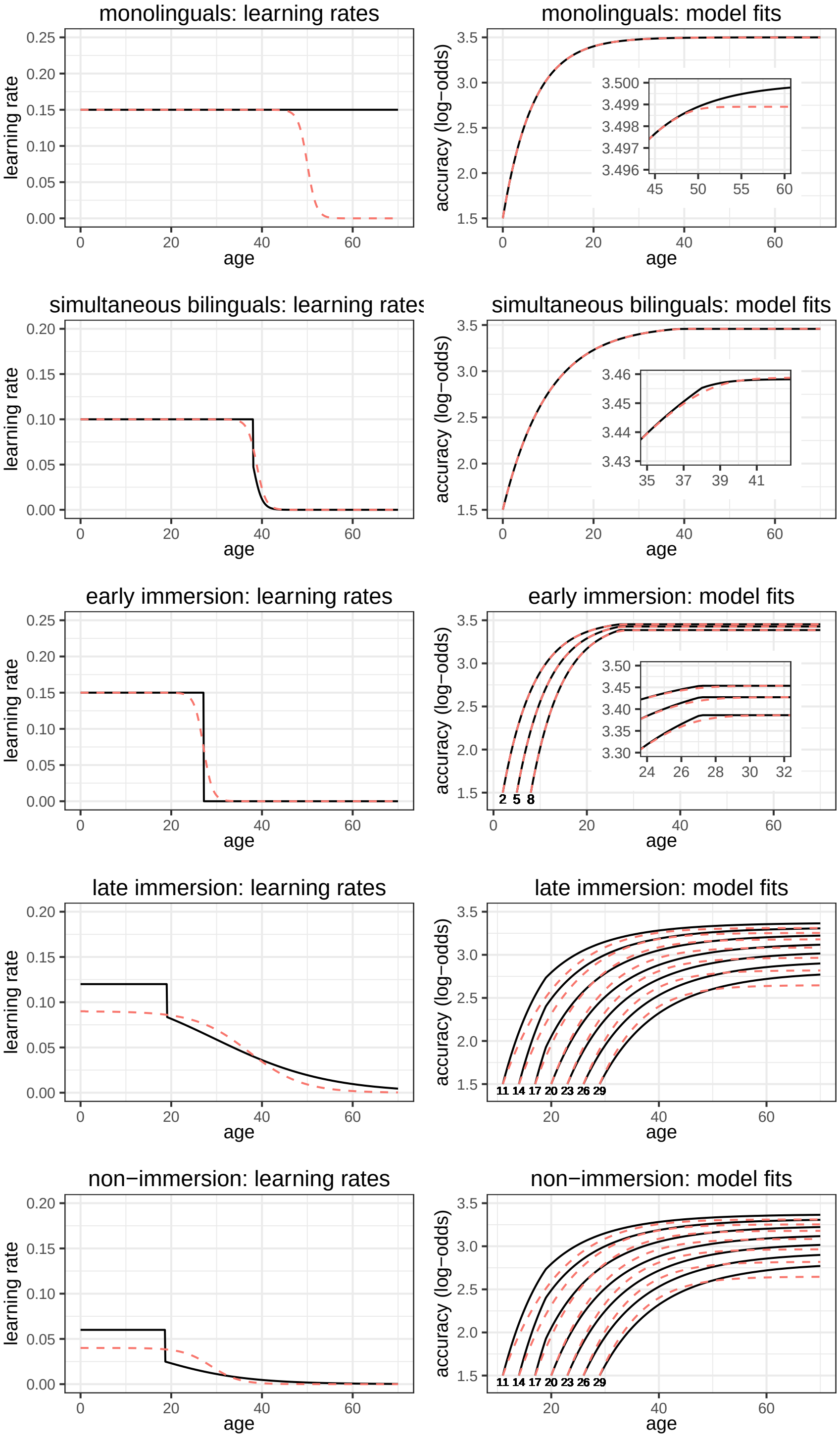


Figure 4. ***Model fits for each learner group, conducted separately.*** **Left panels*: Inferred age-related changes in learning curves.*** **Right panels*: fitted models (insets shows magnified view, where necessary).*** **Solid black*: HTP's model (t_c as a free parameter).*** **Dashed red*: SSBH's continuous model ($t_c = 1$).***

ing rate to be the same for all subject groups (Fig. 5, solid black lines) or allowing them to vary across the five learner groups defined by SSBH (Fig. 5, dashed red lines). Allowing for independent age-related change for each learner group actually led to a significantly *worse* fit after correction for number of parameters (AIC_{diff} = -14.49; relative log-likelihood: 1402 to 1). In any case, fitting each group separately nontheless reliably results in a sharp decline in late adolescence, with the sole exception of monolinguals, where the decline was much later Fig. 5. However, given that monolinguals are quite close to ceiling by late adolescence, the uncertainty about exactly where the decline is likely to be substantial, so one should be cautioned against making too much of this result even if it were significant, which it is not.

In short, there seems little reason to suppose more than minimal differences in age-related change in learning rate across the different learner groups, at least in this dataset.

Summary and Conclusions

SSBH suggest that different learner groups exhibit distinct effects of age on learning, with some groups declining continuously from birth (monolinguals, simultaneous bilinguals, early immersion learners) while others show a marked decline in adolescence (late immersion learners, non-immersion learners). However, none of their arguments or results hold up under scrutiny. If anything their results support HTP's conclusions more than their own. A revised analysis, more appropriate to SSBH's questions, supported HTP's conclusions even more strongly.

This does not mean that the case is closed. Although the dataset and the analyses support a rapid drop in morphosyntax learning ability in late adolescence, there are significant limitations. While a lot of care went into creating HTP's quiz, I doubt that any 95-question quiz can assess an infinitely expressive human grammar precisely and without bias. If such a quiz can be created, we certainly lack the theoretical understanding of morphosyntax needed to construct it at the moment. Moreover, HTP's quiz probes meta-linguistic grammaticality judgments. This is certainly an important linguistic phenomenon – the spectacular failure of late learners to acquire native-like meta-linguistic knowledge is part of what we wish to explain! – but it clearly involves cognitive mechanisms not required for other linguistic phenomena, which themselves depend on cognitive mechanisms not needed for meta-linguistic judgment. To the extent this mechanisms are themselves affected by age, the picture will depend on which phenomenon we study.

In terms of the analytic model, even the best of the models explored above do not fit the data perfectly, and they alide some known issues. It has no previsions for senescence, which turns out to begin much earlier than had been visible in HTP's data (compare Fig. 3A with Fig. 5, top), starting perhaps as early as the mid 40s. Unfortunately, our original

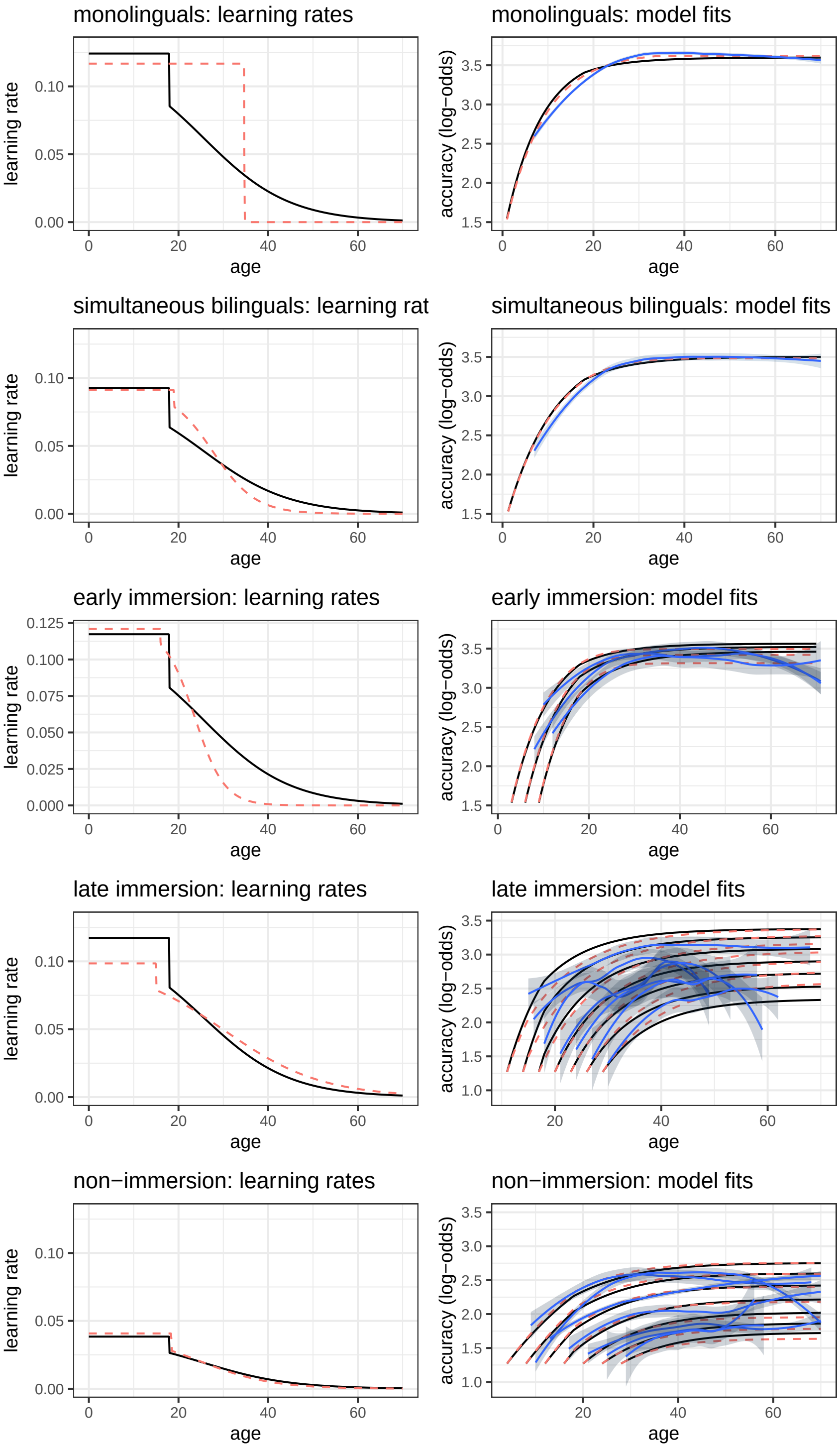


Figure 5. ***Comparisons of the revised model trained on all data (solid black lines) and on individual learner groups (dashed red lines), and LOESS-smoothed data (blue lines with shaded 95% confidence intervals). For nonimmersion learners, only a subset of curves are shown.***

strategy of simply excluding older subjects will not work: excluding subjects older than 45 would vitiate our ability to study later learners. Similarly, the model cannot entertain age-related *increases* in learning rate during childhood, even though these are clear in the present data and in prior work (Asher & Price, 1967; Chan & Hartshorne, in press; Ferman & Karni, 2010; Krashen et al., 1979; Snedeker et al., 2012; Snow & Hoefnagel-Höhle, 1978). Relatedly, the models assume that age-related change is driven by a single underlying factor. This is certainly at least somewhat wrong; in the limit, the r curves we see could be an epiphenomenon of age-related changes in many different underlying mechanisms, each of which looks quite different. Finally, the models assume learning is asymtotic, whereas Frank (2018) correctly notes that many modern theories (especially construction grammars) posit that the set of grammatical structures is a) unbounded, and b) a moving target due to language change.

More broadly, as highlighted by HTP, our analyses estimate age-related change in learning rate. They cannot speak to whether this represents biologically-determined change, age-related changes in environment, or something else. The results certainly constrain the possibilities (e.g., factors that do not change rapidly in late adolescence are unlikely to explain the results), but ultimately that evidence is indirect. We need studies directly testing the causal role of candidate influences on learning.

All of which is to say that HTP and follow-up papers (Chen & Hartshorne, 2021; Frank, 2018; Hernandez, Bodet, Gehm, & Shen, 2021; van der Slik et al., 2021) are just the start of a conversation. We will need many more studies of similar scale and scope to resolve the open theoretical questions.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723.

Asher, J. J., & García, R. (1969). The optimal age to learn a foreign language. *The Modern Language Journal, 53*(5), 334–341.

Asher, J. J., & Price, B. S. (1967). The learning strategy of the total physical response: Some age differences. *Child Development*, 1219–1227.

Birdsong, D. (2018). Plasticity, variability and age in second language acquisition and bilingualism. *Frontiers in Psychology, 9,* 81.

Burnham, K. P., & Anderson, D. R. (1998). Practical use of the information-theoretic approach. In *Model selection and inference* (pp. 75–117). Springer.

Chan, J., & Hartshorne, J. K. (in press). *Is it easier for children to learn english if their native language is similar to english?* Cascadilla Press.

Chen, T., & Hartshorne, J. K. (2021). More evidence from over 1.1 million subjects that the critical period for syntax closes in late adolescence. *Cognition, 214,* 104706.

Ferman, S., & Karni, A. (2010). No childhood advantage in the acquisition of skill in using an artificial language rule. *PloS One, 5*(10), e13648.

Flege, J. E. (2019). A non-critical period for second-language learning. *A Sound Approach to Language Matters: In Honor of Ocke-Schwen Bohn, Aarhus University. Open Access e-Book at Aurhus University Library*.

Frank, M. C. (2018). With great data comes great (theoretical) opportunity. *Trends in Cognitive Sciences, 22*(8), 669–671.

Hartshorne, J. K. (2020a). *Data: A critical period for second language acquisition: Evidence from 2/3 million english speakers*. OSF. Retrieved from osf.io/pyb8s

Hartshorne, J. K. (2020b). How massive online experiments (MOEs) can illuminate critical and sensitive periods in development. *Current Opinion in Behavioral Sciences, 36*, 135–143.

Hartshorne, Joshua K. and Tenenbaum, Joshua B., & Pinker, S. (2018). A critical period for second language acquisition: Evidence from 2/3 million English speakers. *Cognition, 177*, 263–277. https://doi.org/10.1016/j.cognition.2018.04.007

Hernandez, A. E., Bodet, J. P., Gehm, K., & Shen, S. (2021). What does a critical period for second language acquisition mean?: Reflections on Hartshorne et al. (2018). *Cognition, 206*, 104478. https://doi.org/10.1016/j.cognition.2020.104478

Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of english as a second language. *Cognitive Psychology, 21*(1), 60–99.

Krashen, S. D., Long, M. A., & Scarcella, R. C. (1979). Age, rate and eventual attainment in second language acquisition. *TESOL Quarterly*, 573–582.

Snedeker, J., Geren, J., & Shafto, C. L. (2012). Disentangling the effects of cognitive development and linguistic expertise: A longitudinal study of the acquisition of english in internationally-adopted children. *Cognitive Psychology, 65*(1), 39–76.

Snow, C. E., & Hoefnagel-Höhle, M. (1978). The critical period for language acquisition: Evidence from second language learning. *Child Development*, 1114–1128.

van der Slik, F., Schepens, J., Bongaerts, T., & Hout, R. van. (2021). Critical period claim revisited: Reanalysis of hartshorne, tenenbaum, and pinker (2018) suggests steady decline and learner-type differences. *Language Learning*.

Data, Code and Materials Availability Statement

Data and analysis code can be found at osf.io/u6fq5.

Ethics statement

This paper does not contain human subjects research.

Authorship and Contributorship Statement

JKH is solely responsible for this work.

Acknowledgements

The author thanks F. W. P. van der Slik, Tianhua Chen, David Birdsong, Lisa Pearl, Michael C. Frank, Emma Marsden, two anonymous reviewers, and the attendees of talks at the Reading/Tromso joint bilingualism meeting, the BMB Lab at UC-Irvine, the Stanford Language and Cognition Lab, the Brain and Language Lab at Georgetown University, and the Institute of Linguistics and National Taiwan University for comments and suggestions. The author declares no competing interests.

Appendices

Appendix A: Supplementary Figures

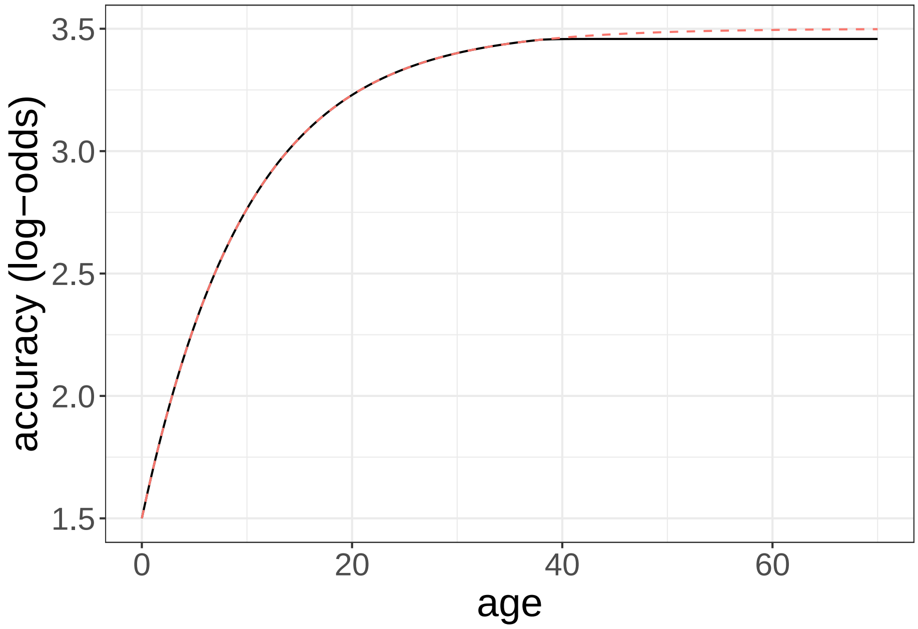


Figure A1. ***When fit to the simultaneous bilinguals only, both the SSBH and HTP models suggest a decline in learning rate at around the age of 40. Here, we plot the predicted learning curves for with (solid black) and without (dashed red) that age-related change. As can be seen, the differences are quite small.***

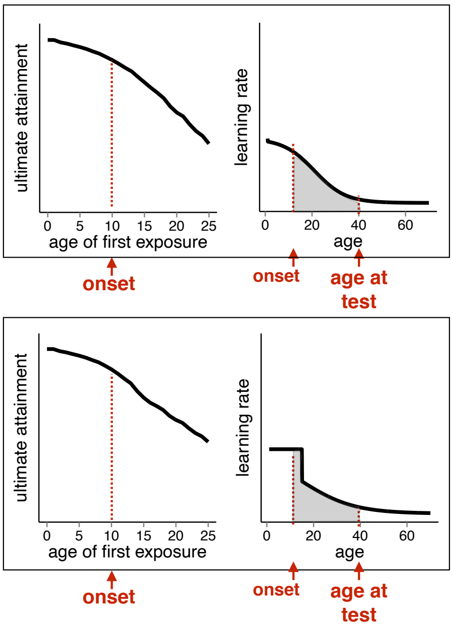


Figure A2. ***Each point on an ultimate attainment curve (left panels) is related to an integral under the learning rate curve from onset of learning through the age at test (shaded portion of right panels). Most studies have assumed that it is possible to infer the shape of the theoretically-critical learning rate curves from easier-to-measure ultimate attainment curves. However, simulations from HTP show that highly similar ultimate attainment curves (left: top vs. bottom) can actually be explained by very different learning rate curves (right: top vs. bottom).***

Appendix B: Limitations of Studying Ultimate Attainment

The oldest age at which one can start learning a new language and still reach nativelike proficiency is not (necessarily) the age at which learning rate declines. To give an intuitive example, suppose we know that if Agnes leaves her home at 8:15 in the morning, she makes it to work comfortably before 9:00. If she leaves after 8:15, she runs into a traffic jam and arrives much later. Does this mean that the traffic picks up at exactly 8:15? Perhaps. Even a slight decrease in speed, applied over the entire travel distance, could be enough to make her tardy. Alternatively, the traffic may grind to a halt at 8:45, so if Agnes hasn't arrived by then, she is out of luck. The point is that if we know what time she left home and how far she got, we know her *average* speed, but not her speed at any given point along the way.

Similarly, if Bartholomew starts learning Swahili as an adult and manages only 80% the proficiency of a native speaker, this does not mean that he started out learning more slowly than a Swahili-acquiring infant. In fact, as mentioned above, during the initial stages of learning, older learners actually learn second languages faster (Asher & Price, 1967; Chan & Hartshorne, in press; Ferman & Karni, 2010; Krashen et al., 1979; Snedeker et al., 2012; Snow & Hoefnagel-Höhle, 1978). Thus, all we know for sure is that at some point along the way, his learning rate decayed to the point where he ultimately was unable to get to the finish line.

More formally, very different age-related changes in the ability to learn language can give rise to indistinguishable ultimate attainment curves (Fig. A2).

Appendix C: Additional errors and imprecisions in SSBH

SSBH make a number of factual misstatements and mathematical errors. The following list may not be exhaustive.

SSBH use Akaike Information Criterion (AIC) for model comparison, but in almost every case appear to have miscounted the number of parameters in the models (a key part of calculating AIC). For most of their analyses, the "continuous" model has 4 free parameters (r_0, α, δ, and the error variance), though in all but one case, they count it as having 5. The "discontinuous" model has one additional free parameter (t_c) but for some reason is counted as having 7. The exceptions are as follows: In the case of the monolingual analysis, they correctly assign the continuous model 4 parameters, but again over-count the discontinuous model (6 instead of 5). When fit to all data, there are 3 additional parameters (the three E parameters), which should give the "continuous" model 7 parameters (which they code correctly) and give the "discontinuous" model 8 (they count 9).

(Note that they explain in Footnote 5 that "the discontinuous model needs to fit three components, the continuous model only one (cf. Figure 1). That explains the difference of two degrees of freedom." It is not possible to count degrees of freedom by inspecting a graph, and the numbers here do not match the numbers in their code.)

These errors tend to overstate the evidence for the "continuous" model. For instance, the relative likelihood for the monolingual analyses in their Table 3 is reported as 0.16. Using the correct number of parameters, it is 0.41. That is, using AIC correctly, rather than the "continuous" model being nearly 7 times more likely, it is only about 3 times more likely. (Strangely, using SSBH's counting of parameters, the ratio is actually 0.15; I have not yet identified the source of that error.)

As described in the main text, the "continuous" model is simply the HTP model (which they call "discontinuous") with the t_c parameter fixed. Across analyses, it is sometimes fixed to 1 and sometimes to 0. SSBH do not provide any explanation, and indeed do not even mention this variation. Inspection suggests that the choice of 1 or 0 probably does not make much difference, though I did not test this systematically. Note that strictly speaking SSBH's "continuous" model is only a special case of HTP's model when t_c is set to 1, since HTP fit HTP's model with a restriction that $t_c > 0$.

SSBH report that HTP defined immersion learners as either simultaneous bilinguals or "later learners who spent at least 90% of their life in an English-speaking country" (SSBH, p. 7). In fact, later immersion learners were required to have spent at least 90% of their life *since starting to learn English* in an English speaking country (HTP, p. 266). This makes a considerable difference: analyses include immersion learners who began learning English as late as 30, so under SSBH's definition they would need to be at least 300 years old at time of testing. Similarly, SSBH incorrectly report that non-immersion learners were those "who spent at most 10% of their life in an English-speaking country" (SSBH, p. 8), whereas the actual definition is "spent at most 10% of post-exposure life in an English-speaking country and no more than 1 year in total" (HTP, p. 266). Note that SSBH do use the correct definitions in their own analyses, so this does not affect their results.

Probably because of their confusion about how subject groups were defined, SSBH mistakenly report that "more than 100,000 language learners in the HTP database could not be classified as belonging to one of the four groups *because key information was missing*" (emphasis added; p. 20). They assert that this high rate of missing data should cast doubt on the validity/accuracy of the HTP data. However, these subjects were not excluded for missing data but rather for having amounts of immersion intermediate between the "immersion" and "non-immersion" learners (see sentence spanning pp. 266-267).

SSBH misdescribe the stimuli. They report that HTP's test included 132 items, of which 95 were used for analysis "based on the criterion that at least 70% of the native English-speaking adults gave the same response" (SSBH, p. 8). In fact, the criterion was that the same response was given by at least 70% of native English-speaking adults in each of 13 dialect groups (HTP, p. 267). The reason was to exclude items for which there was significant dialectal variation. They also assert that HTP measures accuracy on the grammaticality judgment test on a scale of 0 to 1, reflecting "a proportion of correct answers (g)" (SSBH, p. 7). In fact, *g* represents log-odds accuracy on HTP's syntax test and runs from 1.5 to 3.5 (see HTP Supplementary Materials, p. 2). They misstate how HTP (and, it appears, they) calculated log-odds, asserting that it was based on proportion (*log(p/[1-p])*) (SSBH, p. 7) rather than the empirical logit transformation (*log((num correct+.5)/(num incorrect+.5))*).

In Table 2 and surrounding text, they report some discrepancies between the number of subjects per condition for the critical analyses reported by HTP (p. 266) and in SSBH's own analyses. The problem seems to be that they ran their exclusions in a different order from HTP. Specifically, both papers bin subjects by age, age of acquisition, and condition. We then restrict analyses to consecutive ages for which there were at least 10 participants in a 5-year window. HTP excludes subjects over the age of 70 before this binning, whereas SSBH exclude subjects over the age of 70 *after* binning. This means subjects over the age of 70 count towards binning for SSBH but not for HTP, allowing inclusion of more bins for SSBH. Thus, as they report, they end up with 38 more total included subjects. Since we provided them with the original code, it is not clear why they were unaware of these differences.

When replicating one of HTP's analyses, they report that they obtained "a slightly higher R^2 value of .92 (HTP found .89)" (SSBH, p. 10). This likely reflects the fact that while HTP report cross-validated R^2 values in order to address over-fitting, SSBH do not. This will necessarily result in higher R^2 values. In a personal communication, van der Slik suggested that because they ran the optimization algorithm for more iterations than HTP did, this should obviate the need for cross-validation. This is exactly backwards. It is a necessary fact that the more closely the model is fit to the data, the worse over-fitting gets. In any case, the result is that their R^2 values must be treated with caution: a particular model may achieve a better R^2 simply due to overfitting.

In Footnote 7, they write that Chen and Hartshorne "did not test if the application of their segmented model has resulted in a significant improvement in model fit as compared to the continuous model or even the original HTP discontinuous model." In fact, we provided two such metrics. First, the model fits available to ELSD are a proper subset of those available to Chen & Hartshorne's segmented sigmoid model, and thus fitting the revised model is *per se* a comparison of model fit. Second, Chen and Hartshorne also provide cross-validated R^2 statistics for both their model and the HTP model, allowing direct comparison.

While SSBH present these differences between early and late immersion learners as a novel observation, they were reported first by HTP. In particular, HTP in fact reported two sets of analyses showing that immersion learners who began before the age of 10 learn at least as rapidly and successfully as simultaneous bilinguals (HTP p. 270).

Note that I did not rerun SSBH's model fits themselves and instead copied those numbers from their tables. I cannot guarantee they are correct.

License

Can sign-naïve adults learn about the phonological regularities of an unfamiliar sign language from minimal exposure?

Julia Hofweber
University College London, UK; Northeastern University London, UK

Lizzy Aumônier
University of Kent, UK; Northeastern University London, UK

Vikki Janke
University of Kent, UK

Marianne Gullberg
Lund University, Sweden

Chloë Ruth Marshall
University College London, UK

Abstract: Adults can extract phonological regularities from just several minutes' exposure to naturalistic input of an unknown spoken language (Gullberg et al., 2010). We examined whether such implicit statistical learning mechanisms also operate in the sign language modality. The input materials consisted of a continuous sign stream in the form of a weather forecast in Svenskt Teckenspråk (STS). L1-speakers of English with no prior knowledge of a sign language were assigned to two experimental groups who watched the forecast once (N=43) or twice (N=38), and a control group who did not watch it (N=40). Participants completed a 'surprise' lexical decision task designed to tap into their awareness of the phonological properties of the core STS lexicon. They viewed individual signs and indicated whether or not these could be real STS signs. The signs comprised four sets: STS signs that (1) were presented, and (2) were not presented, in the forecast; and signs that are not STS signs and (3) contain handshapes outside the STS handshape inventory, and (4) contain sets of phonological features that are dispreferred across sign languages. We found *no* evidence of any learning of STS phonological regularities. Considered in conjunction with two companion studies which *did* demonstrate some learning of sign forms and their meanings from these same input materials, our findings suggest limits to what can be learnt after just a few minutes of implicit and naturalistic exposure to language in an unfamiliar modality: information about specific lexical items is learnable, but information that requires generalisation across items may require greater amounts, or a different quality, of input.

Keywords: sign language; implicit learning; first exposure; phonological regularities; iconicity.

Corresponding author: Chloë Marshall, UCL Institute of Education, University College London, 20 Bedford Way, London WC 1H 0AL, UK. Email: chloe.marshall@ucl.ac.uk

ORCID ID: https://orcid.org/0000-0003-2405-1999

Citation: Hofweber, J., Aumônier, L., Janke, V., Gullberg, M., & Marshall, C. R. (2024). Can sign-naïve adults learn about the phonological regularities of an unfamiliar sign language from minimal exposure? *Language Development Research*, *4*(1), 455–480. http://doi.org/10.34842/hofweber

Introduction

The universality – or not – of language acquisition mechanisms has far-reaching implications for the disciplines of psycholinguistics, theoretical linguistics and language pedagogy. Although a substantial body of research has investigated language acquisition in both children and adults, the focus has been almost exclusively on spoken languages (Kidd & Garcia, 2022; Schönström, 2021). Sign languages – which are perceived and produced in the visuo-gestural modality – have been relatively neglected, meaning that little is known about the extent to which their acquisition and developmental trajectory resemble spoken languages. This is problematic because theories of first and second language acquisition based solely on spoken languages (and on the written form of some of those languages) make universal claims, yet it is not known whether those theories hold for sign languages too (Gullberg, 2022; Hou & Morford, 2020; Lillo-Martin & Hochgesang, 2022).

The focus of the current paper is on adult language acquisition mechanisms. An important issue in spoken second language acquisition research concerns how adults are able to learn new languages implicitly, and the type of linguistic knowledge they are able to acquire in this way. We extend this inquiry to sign language acquisition, and specifically to whether sign-naïve adults can learn about the phonological regularities of a target sign language at first exposure. In this introduction we review literature on implicit language learning at first exposure, before summarising the key findings of two studies on implicit sign language learning at first exposure that relate directly to the current one and so help situate and motivate it.

Implicit Language Learning on First Exposure to Naturalistic Input

In the field of adult second language acquisition, distinctions have traditionally been made between explicit and implicit learning. Explicit learning (often instructed learning in classrooms) is associated with intentional, deliberate attempts to memorise something with control, effort, and awareness, for example, in a setting where there has been advance warning of a test of learning. In contrast, implicit or incidental learning is characterised as learning without conscious attention to the input or effort, including in an experimental set-up in which participants are unaware that there will be a test of learning, and where they therefore have no intention to learn and no awareness of what is being learnt (e.g., Andringa & Rebuschat, 2015; DeKeyser, 2003; Godfroid, 2021; Hulstijn, 2005; Rebuschat & Williams, 2012; Williams, 2009, *inter alia*). The field is still rife with debate, especially concerning whether implicit learning can ever actually be tested since the very face of testing draws attention to language and potentially to learning. However, it is now widely recognised recognized that adult learning can take place under both explicit and implicit conditions. What is under discussion are the details, and which type of learning is optimal for which types of knowledge.

The current study focuses on *implicit* learning. We use the term simply to refer to learning taking place without instruction or training, recognising the difficulties of assuring that there is no conscious effort to learn or awareness of what is being learnt. Moreover, we focus on the very earliest stage of learning a new language, when the learner is a novice exposed to the language for the *first time*. Our interest is in how language learning gets off the ground, and specifically, how it does so when the language input is *naturalistic* in form. Second[1] language learning often takes place in instructed contexts, for example, in a classroom where linguistic input is broken down into manageable chunks, such as individual words, and often accompanied by explicit explanations, such as translations into the learner's first language. However, this is not the only way in which learners might encounter a new language. In many contexts, such as migration to another country for personal, political or economic reasons during adult life, learners might instead encounter a new language in more informal contexts. This will involve implicit learning through interactions with colleagues, watching television, listening to song lyrics, playing video games and engaging with social media (see pioneering studies by Meisel et al., 1981; Perdue, 1984; and studies in the emerging field of informal second language learning, e.g., Arndt, 2019; Dressman, 2020; Sockett, 2022). An interesting question is how learning takes place in these scenarios, where input is more continuous in nature, and what exactly can be learnt.

The process of breaking into a new language, what Klein (1986, p. 59) called the learner's "problem of analysis", contains three crucial aspects: (1) segmenting the continuous speech stream to identify relevant strings such as words, (2) identifying meaning that can be mapped onto those sound strings, and (3) generalising beyond the input exemplars to novel items so as to form linguistic categories and extract regularities. A set of three linked studies by Gullberg et al. (2010) investigated this very process by asking adult L1-speakers of Dutch to watch a 7-minute weather forecast presented in a typologically different, and previously unknown-to-them, language, namely Mandarin Chinese. Immediately after viewing the forecast, participants undertook 'surprise' tests tapping into form recognition, meaning assignment and phonotactic generalisations. Although participants found these tasks very challenging, and their overall performance was low, they did nevertheless show evidence of being able to extract word-form-related information, and of managing to extract lexical meaning from the context and map it onto word forms thus identified. They were also able to extract abstract, phonotactic information and generalize it to novel items not encountered in the input, a finding replicated by Ristin-Kaufman and Gullberg (2014) with the same materials but with Swiss-German speakers. These findings suggest that

[1] We use 'second language learning' to cover both second and foreign language learning, because the distinction between them does not matter for our purposes. Nor do we make a distinction in this article between 'acquisition' and 'learning'.

adult learners can deal efficiently and quickly with very complex language input at first exposure, even in the absence of instructions.

A question that arises, however, is whether such results would obtain in languages other than spoken Mandarin. For example, could they be found for a sign language in participants unfamiliar with sign languages? In the next section we review evidence from two recent studies that suggests they can.

Implicit Learning in a New Language Modality: Situating the Current Study

We adapted Gullberg et al.'s (2010) implicit learning paradigm to a sign language (specifically, Swedish Sign Language: Svenskt Teckenspråk, STS) to investigate whether the learning mechanisms identified in learners of spoken language at first exposure to uninstructed, naturalistic and continuous input are evident across the modality boundary. We created an STS weather forecast and produced STS versions of Gullberg et al.'s tasks for identifying word forms and lexical meaning. In two studies linked to the current study, we showed that (1) sign novices can distinguish between signs that they have and have not seen in the STS weather forecast, revealing that they are able to identify sign forms in the sign stream (Hofweber et al., 2022), and that (2) they are able to assign meaning to signs more accurately than control participants who have not viewed the forecast (Hofweber et al., 2023). As in Gullberg et al.'s studies, performance was not high[2], which indicates that the tasks are challenging. Nevertheless, the findings do provide evidence of implicit language learning.

The current study completes the trio of experiments by asking whether adults with no previous sign language exposure can generalise beyond encountered signed exemplars on first exposure. Can they, when presented with signs that they have not previously seen, make accurate judgements about which signs could be possible signs of the sign language they have viewed and those which could not? This question is particularly important with respect to implicit learning because whereas Hofweber et al.'s two previous studies focused on the encoding of lexical items in memory, here we ask whether sign novices can extract regularities across those items. Language acquisition – whether in an individual's first or subsequent languages – crucially involves learners being able to generalise beyond exemplars that they have encountered in the input, in order to form categories and establish regularities (see Ambridge, 2020, for an in-depth and up-to-date treatment of this topic). Hofweber et al. (2022, 2023) demonstrated that learners recognised and assigned meaning to new exemplars of lexical items viewed in the test phase of their studies. A remaining question, however, is whether they can establish regularities.

[2] In Study (1), accuracy rates ranged from 53% to 64% across conditions, with 50% representing chance performance, while in Study (2) participants were only able to generate the correct meaning for 14% of signs.

Recall that Gullberg et al. (2010) showed that L1-Dutch-speaking adults could extract phonotactic regularities, i.e., highly abstract information about sound structures, from input in another newly encountered spoken language (Mandarin Chinese). More specifically, learners rejected consonant-vowel-consonant forms with a phonotactically-illegal final consonant (e.g., *gam*). The ability to identify these words as impossible in Mandarin must have stemmed from participants analysing the new language input rather than from transferring their L1 phonotactic knowledge. This is because consonant-vowel-consonant words of this type are acceptable in Dutch (and, indeed, were accepted as possible Mandarin words by a control group of participants who had not viewed the weather forecast).

At this point, one might reasonably ask how 'phonotactics' - a term coined in spoken language linguistics for the rule-based ways in which phonemes can be combined - relates to sign languages. In fact, just as there are phonotactic constraints in spoken languages, where not all combinations of phonemes are possible, so are there constraints on how the formational units of signs (i.e., handshapes, movements and locations) are combined. Certain formational properties are dispreferred in lexical signs across the world's sign languages (Johnston & Schembri, 2009; Sandler, 2012; Sandler & Lillo-Martin, 2006). Examples of phonotactically-dispreferred signs often involve a change of handshape, i.e., the second handshape involves a different set of selected fingers to the first (thereby violating the 'selected fingers constraint', Mandel, 1981), or two moving hands have different handshapes (thereby violating the 'symmetry condition', Battison, 1978).

Preliminary evidence indicates that people might indeed be sensitive to phonotactic regularities in sign when exposed to sign language for the first time. Hofweber et al. (2022) conducted a recognition test, whereby participants first watched a 4-minute weather forecast in STS and were then presented with individual signs. Their task was to decide whether or not each of those signs had or had not appeared in the forecast. There were two sets of items that participants had not seen in the forecast. One comprised phonologically plausible signs in that they were real signs of STS that shared some phonological features with the target signs. The second set, however, were phonologically implausible in that although they were real signs in other sign languages, they were not STS signs and contained phonological features that are dispreferred across the world's sign languages (because they violated the selected fingers constraint or symmetry condition explained above). Participants were more accurate than chance at responding 'no' to both sets of signs that they had not seen in the forecast, but the fact that this effect was greater for the implausible signs than the plausible signs suggests some sensitivity to sign phonotactics.

Further with respect to sign language phonology and the phonotactic constraints discussed above, the core lexicon[3] of each sign language uses a specific set of handshapes. The handshape inventory of any particular sign language is smaller than the larger set of handshapes attested among sign languages of the world (just as each particular spoken language uses only a portion of the phonemes attested worldwide) (Brentari & Eccarius, 2010). This inventory may be augmented by handshapes that represent letters in the one-handed manual alphabet ('fingerspelling'). The inventory for STS can be found here https://teckenspraksleexikon.su.se/handformer. We investigated whether our participants would be sensitive to the handshapes that had occurred in the weather forecast so, when shown signs that they had *not* seen in the input materials, they would accept signs containing handshapes they had seen as possible signs of the language but reject signs containing handshapes they had not seen. Because in the current study we look at both phonotactics and handshapes, we use the broader term 'phonological regularities' from here on.

For the current task, we used the same STS weather forecast exposure video as in Hofweber et al. (2022, 2023) but administered a different experimental task to determine whether sign novices are sensitive to phonological regularities in a newly encountered sign language. Our experimental task required participants to decide whether the signs being shown were real signs of STS or not. There were four sets of signs. Two of these sets comprised real STS signs, one of which appeared in the forecast and one of which did not. The remaining two sets comprised signs that are not signs of STS (and therefore which also did not appear in the forecast). More specifically, one of these sets comprised signs with handshapes outside the STS handshape inventory, and the other comprised signs with sets of phonological features that are dispreferred across sign languages (because they violated the selected fingers constraint or symmetry condition).

Of course, it might be the case that people who have never experienced a sign language nevertheless have expectations of what signs might look like. After all, they bring with them a lifetime of using their hands to make co-speech gestures and of watching the co-speech gestures of others. Furthermore, there is growing evidence that the phonological system is, at least in part, amodal, and that sign-naïve adults can transfer knowledge of regularities in their spoken L1 to signs (Berent & Gervain, 2023, and references therein). It is possible that certain hand configurations and movements are more plausible than others when people are required to make an explicit

[3] The lexicon of many sign languages, including STS, has a tripartite lexicon, comprising core signs, non-core (depicting and pointing) signs and borrowed (fingerspelt) signs (Johnston & Schembri, 2009). Phonological constraints are most strongly attested in the core lexicon (Brentari & Eccarius, 2010). Apart from the occasional pointing sign (pronouns signed with the index finger, and flat-handed points to the weather map), our weather forecast contains just core lexical signs.

judgement on what could or could not be a sign. To control for this possibility, a group of participants undertook the lexical decision task without having watched the weather forecast.

We expected the task to be challenging (as it was for the participants in the studies by Gullberg et al., 2010; Hofweber et al., 2022; Hofweber et al., 2023; Ristin-Kaufmann & Gullberg, 2014) but we predicted that participants who had viewed the forecast would be more accurate at distinguishing between STS and non-STS signs than those who had not viewed it. Importantly, if participants were able to accept both real STS signs that they had just viewed and signs that they had not, and also reject signs that are not signs of STS (which, by definition, would not have appeared in the forecast), this would indicate some level of generalization across the input and not just the recognition of viewed exemplars.

Methods

The materials used in this study can be accessed on the Open Science Framework site via the following link: https://osf.io/8hrp6/. It should be noted that video materials need to be downloaded for viewing.

Participants

This study was originally designed to be run face-to-face in the lab. However, due to testing restrictions in place during the Covid-19 pandemic, we adapted it for online administration. Hence, we uploaded the experiment originally designed in PsychoPy onto the Pavlovia platform, https://pavlovia.org/. This allowed us to reach a more diverse participant pool than is common in lab-based psychology studies. However, the online format also meant that compatibility issues between Pavlovia and participants' domestic software set-up meant that data from some participants (7 in the 0x, 10 in the 1x, and 3 in the 2x Exposure groups) could not be collected. Hence, this paper only reports data collected from participants in which no compatibility issues occurred and the full experimental task could be administered (final *N*=121). In addition, some of the Wechsler Adult Intelligence Scale (WAIS) vocabulary task data had to be discarded (data from 3 participants each in the 0x and 2x Exposure groups, and 15 participants in the 1x Exposure group) due to problems with the audio quality of the voice recording that made transcription unreliable.

Participants were recruited using the website "Call for participants" (https://www.callforparticipants.com). None of them had participated in either of the companion studies to this study (Hofweber et al., 2022; Hofweber et al., 2023). They were adult native-speakers of English with no prior knowledge of any sign languages or of Swedish. Given that the tasks were visual, and to avoid confounds from age-related decline of vision, the maximum age for participants was set at 40 years.

Participants were randomly assigned to one of the following three groups, with group sizes designed to be twice as large as those used by Gullberg et al. (2010) and commensurate with those used by Hofweber et al. (2022) and Hofweber et al. (2023):

- 1x exposure group (*N*=43; 17 males): exposed to the STS forecast once
- 2x exposure group (*N*=38; 19 males): exposed to the STS forecast twice
- 0x exposure group (*N*=40; 21 males): control group who were not exposed to the STS forecast.

Using a detailed demographic and language background questionnaire, each participant was assessed for general demographic variables, such as age and education (an indicator of socio-economic status), as well as for existing knowledge of languages. We also assessed non-verbal reasoning using the matrices subtest of the WAIS III, and English (i.e., the participants' L1) vocabulary knowledge using the vocabulary subtest of the WAIS IV. Between-subject ANOVAs were conducted to compare the participant groups on these background variables. As can be seen in Table 1, the three groups were matched for age, education and number of known languages. However, the 1x exposure group displayed significantly higher non-verbal reasoning and English vocabulary scores than the other groups. Nevertheless, these factors did not correlate with the dependent variable of the study (accuracy in the experimental task), so the difference was not interpreted further.

Table 1. ***Participant background variables***

		Exposure group			*F*	*p*	η^2
		0x	1x	2x			
Age (years)	Mean	23.87	25.88	24.30	2.06	.13	.05
	SD	2.56	5.10	3.51			
Education (years)	Mean	16.53	16.58	16.97	0.23	.79	.005
	SD	2.79	3.37	2.21			
English vocabulary knowledge	Mean	33.17	38.42	34.94	3.93	.02*	.09
(WAIS IV vocabulary raw score)	SD	6.85	7.31	6.63			
Number of known languages	Mean	1.87	2.04	2.24	0.58	.57	.01
	SD	1.38	1.16	1.54			
Non-verbal reasoning ability	Mean	18.67	22.21	19.73	5.83	.004**	.12
(WAIS III matrices raw score)	SD	3.48	4.38	3.75			

Key: * = p < .05; ** = p < .01

Experimental Materials

Weather Forecast

The exposure materials consisted of a 4-minute weather forecast in STS, recorded by a hearing native signer of STS who is a qualified and highly experienced interpreter. These materials were created specifically for this study and for those reported in Hofweber et al. (2022, 2023). Because our participants had English as their L1 and because we were particularly interested in the manual aspects of sign language, we did not use British Sign Language (BSL): the mouthing of many BSL signs is based on the lip patterns for the corresponding English words, so our participants might have relied on lip-reading to make sense of the input materials. For this reason, we chose another language – STS – as the target language.

Amongst a range of other signs, the forecast incorporated 22 target signs which featured in the lexical decision task (designed to be similar in number to the 24 target words used by Gullberg et al., 2010, in their Mandarin weather forecast). These 22 target signs occurred with different frequencies in the forecast. We presented 11 'high frequency' target signs (8 occurrences) versus 11 'low frequency' target signs (3 occurrences; with the exception of one item, SÖDER 'south', which occurred 4 times in error). The high and low frequency sets were matched for a range of crucial aspects of sign language, such as phonology (i.e., locations and hand configurations, and the number of one-handed versus two-handed signs) and iconicity. Iconicity ratings collected from an independent group of 24 sign-naïve participants confirmed the experimenters' intuition that high (M = 3.64, SD = 1.55) and low frequency (M = 3.68, SD = 1.76) signs did not differ in their level of perceived iconicity, $F(1,22) = 0.003$, $p = .96$, $\eta^2 = 0.000$ (see Hofweber et al., 2023, for further details). The two types of target items were also matched for the occurrence frequency of their English translation equivalents using CELEX corpus (Baayen et al., 1995): low: M = 32,759, SD = 51,978; high: M = 27,027, SD = 22,771, $F(1,22) = 0.11$, $p = .74$, $\eta^2 = 0.006$.

Lexical Decision Task

Participants saw 88 short video clips of individual signs, each signed by the same signer who signed the forecast. After each clip, they made a meta-linguistic judgement as to whether or not the sign was a real sign of sign languages. All participants were presented with the same video stimuli but the instructions differed slightly, depending on whether participants had seen the forecast or not. Participants in the 1x and 2x exposure groups were asked if the signs could be real signs of STS specifically, whilst participants in the control group were asked whether the signs could be real signs of sign languages more generally. Control participants received different instructions because it would have been pragmatically odd to ask them to make a judgement relating to a specific sign language they had never seen before, but they could be assumed to have some expectation of what sign languages in general look like.

The stimuli consisted of four sets of signs. None of these signs contained mouthings (i.e., silent mouth patterns from spoken words that signers sometimes use to accompany manual signs), so participants were not able to gain any (spoken) language information from viewing the signer's mouth. Set 1 included 22 stimuli that were the target signs from the forecast, so the accurate answer to these was 'yes', given that they are real signs of STS. Set 2 had 22 stimuli which were also real signs of STS. They contained handshapes, hand orientations, movements and locations that had occurred in the forecast but that were combined in different ways from those shown in the forecast. Therefore, for this set, too, the correct response was 'yes'. The remaining two sets of signs required the response 'no'. Set 3 comprised 22 signs with handshapes (different for each sign) that are not part of the core STS lexicon, and had therefore not appeared in the forecast. However, these handshapes do occur in other sign languages and the signs were indeed real signs (e.g., American Sign Language, Chinese Sign Language, Kenyan Sign Language and Khmer Sign Language). The fourth and final set comprised 22 signs with the sorts of handshape changes and movements that violate constraints of sign formation and are therefore dispreferred across the world's sign languages. None of the signs shown in the forecast violated these constraints. Again, however, signs in this set were real signs from other sign languages. Table 2 summarises the properties of each set of signs. Note that all other phonological differences between the four sets were minimal: all four sets were matched for the number of one- and two-handed signs, and Set 2 was matched to Set 1 for the hand configurations used.

Supplementary Tasks

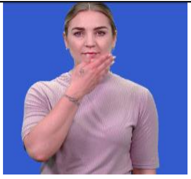
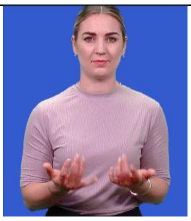
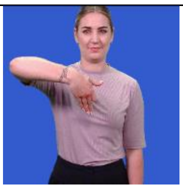
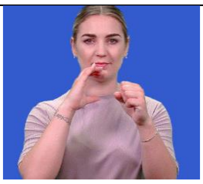
English Vocabulary Knowledge

Participants' knowledge of their first language was measured using the English vocabulary subtest of the Wechsler Adult Intelligence Scale WAIS-IV (Wechsler et al., 2008; not available on the osf site because it is proprietary). In this test, participants are presented with 26 lexical items auditorily and orthographically and asked to define each one. Their responses were recorded using the audio software *Audacity* and scored based on the test manual. The responses from a subset of participants (N=10) were scored by two independent judges, resulting in an interrater reliability score of Spearman's *Rho* = .85, p = .002.

Non-verbal Reasoning Ability

The matrices subtest of the Wechsler Adult Intelligence Scale WAIS-III (Wechsler, 1997; not available on the osf site because it is proprietary) taps into individuals' visual ability to recognise patterns. Participants view designs of shapes and colours. Each design contains a gap. In a multiple-choice style, participants choose one from several options to complete the design. The task was administered online using google docs.

Table 2. ***Properties of the stimulus items in each sign set***

Set	Signs are included in the weather forecast	Signs are in the core STS lexicon	Signs use handshapes from the core STS lexicon	Signs violate the phonotactic constraints of sign languages	Example (photos illustrate each sign's hand configuration(s) and location, and the movement is described in the text. The item ID can be used to locate the video on the osf site)
1	yes	yes	yes	no	 Sweeping movement across chin Gloss: This sign means 'warm' in STS THF1
2	no	yes	yes	no	 Repeated finger wiggle Gloss: This sign means 'simmer' in STS LDD1THF3
3	no	no	no	no	 Double short downward movement Gloss: This sign means 'Namibia' in Namibian Sign Language. LDD2THF1
4	no	no	yes	yes	 Repeated asynchronous movement of the two hands that have different handshapes Gloss: This sign means 'SimCom' in American Sign Language LDD3THF2

Procedure

Since the data collection for this study took place in 2021 when in-lab testing was not permitted due to the Covid-19 pandemic, the study was conducted online using the Microsoft Teams software. The experimenter met and observed each participant individually on Teams. Participants in the exposure groups first watched a short video of a weather forecast in STS. Because the study was designed to tap implicit learning, the instructions were minimal to avoid explicit reference to learning. Participants were simply told to watch the signer as she signed the forecast. Immediately after watching the forecast, they completed the surprise lexical decision task. The control group proceeded with the lexical decision task straight away, without having watched the forecast first. Upon completion of the main experimental tasks, participants completed the WAIS non-verbal reasoning and vocabulary tests. Finally, they filled in the demographic and language background questionnaire on Surveymonkey.

Data Analysis

To analyse the lexical decision task results, we investigated both accuracy rates and yes responses for each item and participant. The summary tables of our results are presented using the style adopted by Ortega et al. (2019). The full data set is available on the osf site: https://osf.io/8hrp6/. Generalised mixed model analyses were conducted in R studio using the lmer.test package in R (Kuznetsova et al., 2017), which automatically generates significance levels for each effect. We initially assumed a maximally conservative approach to random effects, allowing both items and subjects to vary by both intercept and slope (Winter, 2019). However, this resulted in failure to converge due to the model complexity, so we simplified the models to vary only by intercept. The alpha level was set at .05. Due to limitations regarding modelling pair-wise comparisons involving variables with three or more levels in lme4 (Winter, 2019), we conducted the analyses in several steps.

The first step was to assess the overall effect of exposure by comparing results in the control group (0x) to results in the two exposure groups (1x, 2x). Secondly, we compared the two exposure groups to each other. Finally, we assessed the effect of Set for exposure and non-exposure groups separately. Since, there were no differences in pattern across the 1x and 2x exposure groups, the two groups were combined for the analyses by Set. In these analyses by Set, the intercept was set as the values of Set 1 (i.e., target) items.

The analyses on effects of Set were conducted separately for Accuracy and Yes responses, to reveal response biases. Any differences in accuracy between the two sets of items that required a ‘yes’ response (Sets 1 and 2) and the two sets that required a ‘no’ response (Sets 3 and 4) could potentially be driven by participants not actually making a distinction between any of the sets at all, and responding ‘yes’ at similarly high levels throughout, regardless of the phonological properties of the signs. In order to investigate this possibility, we repeated the analysis using Yes responses rather than Accuracy as the dependent variable. We note that the variable 'Accuracy' makes less sense for the control group who did not see the forecast, because their task instructions asked not about STS specifically but about sign languages more generally; because the signs used in Sets 3 and 4 were real signs (albeit ones that are formationally rare) the correct response for these signs was - like it was for Sets 1 and 2 - 'yes'. However, we include the data from the control group in the Accuracy analyses for the sake of completeness.

Results

Results for Accuracy Rates

Table 3 presents the descriptive statistics for Accuracy rates by exposure group and sign set. Figures 1 and 2 illustrate the distribution of accuracy rates. Tables 4, 5, 6 and 7 present the inferential statistics based on linear mixed effects models. Table 4 presents the effect of overall exposure on accuracy. Table 5 presents the effect of number exposures (1x versus 2x) on accuracy. Table 6 presents the effects of Set on accuracy in the 0x control group. Table 7 presents the effects of Set on accuracy in the two exposure groups. Based on Tables 4 and 5, accuracy did not differ by exposure group. As a result, subsequent analyses did not differentiate between 1x and 2x exposure groups. However, the different sign sets yielded different accuracy rates. Whilst there were no accuracy differences between Sets 2 and Sets 1 (target signs), accuracy in Sets 3 and 4s was lower than in Set 1 (target signs). This effect applied across all exposure groups.

Table 3. ***Accuracy rates by exposure group and sign set***

		0x Exposure group	1x Exposure group	2x Exposure group
Accuracy rate (%)	Mean	59.52	55.60	61.41
Set 1 signs	SD	23.07	16.73	16.78
Accuracy rate (%)	Mean	59.57	61.91	59.50
Set 2 signs	SD	24.75	13.86	17.11
Accuracy rate (%)	Mean	42.74	44.64	46.87
Set 3 signs	SD	22.40	16.20	19.81
Accuracy rate (%)	Mean	40.63	44.55	47.12
Set 4 signs	SD	19.63	16.15	19.87

As explained in the text, accurate responses for Sets 1 and 2 were 'yes' and for Sets 3 and 4 were 'no'.

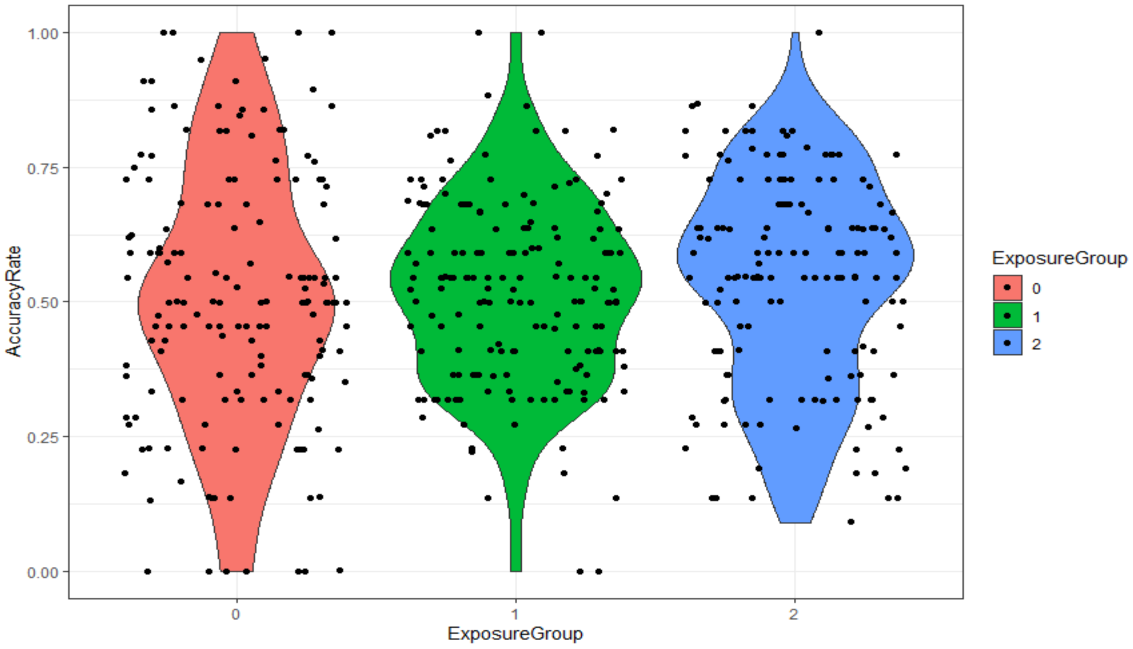


Figure 1. ***Accuracy rates by exposure group***

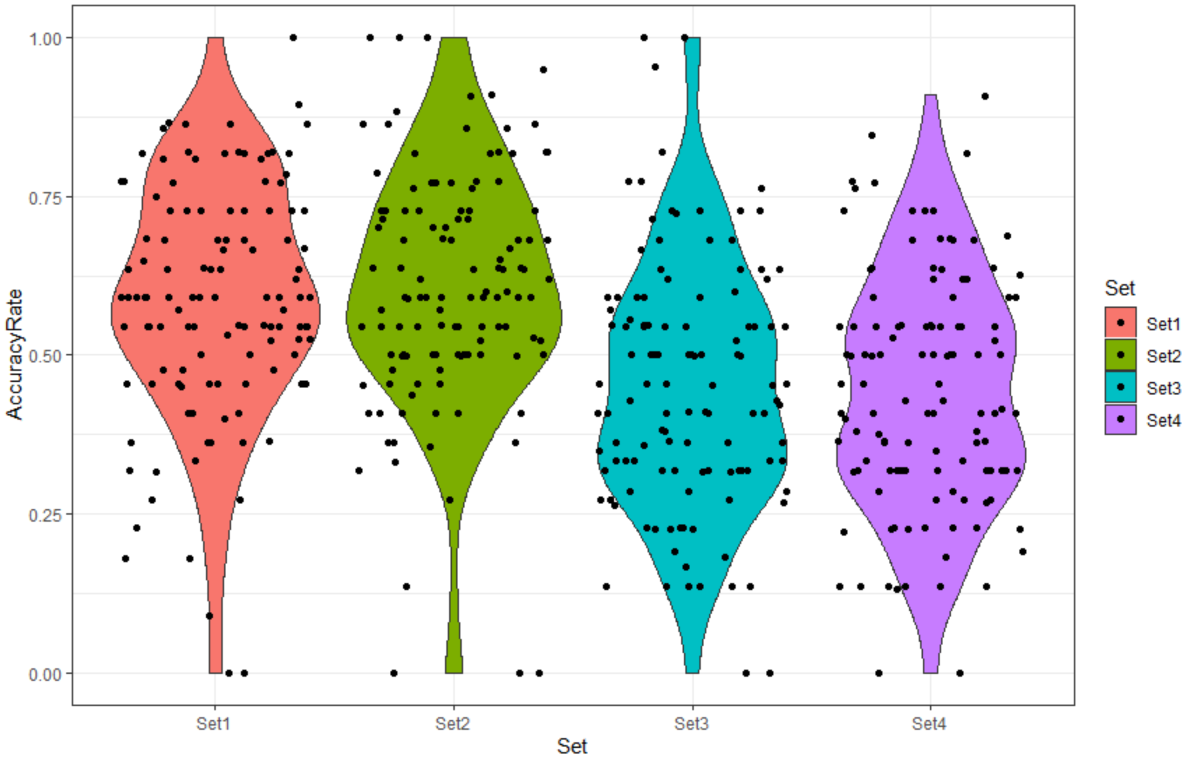


Figure 2. ***Accuracy rates by set***

Table 4. ***Model output for Accuracy: Effect of Exposure (0x versus 1x/2x)***

Predictors	β	SE	Z	*p*
Intercept (0x group)	0.04	0.07	0.67	.50
Exposure (1x/2x groups)	0.01	0.03	0.33	.74

glmer (accuracy ~ exposure+(1|item)+(1|subject),data=Data, family=binomial)

Table 5. ***Model output for Accuracy: Effect of Exposure times (1x versus 2x)***

Predictors	β	SE	Z	*p*
Intercept (1x group)	0.01	0.08	0.17	.86
2x exposure	0.09	0.08	1.13	.26

glmer (accuracy ~ exposure_times+(1|item)+(1|subject),data=Data_exp, family=binomial)

Table 6. ***Model output for Accuracy: Effect of Set in 0x exposure group***

Predictors	β	SE	Z	*p*
Intercept (set 1)	0.44	0.08	5.52	<.01
Set 2	0.05	0.11	0.43	.67
Set 3	-0.80	0.11	-7.61	<.01
Set 4	-0.86	0.11	-8.12	<.01

glmer (accuracy ~set+(1|item)+(1|subject),data=Data_0, family=binomial)

Table 7. ***Model output for Accuracy: Effect of Set in 1x/2x exposure group***

Predictors	β	SE	Z	*p*
Intercept (Set 1)	0.32	0.10	3.04	<.01
Set 2	0.15	0.14	1.08	.28
Set 3	-0.51	0.14	-3.66	<.01
Set 4	-0.67	0.14	-4.79	<.01

glmer (accuracy ~set+(1|item)+(1|subject),data=Data_exp, family=binomial)

Additional analyses were conducted to explore the predictors of Accuracy in Set 1 items. No significant effects of input-related factors, such as frequency or iconicity (p-values > 0.05), or correlations with individual differences (age, education, non-verbal reasoning ability, English vocabulary, number of languages known) were observed (all r-values < .10).

Results for Yes responses

Table 8 presents the descriptive statistics for Yes response rates by exposure group and set. Figures 3 and 4 illustrate the distribution of Yes responses. Whilst Tables 9 and 10 present the inferential statistics based on linear mixed effects models. the effect of group will be identical for accuracy and Yes response rates, the effect of Set may not be. Table 9 presents the effects of Set on Yes response rates in the 0x control group. Table 10 presents the effects of Set on Yes response rates in the two exposure groups. In neither analysis was there a significant effect of Set, indicating that participants – whether or not they had watched the forecast – did not respond differently to signs that were or were not signs of STS. There is therefore no evidence for the learning of phonological regularities during the viewing of the input materials.

Table 8. ***Yes response rate by exposure group and sign set***

		0x Exposure	1x Exposure	2x Exposure
Yes rate (%)	Mean	59.52	55.60	61.41
Set 1 signs	SD	23.07	16.73	16.78
Yes rate (%)	Mean	59.57	61.91	59.50
Set 2 signs	SD	24.75	13.86	17.11
Yes rate (%)	Mean	57.26	55.36	53.13
Set 3 signs	SD	22.40	16.20	19.81
Yes rate (%)	Mean	58.37	59.84	56.88
Set 4 signs	SD	23.54	16.93	18.46

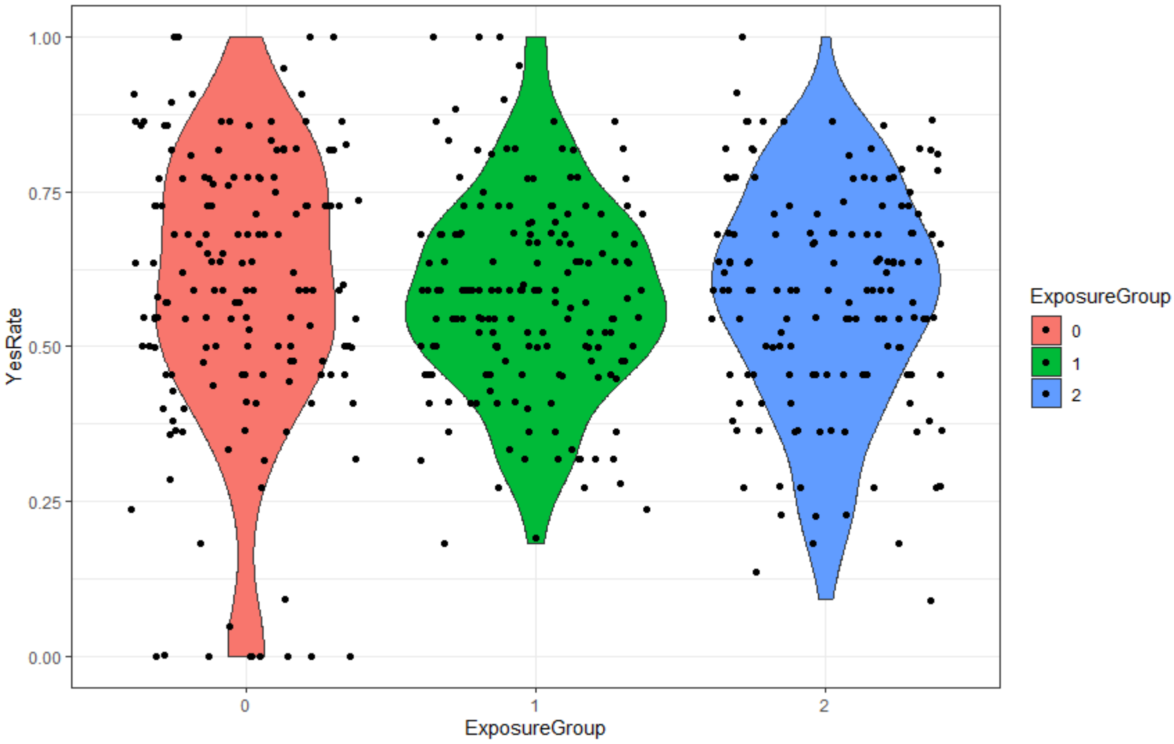


Figure 3. ***Yes rates by exposure group***

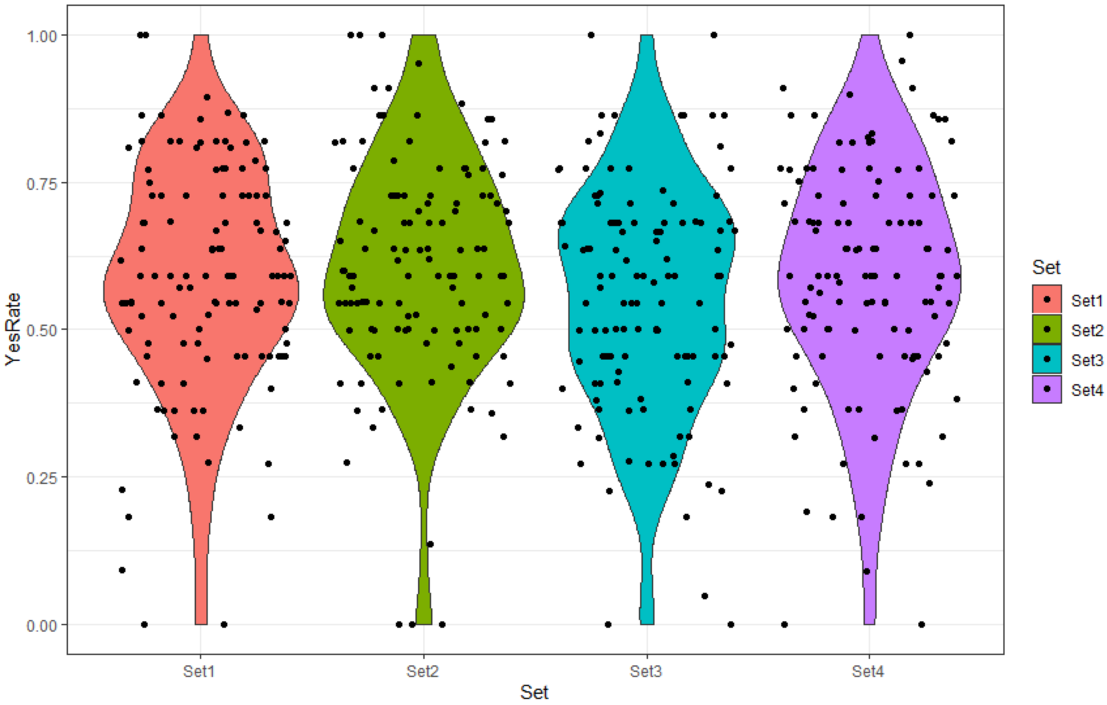


Figure 4. ***Yes rates by set***

Table 9. ***Model output for yes response rate: Effect of set in 0x exposure group***

Predictors	β	SE	Z	*p*
Intercept (Set 1)	0.40	0.20	2.02	.04
Set 2	0.03	0.12	0.23	.82
Set 3	0.09	0.12	-0.73	.47
Set 4	-0.03	0.12	-0.25	.81

glmer (response1 ~set+(1|item)+(1|subject),data=Data_0, REML=false)

Table 10. ***Model output for yes response rate: Effect of set in 1x/2x exposure group***

Predictors	β	SE	Z	*p*
Intercept (Set 1)	0.34	0.12	2.85	<.01
Set 2	0.16	0.15	1.07	.29
Set 3	-0.13	0.15	-0.86	.38
Set 4	0.04	0.15	0.30	.76

glmer (response1 ~set+(1|item)+(1|subject),data=Data_exp, REML=false)

Discussion

In this study, we investigated whether sign-naïve adults are able to learn about the phonological regularities of a target sign language at first exposure in an implicit, uninstructed learning context. The input materials consisted of a continuous sign stream in the form of a 4-minute weather forecast video in Swedish Sign Language (STS). After having watched the forecast either once or twice, participants completed a 'surprise' lexical decision task designed to tap into their understanding of the phonological properties of STS and sign languages in general. The participants viewed individual signs and were asked to indicate whether each sign could be a 'real sign of STS'. Stimulus items comprised four sets: (1) STS signs that were presented in the forecast, (2) STS signs that were not presented in the forecast, (3) signs that are not STS signs and contain handshapes outside the STS handshape inventory, (4) signs that are not STS signs and contained phonological features that are dispreferred across the world's sign languages. Correct answers to Sets 1 and 2 were 'yes' and to Sets 3 and 4 were 'no'. We also tested a group on the experimental task who had not viewed the forecast.

Our predictions were not borne out by the data. First of all, we had predicted that being exposed to an unfamiliar language in an unfamiliar modality – albeit under implicit learning conditions and for only a short time – would lead to greater accuracy in distinguishing between signs that were STS signs versus those that were not, in comparison to participants who had had no exposure. In other words, we had predicted that participants would be able to learn something about the phonological regularities of sign in a short period. However, this was shown not to be the case. In all three groups (0x exposure, 1x exposure, 2x exposure), the rate of accepting signs as possible signs of STS was around 58%. In fact, this rate of acceptance was fairly consistent across different sign sets too (Set 1, 59%; Set 2, 60%; Set 3, 55%; Set 4, 58%). In other words, participants did not distinguish between the different sets of signs in their responses, and so did not behave as we had expected – they did not reject the two sets of signs that we had predicted would be phonologically implausible to them. We had hypothesized that if participants who had viewed the weather forecast were able to accept both real STS signs that they had just viewed in the forecast and signs that they had not, and also reject signs that are not signs of STS (which, by definition, would not have appeared in the forecast), then this would indicate some level of generalization across the input and not just the recognition of viewed exemplars. We did not find evidence to support this hypothesis.

Taking the results of this study together with our two earlier studies (Hofweber et al., 2022; Hofweber et al., 2023), we appear to have found some limits on what can be learnt of a sign language at first exposure to brief and naturalistic input: lexical information – namely sign forms, and the meaning of signs – can be learnt, but it appears that phonological regularities cannot. This is in interesting contrast to Gullberg et al.'s spoken language studies of Mandarin Chinese learning by Dutch and Swiss-German speakers (Gullberg et al., 2010; Gullberg et al., 2012; Ristin-Kaufmann & Gullberg, 2014), where phonotactic restrictions on syllable-final consonants *were* learnt. And yet, it is not the case that our participants were responding 'yes' and 'no' at chance: we found a bias towards responding 'yes'. Participants were therefore erring on the side of being more rather than less accepting of the types of signs that could be part of the STS lexicon. It appears that just four minutes of exposure to naturalistic input is below the threshold for any learning of phonological regularities to occur.

And yet, we did have some hints from a previous study that learners might be able to extract phonological regularities after brief exposure to the same signed input materials. In Hofweber et al.'s (2022) recognition study – where participants had to make a decision as to whether signs had or had not appeared in the weather forecast – there were two sets of items that they had not seen in the forecast. One set comprised signs that were phonologically plausible in that they were real signs of STS and they shared some phonological features with the target signs. The second set of signs, however, were phonologically implausible because they contained phonological features that are dispreferred across the world's sign languages (like the signs in Set 4 in the current study, they broke the selected fingers constraint or symmetry condition). Hofweber et al.'s (2022) participants were more accurate than chance at responding 'no' to both sets of signs that they had not seen in the forecast, but this effect was greater for the implausible signs than the plausible signs, suggesting some sensitivity to phonological regularities.

The difference in findings could perhaps be explained by differences in task and instructions. The task in the current study required higher levels of meta-linguistic awareness. Whilst in Hofweber et al.'s (2022) study, participants were simply asked whether they recognized the sign from the forecast, in the current study they were asked to make a complex and abstract judgement, i.e., 'could this be a real sign of Swedish Sign Language or not'. It is possible that this type of judgement is too challenging for someone with no expertise in language studies. Moreover, it may have been difficult for participants without any prior experience with sign languages to make a particular judgement relating to one sign language. This would have been less of a consideration for speakers being asked about a particular spoken language such as Mandarin Chinese, as was the case in Gullberg and colleagues' studies. Their speakers already had familiarity with spoken languages and will have developed an awareness of the fact that languages can 'sound different' from each other. Can we assume the same for someone who has never learnt a sign language? After all, they were being exposed not just to an unfamiliar language but to an unfamiliar modality. Maybe they interpreted 'could this be a real sign of Swedish Sign Language or not' as a question about modality rather than a particular language, e.g., 'could this be a real sign of a sign language or not' (i.e. the question that the control group, who did not view the forecast, were asked), in which case it might seem strategic to be relatively generous with what can be accepted. That might also explain why the exposure groups did not differ from the control group in their responses.

An alternative explanation of our findings is that the phonological constraints posited for the core lexicon of sign languages are not as strong as is assumed in the literature. It is certainly true that these constraints are broken outside the core lexicon in classifier constructions and in signs that incorporate elements of fingerspelling (Johnston & Schembri, 2009; Sandler & Lillo-Martin, 2006). Furthermore, there is a greater range of handshapes outside the core lexicon (Brentari & Eccarius, 2010). Therefore, it might be that learners need longer than just a few minutes of exposure to learn how these constraints apply to lexical signs. A limitation of our study is that we did not check how strong these phonological constraints are for native signers of STS: we did not investigate how sign-like (or un-sign-like) they would judge the signs in Sets 3 and 4 to be, and this would be a useful addition to any future studies using our experimental paradigm.

A further consideration is whether participants paid sufficient attention to the weather forecast and the experimental task. If not, that could have contributed to our lack of differences between the different sign sets. However, two points suggest that poor attention is not necessarily an issue: (1) in the online presentation of our companion study tapping lexical meaning (Hofweber et al., 2023), a learning effect *was* obtained, and (2) the 'yes bias' in the current study suggests that participants were not responding completely at chance and were trying to respond accurately. With hindsight, the inclusion of catch trials would have been useful in order to determine more directly whether our participants were paying attention, and to exclude any who were not.

An interesting next step would be to explore whether, in contrast to non-signers, people with experience of a sign language unrelated to STS are able to learn its phonological regularities from our same input materials. As Chen Pichler and Koulidobrova (2023) discuss, the current literature on sign language learning has focused on hearing adults learning their first language, but deaf signing adults who are learning a new sign language are a key group for fully understanding the impact of modality on second language learning. In the case of the present task and those reported in our related papers (Hofweber et al., 2022; Hofweber et al., 2023), such a group would allow us to disentangle second-language learning effects within the signed modality from learning effects in a new modality.

The findings from our study also raise questions about possible differences between child and adult acquisition mechanisms, and between first and second language acquisition of phonotactic regularities in a visual modality. In the domain of spoken language, a large body of work has shown that both children and adults are able to successfully extract phonotactic regularities from spoken/auditory input (see Frost et al., 2019, for an overview), although most studies have operated with training paradigms rather than exposure to continuous input without training (but see Ristin-Kaufmann & Gullberg, 2014). Much less is known about differences between children and adults for the implicit learning of sign language phonotactics, let alone in a statistical learning paradigm. We know of no studies that test children's capacity for extracting phonotactic regularities in the visual modality, which means that we cannot tell whether the adults in this study are worse at this task than children. It therefore remains an empirical question with interesting theoretical ramifications to compare first and second language learners in this domain.

In conclusion, we found no evidence that hearing adults, after brief exposure to an unfamiliar language in an unfamiliar modality, were able to demonstrate learning of the phonological regularities explored in our study. Considered in conjunction with two companion studies revealing that participants *were* able to demonstrate learning of sign forms and their meanings after viewing these same input materials, we argue that our findings demonstrate the limits of what can be learnt: information about specific lexical items is learnable, but information that requires generalisation across items may require greater quantities of input or a different quality of input. All three of our studies need replication, preferably with different input materials, to establish their robustness. Furthermore, different conditions that might support the learning of phonological regularities need to be explored, for example, longer exposure time and explicit pointers or explanations, and learning might be better demonstrated using different tasks. Finally, we acknowledge that evidence for the implicit learning of phonological regularities at first exposure to an unfamiliar spoken language rests on just two studies of Chinese, and that these findings need to be replicated in other spoken languages (and preferably in languages with phonotactic properties that are very different to those of Chinese). Taken together, such studies in signed and spoken languages will help clarify the extent to which adult language acquisition mechanisms operate similarly or differently across modalities.

References

Ambridge, B. (2020). Against stored abstractions: a radical exemplar model of language acquisition. *First Language, 40,* 509-559. https://doi.org/10.1177/0142723719869731

Andringa, S., & Rebuschat, P. (2015). New directions in the study of implicit and explicit learning. *Studies in Second Language Acquisition, 37,* 185-96. https://doi.org/10.1017/S027226311500008X

Arndt, H.L. (2019). *Informal second language learning: The role of engagement, proficiency, attitudes, and motivation* [PhD thesis]. University of Oxford.

Battison, R. (1978). *Lexical borrowing in American Sign Language*. Linstok Press.

Berent, I., & Gervain, J. (2023). Speakers aren't blank slates (with respect to phonology)! *Cognition, 232,* 105347. https://doi.org/10.1016/j.cognition.2022.105347

Brentari, D., & Eccarius, P. (2010). Handshape contrasts in sign language phonology. In Brentari, D. (Ed.) *Sign languages* (pp. 284-311). Cambridge University Press.

Chen Pichler, D., & Koulidobrova, E. (2023). The role of modality in L2 learning: The importance of learners acquiring a second sign language (M2L2 and M1L2 learners). *Language Learning, 73,* Issue S1, 197-233. https://doi.org/10.1111/lang.12607

DeKeyser, R.M. (2003). Implicit and explicit learning. In C.J. Doughty & M.H. Long (Eds.), *The handbook of second language acquisition* (pp. 313-348). Blackwell.

Dressman, M. (2020). Introduction. In M. Dressman & R.W. Sadler (Eds.) *The handbook of informal language learning* (pp.1-12). Wiley.

Ellis, N.C. (2012). What can we count in language, and what counts in language acquisition, cognition, and use? In S. T. Gries & D. Divjak (Eds.), *Frequency effects in language learning and processing* (Vol. 1, pp. 7-34). Mouton de Gruyter.

Frost, R., Armstrong, B.C., & Christiansen, M.H. (2019). Statistical learning research: A critical review and possible new directions. *Psychological Bulletin, 145*(12), 1128-1153. https://doi.org/10.1037/bul0000210

Godfroid, A. (2021). Implicit and explicit learning and knowledge. In H. Mohebbi & C. Coombe (Eds.), *Research questions in language education and applied linguistics* (pp. 823-829). Springer.

Gullberg, M. (2022). Why the SLA of sign language matters to general SLA research. *Language, Interaction and Acquisition, 13*(2), 231-253. https://doi.org/10.1075/lia.22022.gul

Gullberg, M., Roberts, L., Dimroth, C., Veroude, K., & Indefrey, P. (2010). Adult language learning after minimal exposure to an unknown natural language. *Language Learning, 60,* 5-24. https://doi.org/10.1111/j.1467-9922.2010.00598.x

Hofweber, J., Aumônier, L., Janke, V., Gullberg, M., & Marshall, C.R. (2022). Breaking into language in a new modality: The role of input and individual differences in recognizing signs. *Frontiers in Psychology.* 13:895880. https://doi.org/10.3389/fpsyg.2022.895880

Hofweber, J., Aumônier, L., Janke, V., Gullberg, M., & Marshall, C.R. (2023). Which aspects of visual motivation aid the implicit learning of signs at first exposure? *Language Learning, 73,* Issue S1, 33-63. https://doi.org/10.1111/lang.12587

Hou, L., & Morford, J. (2020). Using signed language collocations to investigate acquisition: A commentary on Ambridge (2020). *First Language, 40,* 585-591. https://doi.org/10.1177/0142723720908075

Hulstijn, J. H. (2005). Theoretical and empirical issues in the study of implicit and explicit second-language learning. *Studies in Second Language Acquisition, 27*(2), 129-140. https://doi.org/10.1017/S0272263105050084

Johnston, T., & Schembri, A. (2009). *Auslan: an introduction to sign language linguistics.* Cambridge University Press.

Klein, W. (1986). *Second language acquisition.* Cambridge University Press.

Lillo-Martin, D., & Hochgesang, J. (2022). Signed languages – ordinary and unique: A commentary on Kidd and Garcia (2022). *First Language, 42,* 789-803. https://doi.org/10.1177/01427237221098858

Mandel, M. (1981). *Phonotactics and morphophonology in American Sign Language*. PhD dissertation. University of California, Berkeley.

Meisel, J.M., Clahsen, H., & Pienemann, M. (1981). On determining developmental stages in natural second language acquisition. *Studies in Second Language Acquisition, 3*(2), 104-135. https://doi.org/10.1017/S0272263100004137

Ortega, G., Schiefner, A., & Özyürek, A. (2019). Hearing non-signers use their gestures to predict iconic form-meaning mappings at first exposure to signs. *Cognition, 191,* 103996. https://doi.org/10.1016/j.cognition.2019.06.008

Perdue, C. (Ed.) (1984). *Second language acquisition by adult immigrants. A field manual.* Newbury House.

Rebuschat, P., & Williams, J.N. (2012). Implicit learning in second language acquisition. In C. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics*. https://doi.org/10.1002/9781405198431.wbeal0529

Ristin-Kaufmann, N., & Gullberg, M. (2014). The effects of first exposure to an unknown language at different ages. *Bulletin Suisse de Linguistique Appliquée*, *99,* 17-29.

Sandler, W. (2012). The phonological organization of sign languages. *Language and Linguistics Compass*, *6*(3), 162-182. https://doi.org/10.1002/lnc3.326.

Sandler, W., & Lillo-Martin, D. (2006). *Sign languages and linguistic universals*. Cambridge.

Schönström, K. (2021). Sign languages and second language acquisition research: An introduction. *Journal of the European Second Language Association,* 5(1), 30-43. https://doi.org/10.22599/jesla.73

Sockett, G. (2022). Learning beyond the classroom and autonomy. In H. Reinder, C. Lai, & P. Sundquist (Eds.), *The Routledge handbook of language learning and teaching beyond the classroom* (pp. 67-80). Routledge.

Wechsler, D. (1997). *WAIS-III Administration and scoring manual.* The Psychological Association.

Wechsler, D., Coalson, D.L., & Raiford, S.E. (2008). *WAIS-IV technical and interpretive manual.* Pearson.

Williams, J. (2009). Implicit learning in second language acquisition. In W. Ritchie & T. Bhatia (Eds.) *The new handbook of second language acquisition*. (pp. 319-353). Brill.

Winter, B. (2019). *Statistics for linguists: An introduction using R*. Routledge.

Data, Code and Materials Availability Statement

Data: available on the open science framework, https://osf.io/8hrp6/
Code: available on the open science framework, https://osf.io/8hrp6/, in the document 'R_scripts'
Materials: available on the open science framework, https://osf.io/8hrp6/, in the folder 'Experiment_regularities', in 'Weather forecast video' (accompanied by 'The weather forecast script in English'), and in the pdf document 'questionnaire'

The Editor granted exemptions to materials sharing for the following sets of materials, on the basis that they are proprietary and subject to copyright: (1) the English vocabulary subtest of the Wechsler Adult Intelligence Scale WAIS-IV (Wechsler et al., 2008), and (2) the matrices subtest of the Wechsler Adult Intelligence Scale WAIS-III (Wechsler, 1997).

Ethics Statement

This study was granted ethical approval by the Research Ethics Committee of the UCL Institute of Education, number REC 1156, on 18th January 2019. An amendment to allow for online data collection was approved by the same committee on 20th July 2020. All participants gave informed written consent before taking part in the study.

Authorship and Contributorship Statement

Julia Hofweber: conceptualization; data curation; formal analysis; methodology; investigation; visualization; writing – original draft preparation; writing – review and editing.
Lizzy Aumônier: conceptualization; formal analysis; investigation; methodology.
Vikki Janke: conceptualization; formal analysis; funding acquisition; methodology; project administration and supervision; writing – original draft preparation; writing – review and editing.
Marianne Gullberg: conceptualization; formal analysis; funding acquisition; methodology; project administration and supervision; resources; writing – original draft preparation; writing – review and editing.
Chloë Marshall: conceptualization; formal analysis; funding acquisition; methodology; investigation; project administration and supervision; resources; writing – original draft preparation; writing – review and editing.

Acknowledgements

This work was funded by an International Academic Fellowship (IAF2016023 awarded to Chloë Marshall) and a Research Grant (RPG-2018-333 awarded to Chloë Marshall, Marianne Gullberg, and Vikki Janke) from the Leverhulme Trust. We thank Linnéa Lindén, Peter Roslund, Henrik Garde and Josien Greidanus at Lund University Humanities Lab for their assistance with creating the experimental materials, and Vassilis Sideropoulos at the UCL Institute of Education for his statistical advice.

License