Lexicalization in the developing parser

Aaron Steven White
University of Rochester
Rochester, NY
aaron.white@rochester.edu

Jeffrey Lidz
University of Maryland
College Park, MD
jlidz@umd.edu

Word count: 6,970 + 767 (appendix)

Abstract We use children's noun learning as a probe into the nature of their syntactic prediction mechanism and the statistical knowledge on which that prediction mechanism is based. We focus on verb-based predictions, considering two possibilities: children's syntactic predictions might rely on distributional knowledge about specific verbs—i.e. they might be *lexicalized*—or they might rely on distributional knowledge that is general to all verbs. In an intermodal preferential looking experiment, we establish that, by as early as 19 months of age, verb-based predictions are lexicalized: children encode the syntactic distributions of specific verbs and use those distributions to make predictions, but they do not assume that these can be assumed of verbs in general.

Keywords: language acquisition; parsing; prediction; thematic roles

1 Introduction

There is now a wealth of evidence that adult language comprehender's parsing decisions are both predictive and guided, at least in part, by a language's distributional properties (Gordon & Chafetz 1990; Trueswell et al. 1993; MacDonald et al. 1994; Garnsey et al. 1997; Altmann & Kamide 1999). A major question in this literature is how these distributions are encoded and how these encodings are deployed for prediction (McRae et al. 1998; Hale 2001; Elman et al. 2004; Levy 2008; Linzen & Jaeger 2016).

In this paper, we approach this question of encoding and deployment from a developmental perspective, asking how predictive parsing interacts with syntactic bootstrapping. By 4–5 years of age, children appear to use prediction in the course of online sentence comprehension (Trueswell et al. 1999; Snedeker & Trueswell 2004; Fernald & Marchman 2006; Lew-Williams & Fernald 2007; Omaki 2010; Mani & Huettig 2012; Borovsky et al. 2012; Huang et al. 2013; Omaki et al. 2014). The nature of this developing prediction mechanism can often be seen most clearly in cases where children display interpretive biases that disallow them either from accessing a particular adult-like interpretation of a sentence or from accessing an adult-like interpretation in the first place.

Recent work has demonstrated that children utilize such predictive parsing mechanisms for the purposes of both comprehension and learning as early as 19 months of age (Lidz et al. 2017). But it remains unclear whether this predictive parsing mechanism is based on knowledge about the distributional characteristics of particular verbs—i.e. whether distributional knowledge is *lexicalized*—or whether it is based on knowledge of the particular structures that are likely to occur, regardless of the lexical items that occur in that structure—i.e. whether distributional knowledge is *generalized*.

We investigate this question using an intermodal preferential looking experiment, showing that, by as early as 19 months of age, the predictive parsing mechanism children deploy is lexicalized. This experiment builds on a paradigm introduced by Lidz et al. (2017), which we review below.

2 Early predictive parsing and syntactic bootstrapping

Lidz et al. (2017) investigate 16- and 19-month-old children's predictive parsing mechanisms through the lens of syntactic bootstrapping (Gleitman 1990). Beginning with Brown 1957, a broad literature has shown that children use aspects of syntax to drive inferences about word meaning (see Lidz 2022 for a review). For example, children as young as 12 months have been shown to treat a novel word presented as a noun as referring to an object kind (Waxman & Booth 2001), and children as young as 18 months have been shown to expect a novel verb to refer to a category of events (He & Lidz 2017; Carvalho et al. 2019). Moreover, toddlers draw different inferences about verb meaning as a function of whether the novel verb occurs in a transitive or an intransitive clause (Naigles 1990; Yuan & Fisher 2009; Fisher et al. 2010).

Gertner & Fisher (2012) suggest that one way syntactic context is used in inferring verb meaning is through the distinct thematic relations associated with the subject and object position of a clause. The evidence they adduce for this claim is indirect, however, given that it is measured by the meaning children assigned to entire clauses rather than the noun phrases in those clauses. Lidz et al. (2017) test the link between syntactic position and thematic relation more directly by asking what meaning children assigned to a novel noun as a function of its syntactic position. In their experiments, children are exposed to sentences like (1) and (2) along with a scene involving an agent acting on a patient using an instrument.

(1) She's wiping **the tiv**.

(2) She's wiping *with* **the tiv**.

Lidz et al. find that by 16 months of age, children are able to appropriately infer that *the tiv* refers to the patient in (1) and to the instrument in (2), suggesting that knowledge of the link between syntactic position and thematic relation is in place by this age. However, at 19 months of age, children incorrectly infer that *the tiv* refers to the patient in both (1) and (2). The authors argue that 19-month-olds' incorrect inferences are driven by a ballistic predictive parsing strategy that is based on the fact that all the verbs used in the study—and as we show below, most verbs in children's input—are heavily biased toward at least taking a direct object and against only taking a prepositional phrase. This distributional bias, then, overshadows the contribution of the syntactic structure in children's noun learning because it leads them to erroneously represent (2) as though it were a simple transitive clause and consequently treat *the tiv* as though it were the direct object and hence as referring to the patient of the event.

Lidz et al. bolster this argument by showing that when 19-month-olds receive sentences that satisfy the purported prediction of a direct object, as in (3) and (4), they are able to to correctly infer that *the tiv* refers to the patient in (3) and to the instrument in (4).

(3) She's wiping **the tiv** with that thing.

(4) She's wiping that thing *with* **the tiv**.

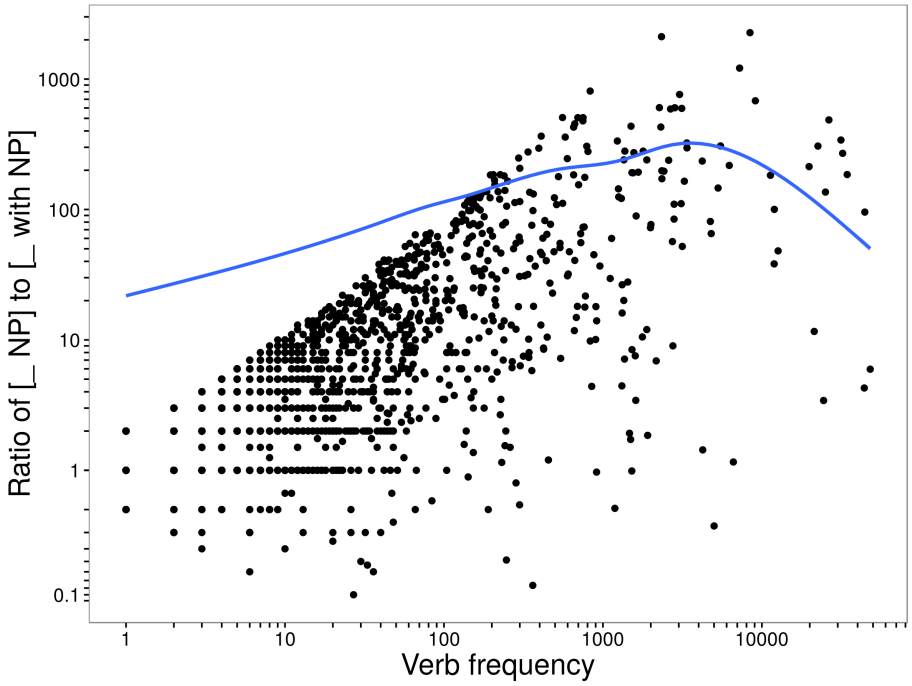

Figure 1: Ratio of [_ NP] count to [_ with NP] count by verb in child-directed speech. Blue line shows unweighted cumulative mean going from right to left. Add 1 smoothing has been applied to each verb's subcategorization frame counts to avoid zeros in the denominator..

Further supporting this predictive parsing account, they show in a post hoc analysis that 19-month-old children with smaller verb vocabularies are better able to associate *the tiv* with the correct referent in (1) and (2) than are 19-month-old children with larger verb vocabularies. One possible explanation suggested by Lidz et al. is that 19-month-old children with smaller verb vocabularies may not know the statistical distribution of known verbs well enough to use them for making predictions.

One implication of this account is that children must track distributional properties in the input. This implication raises the question of how those distributional properties are encoded: as properties of the particular verbs themselves (*lexicalized encoding*) or as properties of the category verb (*generalized encoding*).

The predictions of the generalized encoding hypothesis rely crucially on the distribution of verbs' subcategorization frame distributions in children's input. Nearly all verbs' distributions, at least in child-directed speech, turn out to be heavily biased toward transitive frames relative to intransitive frames with a prepositional phrase. This can be seen in Figure 1, which shows the ratio of [_ NP] frames to [_ with NP] extracted from all CHILDES corpora (MacWhinney 2014a; b) parsed using MEGRASP (Sagae et al. 2007). Each point in this figure is a verb, whose frequency is plotted on the x-axis. The blue line gives the unweighted cumulative mean ratio moving from right to left, with the idea that children are more likely to know higher frequency verbs. We see that this cumulative mean never dips below 10:1, suggesting a very heavy bias toward transitive frames across the frequency spectrum.

Thus, both the lexicalized encoding hypothesis and the generalized encoding hypothesis are plausible descriptions of how children encode syntactic distributions for deployment

during predictive parsing. We now describe an experiment aimed at pulling these two hypotheses apart.

3 Experiment

In this experiment, we examine how children use the syntactic context of a noun phrase (NP) to make inferences about its thematic relation. Using a word-learning task in the intermodal preferential looking paradigm (Spelke 1976; Hirsh-Pasek & Golinkoff 1999), we tested children's abilities to assign a meaning to a novel noun contained in a direct object NP as compared to a prepositional object NP. In adult English, the NP containing the novel word is interpreted as a patient in (5) but as an instrument in (6).

(5) She's meeking **the tiv**.

(6) She's meeking *with* **the tiv**.

If children are able to use this thematic role information to learn the meaning of a novel noun, in (5), we expect them to be able to link *the tiv* to the object being pushed, or in (6), to the object used to do the pushing.

This experiment is identical to Lidz et al.'s Experiment 1 up to the linguistic stimuli: we replace the known verbs they use with novel verbs. The stimuli analogous to (5) and (6) in Lidz et al.'s experiment are (1) and (2), which use the known verb *wipe*.

We do this replacement in order to test two hypotheses about how children make predictions about upcoming arguments. On the one hand, children's predictions might be lexicalized. In this case, children would use distributional information they have about a particular verb to make predictions. On the other hand, children's predictions might be generalized, in which case children would use their knowledge of the distribution of subcategorization frames that occur in all clauses, regardless of the verb found in that clause.

In the case of generalized predictions, we would expect 19-month-old children to use the same predictive mechanism to parse (5) and (6) as they do to parse (1) and (2), which contain the real verb *wipe*. This would mean that 19-month-olds who hear (5) or (6) would always associate *the tiv* with the patient, as they did in Lidz et al.'s Experiment 1. In contrast, in the case of verb-specific or lexicalized predictions, we would instead expect 19-month-old children to use a distinct predictive mechanism—or no predictive mechanism at all—to parse (5) and (6), since children do not have information about the distributional properties of the novel verb *meek*. This means that 19-month-olds that hear (5) or (6) will associate *the tiv* with the correct referent, similar to 16-month-old children in Lidz et al.'s Experiment 1 and 19-month-old children in their Experiment 3.

One possibility that arises here is that vocabulary knowledge may condition the parsing mechanism that children deploy. This is plausible in light of Lidz et al.'s finding that 19-month-old children with smaller verb vocabularies are better able to associate *the tiv* with the correct referent in (1) and (2) than are 19-month-old children with larger verb vocabularies. Here, we assess the possibility that a similar conditioning may be found in our paradigm by collecting information about children's vocabulary knowledge to be used in our analysis.

3.1 Method

3.1.1 Apparatus and procedure

Each child arrived with his/her parent and was entertained by a researcher with toys while another researcher explained the experiment to the parent and obtained informed consent. The child and parent were then escorted into a sound proof room, where the child was either seated on the parent's lap or in a high chair, centered six feet from a 51" television, where the stimuli were presented at the child's eye-level. If the children were on the parents' laps, the parents wore visors to keep them from seeing what was on the screen. Each experiment lasted approximately 5 minutes, and the children were given a break if they were too restless or started crying. In the case that the child did not complete the experiment or were extremely fussy over the entire course, this was noted for later exclusion from the sample.

The child was recorded during the entire experiment using a digital camcorder with a sample rate of 30 frames/second centered over the screen. A researcher watched the entire trial with the audio off on a monitor in an adjacent room and was able to control the camcorder's pan and zoom in order to keep the child's face in focus throughout the trial. Videos were then coded offline frame-by-frame for direction of look by a research assistant blind to the experimental condition and without audio using the SuperCoder program (Hollich 2005).

3.1.2 Design

Our design and stimuli were exactly the same as those used by Lidz et al. (2017) except for the audio stimuli. Participants were presented with eight trials, each involving a different verb and concomitant scene. Each of these trials was separated into two phases: the familiarization phase and the test phase. These phases are described below and Table 1 gives a sample script.

3.1.2.1 Familiarization Phase

During the familiarization phase, children were shown videos of 15 second dynamic scenes involving three objects: a human hand, an instrument manipulated by the hand, and a patient causally affected via the instrument. A recorded linguistic stimulus of the form either *she's* VERB*ing the* NOVEL NOUN (*V NP*) or *she's* VERB*ing with the* NOVEL NOUN (*V with NP*) was associated with each scene. Each of these pairings constitute a level in the between-subjects STRUCTURE factor. VERB and NOVEL NOUN in these frames were replaced with a known verb and a novel noun. All linguistic stimuli were recorded by the same adult female that recorded the stimuli for Lidz et al.'s experiments. The linguistic stimulus was presented three times as the scene progressed with different lead-in words—e.g. *Look!*.

3.1.2.2 Test Phase

A blank screen was then shown for two seconds after each scene, during which the question *where's the* NOVEL NOUN*?* was asked once. The test video began at the offset of the novel noun in the first of these questions, when a screen with separate static images of both the instrument and the patient from the previous dynamic scene was displayed. One of these images took up approximately one third both by-width and by-height of the left

Phase	Length	Video	Audio
Pre-trial	2 seconds 5 seconds	Blank screen Smiling baby	*Silence* [Baby giggle]
Familiarization	15 seconds	Camera being wiped by a cloth	Hey, look at that! She's meeking (with) the tig! Wow, do you see her meeking (with) the tig? Yay, she's meeking (with) the tig!
Test	2 seconds 2 seconds 3 seconds	Blank screen Split screen: camera and cloth	Where's the tig? *Silence* Which one's the tig?

Table 1: An example of a single test trial.

portion of the screen and the other took up approximately one third by-width and by-height of the right portion, with an approximately one-third by-width separation in the middle of the screen. The side on which the instrument appeared was counterbalanced and pseudorandomized such that the instrument did not show up on the same side more than twice in a row.

Two seconds after the two images were presented, the question—*which one's the* NOVEL NOUN*?*—was played. The split screen was presented for five seconds total, after which the screen went blank. After a two second blank screen, either the next learning phase started or an attention-getting phase involving a picture of a child and laughter was presented.

3.1.3 Materials

Eight verbs contained in the MCDI checklist were chosen with the criterion that their associated event concept must support the use of an instrument. Eight novel nouns were constructed and one associated with each verb. Table 1 gives a sample script summarizing the above description. In the *V with NP* conditions, children heard *with* during the familiarization, while those in the *V NP* conditions did not, represented in the table by the parentheses.

Table 2 shows each tuple of verb, novel noun, instrument object, and patient object. To control for possible order effects, we created two presentation orders for the trials by first building one pseudorandomized order according to the above sequencing criterion, then inverting it to create the second order. When crossed with the two linguistic structure levels (STRUCTURE: *V NP*, *V with NP*), this yielded four stimulus sets.

3.2 Participants

We recruited 32 19-month-olds (16 females) with a median age of 19;15.5 (mean: 19;16.1, range: 19;0 to 20;0).[1] Six additional participants were tested but were excluded from the final sample prior to analysis for fussiness or inability to complete the experiment. Participants were recruited from the greater College Park, MD area and were acquiring English as a native language. All participants heard English at least 80% of the time. Par-

[1] Appendix A reports simulation-based *post hoc* power calculations for the crucial statistical test reported in Section 3.4. Based on these calculations, the sample size reported above provides approximately 50% power for that test, and a sample size of 64 would be required for 80% power.

Action	Instrument	Patient	Verb	Noun
wipe	cloth	camera	*meek*	*tig*
throw	cup	ball	*doadge*	*frap*
hit	ruler	cone	*lonk*	*tam*
push	bulldozer	block	*tiz*	*gop*
touch	pipe cleaner	pumpkin	*rem*	*pint*
wash	sponge	toy car	*sloob*	*pud*
tickle	feather	mouse puppet	*chiff*	*seb*
pull	fishing pole	train	*stip*	*wug*

Table 2: The verbs and novel nouns used in the linguistic stimuli and the objects used in the visual stimuli for Exps. 1 and 2.

ticipants within each age group and sex were distributed evenly across the four stimulus sets.

Parents completed the MacArthur-Bates Communicative Development Inventory (MCDI) checklist (Fenson 2007). By this index, participants' median productive verb vocabulary was 5 verbs (mean: 16 verbs, IQR: 1–30 verbs), and their median productive total vocabulary was 63 words (mean: 139.5 words, IQR: 41–251 words). The parent of one participant in the *V NP* condition did not submit an MCDI checklist, and for the purposes of analysis, that participant's verb vocabulary value was set to the mean across participants (but excluded from the above statistics).

3.3 Measures

Following Lidz et al., we compute two measures for each trial each child received. The first measure (FAMILIARIZATION PROPORTION) is the proportion of the time each child was looking at the screen during the familiarization phase for a given trial. This measure provides a proxy for how well the child was paying attention to the pairing of the linguistic stimulus with the scene in the video. We expect that the less a child pays attention during a particular familiarization, the less likely it is that their behavior during the test phase associated with that familiarization provides evidence about the inferences they make based on the linguistic stimuli.

The second measure (OBJECT COUNT) is the number of video frames on which each child was looking at the instrument (LOOKS TO INSTRUMENT) paired with the number of frames on which they were looking at the patient (LOOKS TO PATIENT) on each trial.[2] This was calculated by converting the left-right coding of the test phase into an instrument-patient coding and then computing the relevant counts by trial for each child. Note that, unlike the first measure, this second measure is not a proportion, though we can compute a proportion from it. For the purposes of visualization and basic comparisons of means, we work with proportions computed from these counts; for the purposes of more fine-grained analysis, we work with the counts themselves.

In addition to the measures used by Lidz et al., we also compute two measures of vocab based on verb vocabulary and total vocabulary in MCDI. Because verb vocabulary and

[2] Note that, because children do not necessarily look at the screen during the entire test phase, the sum of LOOKS TO INSTRUMENT and LOOKS TO PATIENT will not necessarily be the number of frames in the test phase. This is in fact a feature of OBJECT COUNT as a measure, since it retains information about the relative amount of data from which a probability is computed, where analyzing the proportion directly does not.

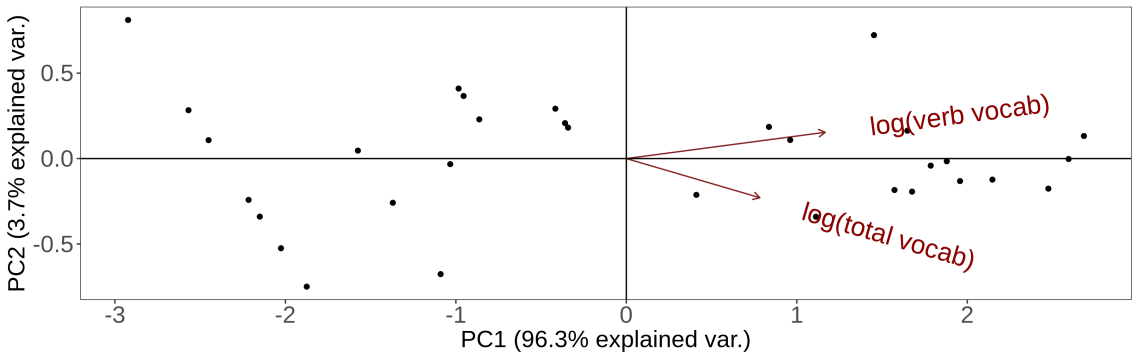

Figure 2: Biplot for principal component analysis of logged vocabulary. Each point shows the score for a child on each component and the red vectors show the loadings for each variable on each component.

total vocabulary are highly correlated ($r = 0.92$), they cannot be entered into our analyses in their raw forms without giving rise to issues of collinearity. As such, we first apply principal component analysis to the logged form of these two measures.

Figure 2 shows the biplot for this analysis. The first principal component (PC1), which explains over 96% of the variance in the logged vocabulary measures, loads positively on both verb vocabulary and total vocabulary. The fact that this component explains so much of the variance in the logged vocabulary measures is unsurprising in light of their extremely high correlation, The second principal component (PC2), which explains less than 4% of the variance in the logged vocabulary measures, loads positively on verb vocabulary but negatively on total vocabulary.

This pattern might be taken to indicate that the first component provides a measure of overall vocabulary knowledge while the second component provides a measure of verb knowledge adjusted for this overall knowledge. But caution is warranted here in light of the fact that the first principal component explains over 96% of the variance in the logged vocabulary measures, likely indicating that it is not possible to distinguish any effects of verb knowledge from the effect of total vocabulary knowledge.[3] Said another way, observing an effect of PC1 is consistent with observing an effect of verb knowledge, though it does not imply it. In the name of due diligence, however, we include both PC1 and PC2 in our statistical analyses with the caveat that PC2 is likely uninteresting because it explains so little variance—indeed, it may merely be capturing noise.

For the purposes of reporting statistics, we use the continuous form of both variables. For the purpose of visualization, we discretize the first principal component at its median, referring to the group of children that have a vocabulary score above the median as the high vocab group and the group of children that have a vocabulary score below

[3] An anonymous reviewer suggests that the effect of verb vocabulary knowledge and total vocabulary knowledge might be distinguished using a residualization strategy: residualize one vocabulary variable against the other then analyze the effects of residualized variable and the raw variable it was residualized against. Unfortunately, this method does not help in this context exactly because the variables are so highly correlated. Residualizing either total vocabulary against verb vocabulary or vice versa will necessarily result in two variables that are very slight rotations of the principal components shown above: whichever variable is left in its raw form will necessarily be very highly correlated with the first principal component—log(verb vocabulary) has a 0.99 correlation with PC1 and log(total vocabulary) has a 0.96 correlation—and whichever variable is residualized will necessarily be very close to the second. Thus, residualization not only introduces an additional researcher degree of freedom—the direction in which to residualize—but also raises the likelihood of misinterpretation: seeing a reliable effect for the raw variable does not mean that that variable indeed has an effect to the exclusion of the other. See Wurm & Fisicaro 2014 for further discussion.

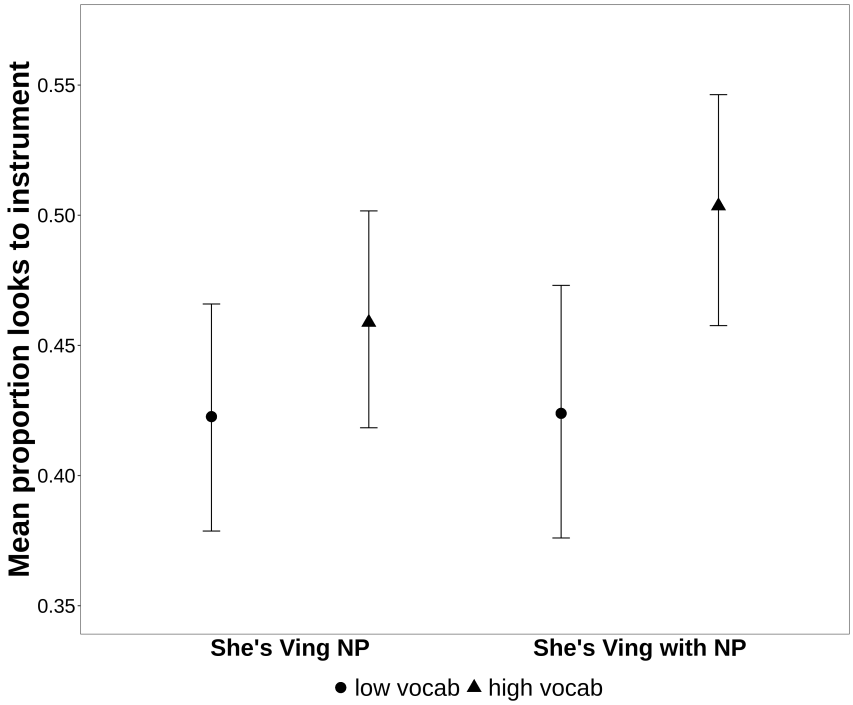

Figure 3: Mean proportion looks to instrument by STRUCTURE and discretized PC1. Error bars show 95% confidence intervals computed from nonparametric bootstrap on participant weighted means.

the median as the low vocab group, since scoring more positively on the first principal component implies having a larger total vocabulary and a larger verb vocabulary.

3.4 Results

Figure 3 plots the mean proportion of looks to instrument by STRUCTURE and discretized PC1. The confidence intervals in Figure 3 are computed from a nonparametric bootstrap of the condition mean with 999 iterations. In this bootstrap, children's mean proportion of looks to instrument across trials, weighted by FAMILIARIZATION PROPORTION, was first computed and then these mean proportions were resampled. Qualitatively, this plot appears to support a hypothesis wherein children with larger vocabularies are able to correctly map direct objects to patients and prepositional objects to instruments, but children with smaller vocabularies are not.

To assess the reliability of this pattern, we follow Lidz et al. in using a logistic mixed effects model with OBJECT COUNT as the dependent variable, random intercepts for child and item, by-item random slopes for STRUCTURE, and a loss weighted by FAMILIARIZATION PROPORTION. We first fit such a model with fixed effects for STRUCTURE, PC1, and PC2 as well as the two-way interaction between STRUCTURE and PC1 and the two-way interaction between STRUCTURE and PC2. We test the reliability of these interactions using a log-likelihood ratio test. We find that the model that includes both interactions is reliably better than the one that does not include the interaction between STRUCTURE and PC1 ($\chi^2(1) = 3.98$, $p < 0.05$) but a similar pattern is not observed for the interaction between STRUCTURE and PC2 ($\chi^2(1) = 0.28$, $p = 0.60$). Thus, the apparent interaction between STRUCTURE and PC1 seen in Figure 3 is reliable.

3.5 Discussion

In a novel verb variant of Lidz et al.'s Experiment 1, we found a pattern of results opposite to what they found with real verbs: 19-month-olds with smaller vocabularies fail to map NPs to the correct referent based on the structure they are found in, while 19-month-olds with larger vocabularies succeed, mapping the NP in the *V NP* condition to the patient and the NP in the *V with NP* condition to the patient. Why might we find such an opposite pattern?

Lidz et al. argue that 19-month-olds with larger vocabularies fail in the real verb experiment due to a predictive parsing strategy in combination with an inability to revise predictions. But the pattern they observe is consistent with this predictive parsing strategy being based on either a lexicalized encoding or a generalized encoding, since 19-month-olds with smaller vocabularies likely do not have sufficient evidence for either type of encoding while 19-month-olds with larger vocabularies likely have sufficient evidence for both. In our novel verb experiment, regardless of vocabulary size, children could not have enough distributional knowledge about the particular verb to deploy it in prediction, since they could not have distributional knowledge about the particular verb at all. We have in effect put all 19-month-olds into the same position 16-month-olds were in in Lidz et al.'s experiments.

The success of 19-month-olds with large vocabularies in this context thus provides evidence that these children's parsing predictions are based on a lexicalized encoding, not a generalized one. If these children's predictions were based on a generalized encoding, they should always predict a direct object and thus fail in the same way 19-month-olds with large vocabularies failed in Lidz et al.'s Experiment 1.

What the failure of 19-month-olds with smaller vocabularies implies is less clear. One possibility is that their failure is not indicative of the predictive parsing strategy these children use at all. It may simply be that having to process two novel words at once—both a verb and a noun—is particularly burdensome for children with smaller vocabularies for whatever reason it is that they have smaller vocabularies in the first place. Depending on what this reason is, this account might predict either that all 16-month-olds would similarly fail in our experiment—e.g. if the failure is simply about amount of vocabulary knowledge—or that, similar to the results of our experiment, 16-month-olds with larger vocabularies would succeed but those with smaller vocabularies would fail—e.g. because differences in vocabulary knowledge at a particular age index cognitive resources relevant to processing two novel words at once.

The second possibility is that 19-month-olds with smaller vocabularies—unlike those with larger vocabularies—make predictions in our experiment based on verb-general knowledge—plausibly because they are less certain about those specific verbs' distributional properties. This uncertainty might arise in two different ways: (i) children who know fewer verbs tend to have less experience with the verbs they do know—e.g. because less vocabulary knowledge is indicative that the verbs that they do know were more recently learned; or (ii) children who know fewer verbs need additional evidence about a specific verb to become certain enough about its distribution to use that distribution in predictive parsing. This second version might be plausible insofar as knowledge of verbs' distributional properties is hierarchical (Perfors et al. 2010) and thus children who know more verbs require less evidence to acquire the distributional properties of a verb whose distribution is prototypical relative to the verbs those children already know.

A major hurdle faced by either version of this account is that, if 19-month-olds with smaller vocabularies make predictive parsing decisions based on generalized encodings and thus fail in our experiment, it is unclear why they do not similarly do so in Lidz

et al.'s Experiment 1. Why should they not fail in that experiment as well? To overcome this hurdle, such an account would likely need to posit that 19-month-olds with smaller vocabularies are unable to deploy predictive parsing for known verbs—e.g. because they attempt to make predictions on the basis of lexicalized encodings but fail to do so in the face of the uncertainty inherent to that knowledge.

One way to test this account might be to turn novel verbs in our experiment into "known" verbs by exposing children to dialogues containing the novel verb and then testing their noun learning using the same stimuli we use here (Yuan & Fisher 2009; Arunachalam & Waxman 2010; Yuan et al. 2011). If the sentences in which these novel verbs are found in these dialogues are heavily biased toward having transitive structures, 19-month-olds may gain a lexicalized encoding for those novel verbs that they can then deploy in predictive parsing, therefore causing them to fail in the same way 19-month-olds with larger vocabularies did in Lidz et al.'s Experiment 1.

The first version of the account predicts that, insofar as the dialogues contain a sufficient number of examples of the novel verb for children to form a lexicalized encoding, they will fail regardless of vocabulary knowledge. In contrast, the second version of the account predicts that, insofar as the dialogues contain a sufficient number of examples of the novel verb for children with larger vocabularies (but not smaller vocabularies) to form a lexicalized encoding, children should behave as they did in Lidz et al.'s Experiment 1: those with larger vocabularies should now fail—because, like their counterparts in Lidz et al.'s Experiment 1, they have formed a lexicalized encoding that they now deploy in predictive parsing—but those with smaller vocabularies should succeed—because, like their counterparts in Lidz et al.'s Experiment 1, they will not make predictions for the newly "known" verbs due to remaining uncertainty about their distributional properties.

A crucial component of designing such an experiment is determining the correct number of items to include in the dialogues. This choice is important for both accounts, but it is particularly important for the second version: there must be enough examples for children with larger vocabularies to form a lexicalized encoding but not so many that children with smaller vocabularies can similarly do so. Thus, insofar as an account is to be pursued wherein the kind of distributional information children deploy in predictive parsing is modulated by vocabulary knowledge, a crucial next step is to develop finer-grained predictions about the amount of evidence children at different stages of development require to construct lexicalized encodings with high certainty. Combining hierarchical models of argument structure knowledge with probabilistic parsers may be a fruitful next step.

4 Conclusion

The study just reported adds support to the view that 19-month-olds have knowledge of the link between syntactic position and thematic relation. The fact that they can use the syntactic position of an NP to assign it an interpretation supports theories of word learning that treat syntactic structure as informative (Gleitman 1990), and more indirectly, theories of verb-learning that use the thematic relations of the NPs in a clause as evidence about the meaning of the verb (Gertner & Fisher 2012; Perkins 2019). However, 19-month-olds' ability to deploy the link between syntactic position and thematic relation can be disrupted during sentence comprehension by lexicalized knowledge of verb-argument structure. Whereas prior work showed that 16-month-olds, but not 19-month-olds, successfully mapped a novel noun phrase to different referents depending on its syntactic position, the current work shows that 19-month-olds' failure in previous

work resulted from their knowledge of specific verb distributions. In the current study, 19-month-olds with larger vocabularies were able to correctly identify the referent of a novel noun phrase as a function of syntactic position even with novel verbs. The fact that having a larger vocabulary helped these children to avoid a parsing error with novel verbs suggests that their prior failures derive from knowledge of specific verb distributions and not from a general knowledge that transitive clauses are more likely than intransitive clauses.

The finding that 19-month-olds' syntactic predictions are driven by lexicalized subcategorization frequencies comports well with work from older children and adults (Trueswell et al. 1993; Trueswell & Kim 1998; Snedeker & Trueswell 2004; Altmann & Kamide 2007; Borovsky et al. 2012). It further adds to this literature by showing that lexically driven syntactic predictions occur from the earliest stages of language development. As soon as children have acquired lexical statistics, they appear to use that information to drive parsing predictions.

Our data also informs a debate concern the origins of children's early syntactic knowledge. To what degree is early syntactic knowledge associated with specific lexical items (Tomasello & Kruger 1992; Theakston et al. 2015; Lieven 2016) and to what degree does syntactic knowledge abstract away from specific lexical items (Gertner et al. 2006; Naigles 2002; Fisher et al. 2010; Viau & Lidz 2011)? Our data suggests that syntactic knowledge begins with abstract categories and that lexically specific distributional information informs the development of parsing strategies, but not the knowledge itself. That knowledge is revealed when we take away children's ability to rely on lexically specific knowledge, as in the current study.

Abbreviations

V = verb, NP = noun phrase

Funding information

This work was supported by NSF BCS-1551629 (*Transitivity of Sentences and Scenes in Early Language Development*).

Acknowledgements

We are grateful to audiences at the University of Maryland, BUCLD 34, BUCLD 35, LSA 2010, CUNY 2011, and SRCD 2013 for useful comments on this work.

Competing interests

The authors have no competing interests to declare

References

Altmann, Gerry & Yuki Kamide. 1999. Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition* 73(3). 247–264.

Altmann, Gerry & Yuki Kamide. 2007. The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language* 57(4). 502–518.

Arunachalam, Sudha & Sandra R. Waxman. 2010. Meaning from syntax: Evidence from 2-year-olds. *Cognition* 114(3). 442–446.

Borovsky, Arielle, Jeffrey L. Elman & Anne Fernald. 2012. Knowing a lot for one's age: Vocabulary skill and not age is associated with anticipatory incremental sentence interpretation in children and adults. *Journal of Experimental Child Psychology* 112(4). 417–436.

Brown, Roger. 1957. Linguistic determinism and the part of speech. *Journal of Abnormal and Social Psychology* 55(1). 1.

Carvalho, Alex de, Angela Xiaoxue He, Jeffrey Lidz & Anne Christophe. 2019. Prosody and Function Words Cue the Acquisition of Word Meanings in 18-Month-Old Infants. *Psychological Science* 30(3). 319–332. https://doi.org/10.1177/0956797618814131. https://doi.org/10.1177/0956797618814131. _eprint: https://doi.org/10.1177/0956797618814131.

Elman, Jeffrey L., Mary Hare & Ken McRae. 2004. Cues, constraints, and competition in sentence processing. *Beyond nature-nurture: Essays in honor of Elizabeth Bates* 111–138.

Fenson, Larry. 2007. *MacArthur-Bates Communicative Development Inventories: User's guide and technical manual*. Baltimore: Paul H. Brookes Publishing Company.

Fernald, Anne & Virginia A. Marchman. 2006. Language Learning in Infancy. In Matthew J. Traxler & Morton A. Gernsbacher (eds.), *Handbook of Psycholinguistics*, 1027–1071. London: Academic Press 2nd edn. https://doi.org/10.1016/B978-012369374-7/50028-6. https://www.sciencedirect.com/science/article/pii/B9780123693747500286.

Fisher, Cynthia, Yael Gertner, Rose M. Scott & Sylvia Yuan. 2010. Syntactic bootstrapping. *Wiley Interdisciplinary Reviews: Cognitive Science* 1(2). 143–149.

Garnsey, Susan M., Neal J. Pearlmutter, Elizabeth Myers & Melanie A. Lotocky. 1997. The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language* 37(1). 58–93.

Gertner, Yael & Cynthia Fisher. 2012. Predicted errors in children's early sentence comprehension. *Cognition* 124(1). 85–94.

Gertner, Yael, Cynthia Fisher & Julie Eisengart. 2006. Learning Words and Rules: Abstract Knowledge of Word Order in Early Sentence Comprehension. *Psychological Science* 17(8). 684–691. https://doi.org/10.1111/j.1467-9280.2006.01767.x. https://doi.org/10.1111/j.1467-9280.2006.01767.x. Publisher: SAGE Publications Inc.

Gleitman, Lila. 1990. The structural sources of verb meanings. *Language Acquisition* 1(1). 3–55.

Gordon, Peter & Jill Chafetz. 1990. Verb-based versus class-based accounts of actionality effects in children's comprehension of passives. *Cognition* 36(3). 227–254.

Hale, John. 2001. A Probabilistic Earley Parser As a Psycholinguistic Model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies* (NAACL '01). 1–8. Stroudsburg, PA, USA: Association for Computational Linguistics.

He, Angela Xiaoxue & Jeffrey Lidz. 2017. Verb Learning in 14- and 18-Month-Old English-Learning Infants. *Language Learning and Development* .

Hirsh-Pasek, Kathryn & Roberta Michnick Golinkoff. 1999. *The Origins of Grammar: Evidence from early language comprehension*. Cambridge, MA: The MIT Press.
Hollich, George. 2005. Supercoder: A program for coding preferential looking.
Huang, Yi Ting, Xiaobei Zheng, Xiangzhi Meng & Jesse Snedeker. 2013. Children's assignment of grammatical roles in the online processing of Mandarin passive sentences. *Journal of Memory and Language* 69(4). 589–606.
Levy, Roger. 2008. Expectation-based syntactic comprehension. *Cognition* 106(3). 1126–1177.
Lew-Williams, Casey & Anne Fernald. 2007. Young children learning Spanish make rapid use of grammatical gender in spoken word recognition. *Psychological Science* 18(3). 193–198.
Lidz, Jeffrey. 2022. Children's use of syntax in word learning. In Anna Papafragou, John C. Trueswell & Lila Gleitman (eds.), *The Oxford Handbook of the Mental Lexicon*, Oxford: Oxford University Press.
Lidz, Jeffrey, Aaron Steven White & Rebecca Baier. 2017. The role of incremental parsing in syntactically conditioned word learning. *Cognitive Psychology* 97. 62–78. https://doi.org/10.1016/j.cogpsych.2017.06.002. http://www.sciencedirect.com/science/article/pii/S0010028516302742.
Lieven, Elena. 2016. Usage-based approaches to language development: Where do we go from here? *Language and Cognition* 8(03). 346–368.
Linzen, Tal & T. Florian Jaeger. 2016. Uncertainty and expectation in sentence processing: evidence from subcategorization distributions. *Cognitive science* 40(6). 1382–1411.
MacDonald, Maryellen C., Neal J. Pearlmutter & Mark S. Seidenberg. 1994. The lexical nature of syntactic ambiguity resolution. *Psychological Review* 101(4). 676.
MacWhinney, Brian. 2014a. *The CHILDES project: Tools for Analyzing Talk*, vol. 2. New York: Psychology Press.
MacWhinney, Brian. 2014b. *The CHILDES project: Tools for Analyzing Talk*, vol. 1. New York: Psychology Press.
Mani, Nivedita & Falk Huettig. 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. ISBN: 0096-1523.
McRae, Ken, Michael J. Spivey-Knowlton & Michael K. Tanenhaus. 1998. Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language* 38(3). 283–312.
Naigles, Letitia. 1990. Children use syntax to learn verb meanings. *Journal of Child Language* 17(2). 357–374.
Naigles, Letitia R. 2002. Form is easy, meaning is hard: resolving a paradox in early child language. *Cognition* 86(2). 157–199. https://doi.org/10.1016/S0010-0277(02)00177-4. https://www.sciencedirect.com/science/article/pii/S0010027702001774.
Omaki, Akira. 2010. *Commitment and Flexibility in the Developing Parser*: University of Maryland dissertation.
Omaki, Akira, Imogen Davidson White, Takuya Goro, Jeffrey Lidz & Colin Phillips. 2014. No fear of commitment: Children's incremental interpretation in English and Japanese wh-questions. *Language Learning and Development* 10(3). 206–233.
Perfors, Andrew, Joshua B. Tenenbaum & Elizabeth Wonnacott. 2010. Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of Child Language* 37(3). 607–642. https://doi.org/10.1017/S0305000910000012. http://www.cambridge.org/core/journals/journal-of-child-language/article/

variability-negative-evidence-and-the-acquisition-of-verb-argument-constructions/D62EDBFF5A8F1ACC821451FEAD3C88FB. Publisher: Cambridge University Press.

Perkins, Laurel. 2019. *How Grammars Grow: Argument Structure and the Acquisition of Nonbasic Syntax*. College Park, MD: University of Maryland, College Park dissertation.

Sagae, Kenji, Eric Davis, Alon Lavie, Brian MacWhinney & Shuly Wintner. 2007. High-accuracy annotation and parsing of CHILDES transcripts. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*. 25–32. Prague, Czech Republic: Association for Computational Linguistics.

Snedeker, Jesse & John C. Trueswell. 2004. The developing constraints on parsing decisions: The role of lexical-biases and referential scenes in child and adult sentence processing. *Cognitive Psychology* 49(3). 238–299.

Spelke, Elizabeth. 1976. Infants' intermodal perception of events. *Cognitive Psychology* 8(4). 553–560.

Theakston, Anna L., Paul Ibbotson, Daniel Freudenthal, Elena VM Lieven & Michael Tomasello. 2015. Productivity of noun slots in verb frames. *Cognitive Science* 39(6). 1369–1395.

Tomasello, Michael & Ann Cale Kruger. 1992. Joint attention on actions: Acquiring verbs in ostensive and non-ostensive contexts. *Journal of Child Language* 19(02). 311–333.

Trueswell, John C. & Albert E. Kim. 1998. How to prune a garden path by nipping it in the bud: Fast priming of verb argument structure. *Journal of Memory and Language* 39(1). 102–123.

Trueswell, John C., Irina Sekerina, Nicole M. Hill & Marian L. Logrip. 1999. The kindergarten-path effect: Studying on-line sentence processing in young children. *Cognition* 73(2). 89–134.

Trueswell, John C., Michael K. Tanenhaus & Christopher Kello. 1993. Verb-specific constraints in sentence processing: separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 19(3). 528.

Viau, Joshua & Jeffrey Lidz. 2011. Selective learning in the acquisition of Kannada ditransitives. *Language* 87(4). 679–714.

Waxman, Sandra R. & Amy E. Booth. 2001. Seeing pink elephants: Fourteen-month-olds' interpretations of novel nouns and adjectives. *Cognitive Psychology* 43(3). 217–242.

Wurm, Lee H. & Sebastiano A. Fisicaro. 2014. What residualizing predictors in regression analyses does (and what it does not do). *Journal of Memory and Language* 72. 37–48. https://doi.org/https://doi.org/10.1016/j.jml.2013.12.003. https://www.sciencedirect.com/science/article/pii/S0749596X13001368.

Yuan, Sylvia & Cynthia Fisher. 2009. "Really? She Blicked the Baby?" Two-Year-Olds Learn Combinatorial Facts About Verbs by Listening. *Psychological Science* 20(5). 619–626.

Yuan, Sylvia, Cynthia Fisher, Padmapriya Kandhadai & Anne Fernald. 2011. You can stipe the pig and nerk the fork: Learning to use verbs to predict nouns. In *Proceedings of the 35th annual Boston University Conference on Language Development*. 665–677. Somerville, MA: Cascadilla Press.

A Simulation-based *post hoc* power analysis

In order to guide future experiments, we conduct a *post hoc* power analysis for the log-likelihood ratio test of the interaction between STRUCTURE and PC1 reported in Section 3.4. Because we cannot compute power for this test analytically, we take a simulation-based approach. The simulation closely follows the assumptions of the underlying mixed effects model within which the interaction is tested: in addition to simulating independent variables—e.g. total vocabulary and verb vocabulary—based on distributions observed in our sample, we also simulate the random effects associated with participants and items based on the (co)variance estimates obtained for the random effects in the full model fits reported in Section 3.4.

A.1 Simulating participants

Simulating participants requires randomly assigning each simulated participant to a condition and sampling three quantities for each simulated participant: total vocabulary, verb vocabulary, and the participant random intercept–i.e. the participant's bias to look more toward patients or instruments, irrespective of the linguistic stimulus.

A.1.1 Simulating vocabulary knowledge

To simulate vocabulary knowledge, we first fit a negative binomial distribution to the total vocabulary counts for the 32 children in our original experiment, and we regress verb vocabulary on logged total vocabulary in a zero-inflated negative-binomial regression. For each simulated participant, we then (i) sample a total vocabulary count from the negative binomial distribution fit to total vocabulary; and (ii) sample a verb vocabulary count given a total vocabulary count by using the zero-inflated negative-binomial regression to compute a distribution over verb vocabulary counts from the logged total vocabulary count and then sampling the verb vocabulary count from that distribution. We convert these vocabulary measures into principal component scores by first logging them, then applying the principal component analysis fit in Section 3.3.

A.1.2 Simulating participant looking biases

To simulate underlying looking biases, we sample from a normal distribution with 0 mean and standard deviation equal to the standard deviation of by-participant random intercepts ($\hat{\sigma}_{\text{part-inter}} = 0.20$) in the full model described in Section 3.4.

A.2 Simulating items

Simulating items requires sampling two quantities: the item random intercept and the item random slope for STRUCTURE. To simulate these two quantities, we sample from a multivariate normal distribution with $[0, 0]$ mean and covariance equal to the covariance of by-item random effects ($\hat{\Sigma} = \begin{bmatrix} 0.11 & -0.08 \\ -0.079 & 0.18 \end{bmatrix}$) estimated in the full model described in Section 3.4, with *V NP* as the reference level in a dummy coding of STRUCTURE. This estimated covariance implies an estimated standard deviation for the random intercept of $\hat{\sigma}_{\text{item-inter}} = 0.33$, an estimated standard deviation for the random slopes of $\hat{\sigma}_{\text{item-slope}} = 0.42$ and a correlation between the two of -0.56.

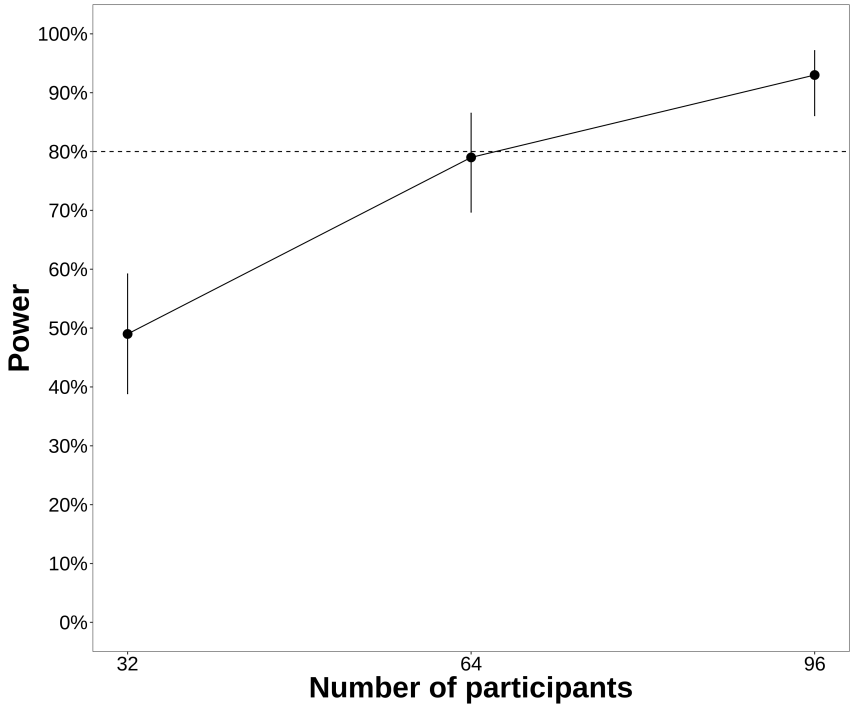

Figure 4: Estimated power for the log-likelihood ratio test of the interaction between STRUCTURE and PC1 reported in Section 3.4, varying the number of participants but keeping the number of items constant at 8. Estimates are based on 100 simulated datasets, and confidence intervals are computed using the Clopper-Pearson method.

A.3 Simulating looks

To simulate the looks each simulated participant gives in a trial containing each simulated item, we use the estimates of the fixed effect coefficients from the full model fit in Section 3.4 in conjunction with the simulated random effects estimates to compute a log-odds of looking to instrument v. looking to patient for each participant in each trial. We then sample 150 looks given the probability of looking to instrument computed from those log-odds. 150 was chosen based on the length of the test phase in each trial—5 seconds, excluding the 2 seconds of blank screen—and the sample rate of the camera used: 30 samples per second. Not all children look at the split screen during the entire test phase, and so not all trials have 150 observations. We do not attempt to simulate looks away from the split screen or differential attention in the familiarization phase.

A.4 Calculating power

We calculate power varying the number of simulated participants but keeping the number of simulated items constant at 8 (the number of items in the actual experiment). We consider simulations with 32 participants (the number in the actual experiment), 64 participants, and 96 participants. For each number of participants, we simulate 100 datasets using the procedure described above and fit to each simulated dataset (a) the full model described in Section 3.4; and (b) the full model without the interaction between STRUCTURE and PC1. We compute the p-value from the log-likelihood ratio test comparing these two models. Figure 4 shows the power estimates, assuming $\alpha = 0.05$. We see that to achieve 80% power, future studies would need approximately 64 participants.