

RESEARCH

Open Access



Assessing receptive verb knowledge in late talkers and autistic children: advances and cautionary tales

Sabrina Horvath^{1*}  and Sudha Arunachalam²

Abstract

Purpose Using eye-tracking, we assessed the receptive verb vocabularies of age-matched late talkers and typically developing children (experiment 1) and autistic preschoolers (experiment 2). We evaluated how many verbs participants knew and how quickly they processed the linguistic prompt. Our goal is to explore how these eye-gaze measures can be operationalized to capture verb knowledge in late talkers and autistic children.

Method Participants previewed two dynamic scenes side-by-side (e.g., “stretching” and “clapping”) and were then prompted to find the target verb’s referent. Children’s eye-gaze behaviors were operationalized using established approaches in the field with modifications in consideration for the type of stimuli (dynamic scenes versus static images) and the populations included. Accuracy was calculated as a proportion of time spent looking to the target, and linguistic processing was operationalized as latency of children’s first look to the target.

Results In experiment 1, there were no group differences in the proportion of verbs known, but late talkers required longer to demonstrate their knowledge than typically developing children. Latency was predicted by age but not language abilities. In experiment 2, autistic children’s accuracy and latency were both predicted by receptive language abilities.

Conclusion Eye gaze can be used to assess receptive verb vocabulary in a variety of populations, but in operationalizing gaze behavior, we must account for between- and within-group differences. Bootstrapped cluster-permutation analysis is one way to create individualized measures of children’s gaze behavior, but more research is warranted using an individual differences approach with this type of analysis.

Keywords Late talkers, Autism spectrum disorder, Vocabulary, Verbs, Assessment, Eye-tracking

Introduction

Advances in eye-tracking technology and research highlight a tantalizing prospect: It may be possible to develop a standardized vocabulary assessment that relies on eye-tracking. Such an assessment would be beneficial

for evaluating language knowledge in populations who have difficulty with pointing or vocalizing to indicate their knowledge, including children with motor impairments such as cerebral palsy [1, 2], autistic children [3–5], or young infants [6, 7]. Children in these populations are known to struggle with standardized assessments because of the behavioral demands of the task, often-times being unable to participate (e.g., [4, 5]). In this way, an eye-tracking assessment is an ideal solution to a major clinical need.

For example, eye-tracking could be used as part of an assessment to collect information on receptive language

*Correspondence:

Sabrina Horvath
horvaths@musc.edu

¹ Medical University of South Carolina, Charleston, USA

² New York University, New York, USA



that a clinician might be otherwise unable to acquire (such as through a standardized assessment with high behavioral demands). Capturing receptive abilities is critical in clinical decision-making. A receptive language eye-tracking assessment could provide insight into whether a late talker or minimally speaking autistic child has delays in their receptive abilities in addition to those that have been observed expressively. This, in turn, would guide goal setting and intervention approaches.

At present, eye-tracking technology has not yet been adapted by clinicians, due largely to the prohibitive cost and availability of technology but also perhaps due to the lack of available assessments and guidelines for interpretation. However, with rapid advances in the fields of technology and artificial intelligence, it is conceivable that eye-tracking assessments will one day make their way into general use. Currently, using eye-tracking to assess vocabulary is regularly used in research settings to address a wide variety of questions about language development. Landmark studies have demonstrated that lexical processing speed, measured by eye-tracking, predicts both concurrent and later language abilities in typically developing children [8–10] and late talkers [11]. Other research has demonstrated that eye-tracking better captures vocabulary knowledge than standardized assessments in autistic children [12, 13]. Lany [14] recently demonstrated that language processing speed as measured through eye-tracking predicts concurrent noun-learning abilities in toddlers. Taken together, these studies demonstrate the value of using eye-tracking to assess receptive vocabulary and its potential as a tool for better understanding how language development unfolds.

In this paper, we focus specifically on how current eye-tracking paradigms and measures may be used to assess receptive vocabulary in populations with diverse language abilities. Our results offer advances toward the development of eye-tracking vocabulary assessment but also highlight remaining gaps in our knowledge. As stated, assessing receptive vocabulary using eye-tracking is at present a clinical pipe dream, but by addressing the gaps we have identified we may one day make such assessments feasible. Moreover, such foundational work can support further research efforts to better understand the ways in which language development does and does not differ between typically developing children and those with diverse language abilities. While it is not possible to address all research questions in this study—leaving for future studies such topics as target selection, task duration, and adaptability of the paradigm based on child performance—we aim in this study to take an initial step of translating current methods into new populations and different kinds of vocabulary words.

Dozens of studies have demonstrated the feasibility of using eye-tracking to assess children's receptive vocabulary (e.g., [4, 8, 15–17]). Broadly, these studies follow the same experimental paradigm. Each trial begins with a “Baseline Phase” where children preview multiple potential candidate scenes (e.g., pictures of a ball and a shoe). In the “Prompt Phase,” children are directed to look to one of the scenes (e.g., “Where's the ball?”), sometimes with the pictures displayed on the scene (as a “Looking While Listening Paradigm,” e.g., [8]) and sometimes without (as an “Intermodal Preferential Looking Paradigm,” e.g., [16]). The time after this linguistic prompt is the “Test Phase” of the trial, from which children's eye-gaze behavior is analyzed to determine knowledge.

Most of these studies have focused only on one type of vocabulary word: nouns labeling objects or animals. However, a comprehensive assessment should include many types of words. We argue that verbs are critically important to consider. Verb meanings are strongly tied to the structure of the sentences in which they appear, and verb knowledge is a better predictor of later language outcomes than noun knowledge [18, 19]. Moreover, limited verb knowledge is considered a warning sign for a later diagnosis of language disorder [20], and it has been hypothesized that understanding late talkers' verb vocabularies, in particular, may help researchers and clinicians ultimately distinguish between those at greatest risk for language disorder and those likely to “catch up” to typically developing peers [21]. Therefore, any assessment of receptive vocabulary must include a large number of verbs.

However, in the context of eye-tracking assessments, testing verb knowledge raises unique challenges. Prior studies using eye gaze to assess noun knowledge have primarily used static images such as a picture of a ball. Because most early-acquired verbs denote dynamic events that unfold over time, static images are not good depictions of them; consider, for example, the difficulty of distinguishing “catching” and “throwing” with a single static image. Although some standardized assessments such as the Peabody Picture Vocabulary Test [22] depict verb referents using static images, these are not ideal for young children, who have difficulty interpreting the symbols used to denote dynamic action (e.g., lines to indicate motion) and extrapolating movement from them [23, 24].

Instead, videos of dynamic scenes better illustrate verb meanings. However, using dynamic scenes requires rethinking how eye gaze measures are operationalized because gaze behaviors differ when viewing dynamic scenes as compared to static images (e.g., [15, 25–27]). Relatively few studies have used dynamic scenes to depict verbs, and among those that have, there is not a

consensus on what eye gaze measures are best nor how they should be operationalized [6, 15–17, 28–30].

What research has been done using dynamic scenes to depict verb referents in vocabulary assessments has, until now, included only typically developing children. However, in developing assessments for children with language and communication disorders, we must consider the possibility that their performance differs substantially from that of their typically developing peers. Although findings are mixed [11, 13, 30–40], several studies suggest that late talkers, children with developmental language disorder, and autistic children are all slower language processors typically developing children [11, 13, 30, 31, 33, 38], meaning that they may take longer to settle their gaze on the scene depicting the correct referent.

It is critical that researchers understand how to adapt receptive vocabulary tasks for children with differing abilities. For example, Brady et al. [3] have argued that because pointing is challenging for autistic children, using eye gaze will more appropriately capture their vocabulary knowledge if we can reliably interpret their gaze behaviors. Several studies have explored the possibility of assessing autistic children's vocabulary using eye gaze (e.g., [3, 4, 41–46]), but they have focused almost exclusively on noun vocabulary and static image targets.

However, several studies have also looked at sentence-level processing in young children with language delays, and critically these studies have hinged on verb processing. For example, some recent studies have examined whether young 3- to 5-year-old autistic children [12, 13, 39] and 2-year-old late talkers [36] engaged in incremental sentence processing, with the verb serving as a cue for the objective noun. Taken together, these studies suggest that young children with language delays can process familiar verbs to anticipate their objects, but that they are generally slower to do so than their typically developing peers, depending on their ages and receptive vocabulary knowledge. These findings suggest that additional work with young autistic children and late talkers, which we undertake in the current study, should examine how they process the lexical item in itself, without the added task of using it to predict an upcoming noun.

In this study, we address these gaps in the literature. Our goal is to advance understanding of the potential for eye-gaze vocabulary assessments. First, we focus specifically on verbs, depicting verb referents as dynamic scenes rather than static images. Second, we include children with language delays and disorders. Specifically, in experiment 1, we study late talkers' receptive verb knowledge, and in experiment 2 we study autistic children's receptive verb knowledge. Prior research indicates that both populations differ in their eye-gaze behaviors during receptive vocabulary tasks as compared to typically developing

children, at least given noun trials and static images [3, 4, 11, 41, 46].

Experiment 1 compares late talkers' and typically developing children's performance on a receptive verb vocabulary task. Approximately 15% of 2-year-olds are late talkers, defined by having atypically small expressive vocabularies for their age with no known cognitive or developmental disorders [47]. Late talkers are at increased risk for developmental language disorder [48–50]. Many late talkers have receptive language delays as well, but others do not [51, 52]. Importantly, there appear to be differences in how late talkers build their vocabularies [53–56], including specifically their verb vocabularies [57, 58].

In experiment 2, we consider autistic children's performance. Autistic children have notoriously heterogeneous language abilities [35, 59–61], ranging from non-speaking to having no expressive or receptive language deficits. However, most autistic children have below-age language abilities, including in vocabulary knowledge (e.g., [61, 62]). Receptive language abilities vary widely, but for some, receptive language is even more impaired than expressive language, at least in early language development and when measured through standardized assessments (e.g., [59, 63]). Further, autistic children have particular difficulty participating in standardized assessments which require overt behavioral responses [3–5] making alternative assessment methodologies an important clinical need.

Operationalizing eye gaze behaviors

Given the richness of the dataset collected during eye-tracking, interpreting gaze behavior to measure vocabulary knowledge is a daunting task. Prior studies have converged on two distinct measures: accuracy and processing speed. Because this is a methods study, we will discuss each measure below in detail, including how it has been operationalized in prior research and how it may need to be modified for (1) verb trials with dynamic scenes and (2) children with language delays and disorders. In so doing, we recognize that “children with language delays and disorders” is not a monolith, and that there is considerable variability across and within disorder profiles. To this point, methods must be adapted to specific populations and even, potentially, to individuals.

Accuracy

Accuracy—or, whether a child knows the target word—is operationalized as a proportion of looking time to the target scene. There are two common ways of calculating accuracy. One approach [8, 11] is to compare the proportion of time spent looking to the target versus distractor scenes during the Test Phase. If the child prefers

the target to the distractor, they are credited with knowing the target word. An alternative approach [7, 9, 64] is to compare the proportion of time spent looking to the target scene before versus after the linguistic prompt is provided (i.e., Baseline versus Test). If this proportion looking increases by a predetermined threshold, typically 15%, then the child is credited with knowing the target word [7]. We use this second method because it accounts for idiosyncratic preferences that children may have for one scene over the other. This second approach has been used with static images and noun targets [9, 64] as well as dynamic scenes and verb targets [9, 17]. Further, accuracy scores derived from this approach correlate with concurrent vocabulary knowledge as measured by the Peabody Picture Vocabulary Test or parent report [9, 17].

Regardless of the approach, researchers must identify a time window within the Test Phase from which to calculate accuracy. Children's gaze is predictably drawn to the target immediately after the linguistic prompt, but afterward, they look around to other locations on the screen. Fernald and colleagues [8, 10, 11, 65], working with typically developing toddlers who view noun trials with static images, use a time window for accuracy calculations of 300 to 1800 ms after the linguistic prompt. Children require 300 ms to coordinate and launch an eye movement toward the target, and their attention patterns are less consistent after 1.5 s of gaze. But, what about for verb trials with dynamic scene stimuli, or children who are not typically developing?

Valleau and colleagues [17] have observed that this 300 to 1800 ms window may be inappropriate for toddlers when they are shown verb trials with dynamic scenes. In a study including typically developing 22- to 24-month-old typically developing toddlers, Valleau et al. found that, on noun trials, toddlers showed evidence of preferring the target scene within 300 to 1800 ms after the target noun was queried, consistent with Fernald and colleagues' research. However, on verb trials, the same toddlers required additional time to orient their gaze. It remains an open question in research as to what window should be used to calculate accuracy given verb targets and dynamic scenes.

Furthermore, the appropriate time window for calculating accuracy may differ for typically developing children and those with language delays and disorders. For example, late talkers are slower lexical processors than typically developing toddlers and may therefore take longer to settle on the correct scene [11]. Autistic children are also slower language processors than typically developing peers [33] and demonstrate highly variable patterns of language ability [61]. We therefore cannot rely on previously established windows to calculate accuracy for these populations.

Given that there is no a priori hypothesis for when or how long the test window should be in accuracy calculations, we identify the appropriate windows using a bootstrapped cluster-based permutation analysis [66]. Bootstrapped cluster-based permutation analysis is a systematic method for identifying when, and for how long, behavior differs between groups or conditions in studies that have time-locked data, such as eye-tracking or EEG studies. This approach has successfully been used in a range of eye-tracking studies, including with autistic individuals (e.g., [34, 39]). In a bootstrapped cluster-based permutation analysis, small time windows are analyzed to determine whether there are differences in gaze behavior between conditions, and consecutive windows that show a difference are grouped together in clusters. Permutation testing is applied to derive a distribution of cluster values in order to determine whether the clusters are statistically significant. However, there are challenges in applying this approach to individual children's performance, as we discuss below.

Processing speed

To evaluate processing speed, researchers typically measure children's latency to look to the target image—that is, how long after the auditory prompt it takes children to first look at the target, irrespective of whether or when they shift gaze away afterward. Latency has been widely used in eye-tracking research addressing a multitude of questions about language development; we focus here on research specifically targeting familiar vocabulary items. Robust evidence from studies involving static images to depict familiar noun targets has shown that this measure predicts concurrent language ability [8], concurrent noun-learning abilities [14], and later language and intelligence [10, 11, 67]. Whether this measure can also be used with dynamic scene stimuli is less clear: If children happen to look at the distractor first, they may take longer than with static images to visually disengage from it. Indeed, Valleau et al. [17] found no association between latency and vocabulary size in typically developing children under 2 years of age for verbs and dynamic scenes. This may mean that because of differences in how dynamic scenes capture the attention compared to static images, their speed to first look to the target is not a useful measure. However, age may also be a factor: Koenig et al. [28] found that, for typically developing 3-year-olds, who may be better equipped to attend to task demands, latency correlated with vocabulary size on both noun and verb trials.

In considering children with language delays and disorders, latency has been shown to be a predictive measure given noun trials with static images. For example, Fernald and Marchman [11] demonstrated that late talkers have

longer latencies than typically developing toddlers, and that late talkers' latencies predicted vocabulary growth between 18 and 24 months. Autistic children, too, have slower language processing skills that correspond with other language measures (e.g., [31, 33, 68, 69]), including their accuracy on eye-tracking receptive vocabulary tasks [4]. However, no studies have considered latency for children with language delays and disorders viewing dynamic scenes and familiar verb stimuli in a vocabulary assessment.

The present study

To begin to address these foundational gaps in our understanding of how eye gaze can be used to examine receptive verb vocabulary in children with language delays and disorders, we present two experiments that adapt tasks used by Valteau et al. [17] and Koenig et al. [28]. Our goal is to explore how accuracy and processing speed measures may be adapted to accurately capture verb knowledge in children with language delays and disorders.

Experiment 1: late talkers and typically developing children

Participants

The final sample included 45 children (17 female, 28 male) with an average age of 28.5 months ($SD=3.0$ months, range=24.5–34.7 months) recruited from the greater Boston area. The sample skewed male because we focused recruitment on late talkers, who are more likely to be male [49]. Participants were pre-screened for a history of hearing loss or tubes; additionally, children were screened at their visit for a high risk of autism spectrum disorder using the Modified Checklist for Autism in Toddlers, Revised (M-CHAT-R: [70]). All participants in the final sample were classified as “low risk.” Per parent report, all children were exposed to English at least 70% of the time. Participants had no reported developmental disorders other than for language: Three children were reported to have a language-related diagnosis, either “expressive language delay” or “language delay.” Seven additional children participated but were excluded from final analysis due to a prolonged history of ear tubes ($n=2$), a history of tongue-tie ($n=1$), diagnosis with autism spectrum disorder within a week following participation ($n=1$), or failure to complete the experimental session due to fussiness ($n=3$).

We asked parents to report on race/ethnicity and parent education. One family did not provide information on race/ethnicity, and three families did not provide information on parent education. Of those who did, participants were primarily White (91%), 2% were Asian, and 5% were more than one race. Nearly all children (96%) had at least one parent with a college degree or more

advanced degree, including 35% of children who had at least one parent with a doctorate (Ph.D., M.D., or J.D.).

Expressive vocabulary was assessed using the MacArthur-Bates Communicative Development Inventories Level 2 Short Form A (MBCDI: [71]). Children were reported to produce, on average, 69 of the 100 words ($SD=27$, range=1–100). The Preschool Language Scales, 5th edition (PLS: [72]) was administered to characterize broader language abilities. Children averaged a standard score of 106 ($SD=17$) on the Auditory Comprehension subscale (PLS-AC) and 107 ($SD=16$) on the Expressive Communication subscale (PLS-EC). Finally, the Visual Reception subscale of the Mullen Scales of Early Learning (MSEL-VR: [73]) was used as a proxy for nonverbal intelligence. Children had an average *T*-score of 53 ($SD=11$); one participant did not complete MSEL-VR testing.

We used the MBCDI to classify each child as either a “late talker” ($n=14$) or “typically developing” ($n=31$). Of the late talkers, eight had a standard score at or below the 15th percentile for their age and gender. The MBCDI is only normed for children up to age 30 months; however, children older than 30 months whose score was at or below the 15th percentile for their gender at 30 months were also classified as late talkers ($n=4$). An additional 2 late talkers were classified based on parent report that they had qualified for speech and language therapy because of late talking. In total, 10 late talkers had qualified for or were receiving speech therapy services at the time of participation; no typically developing children had any reported history of therapy. There were no group differences with respect to average age ($t=1.14$, $p=0.26$, *n.s.*), proportion male ($z=0.85$, $p=0.39$, *n.s.*), or proportion monolingual English language learners ($z=1.5$, $p=0.12$, *n.s.*). All standardized measures showed group differences (see Table 1).

Apparatus

Stimuli were displayed on a 24-inch Tobii T60 XL corneal reflection eye-tracking monitor, which samples gaze approximately every 17 ms, calibrated at the beginning of each experimental session using a 5-point calibration procedure. Children sat in a car seat 20 in from the monitor or in their parent's lap while the parent wore a blindfold.

Stimuli

The stimuli were initially developed by Konishi et al. [74] and modified for an eye-tracking procedure by Valteau et al. [17]. Konishi et al. selected a total of 36 verbs and 14 nouns that are highly imageable and learned early in typical language development. They filmed 6-s video clips depicting the referent action for each verb and selected static images depicting the referent object for

Table 1 Late talkers’ and typically developing children’s performance on standardized assessments (experiment 1)

	Macarthur-Bates communicative development inventories (short form A): raw score	Preschool language scales, auditory comprehension subscale: standard score	Preschool language scales, expressive communication subscale: standard score	Mullen scales of early learning, visual reception subscale: T-scores
Late talkers	M=40 SD=25	M=88 SD=15	M=90 SD=11	M=44 SD=10
Typically developing children	M=82 SD=16	M=114 SD=12	M=114 SD=12	M=58 SD=8
Test statistic	t=6.73, p<0.001	t=6.31 p<0.001	t=6.17 p<0.001	t=4.77 p<0.001








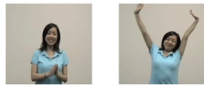

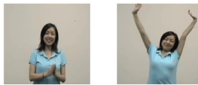
		Preview Phase	Baseline Phase	Prompt Phase	Test Phase	
Noun Trials	Visual Stimuli					
	Auditory Stimuli	<i>Look!</i>	<i>Wow!</i>	<i>Do you see?</i>	<i>Where is the cookie?</i>	(beginning at 13 s) <i>Find the cookie!</i>
	Time from onset (s)	0 – 2 s	2 – 4 s	4 – 7 s	7 – 11 s	11 – 17 s
Verb Trials	Visual Stimuli					
	Auditory Stimuli	<i>Look!</i>	<i>Wow!</i>	<i>Do you see?</i>	<i>Where is she clapping?</i>	(beginning at 20 s) <i>Find clapping!</i>
	Time from onset (s)	0 – 4 s	4 – 8 s	8 – 14 s	14 – 18 s	18 – 24 s

Fig. 1 The trial structure of one trial for experiments 1 and 2

each noun. Valteau et al. recorded accompanying auditory stimuli and arranged the stimuli into the trial structure depicted in Fig. 1. Experiment 1 included a subset of ten of the verb trials from the stimuli used by Valteau et al., described below, including only one item from each pair (e.g., “clap” but not “stretch”), as well as four of the noun trials which served as fillers to break up the session. All participants saw the same 14 trials in the same order. Children saw each trial once. See Additional file 1: Appendix A for a list of trials.

Visual stimuli

Verb trials featured two dynamic scenes side-by-side. Two verb trials featured dynamic scenes with just an

actor (e.g., “clapping” and “stretching”) and eight trials featured dynamic scenes with an actor and an object (e.g., “shaking” and “opening” a present). Within each trial, the actor and object were the same in both dynamic scenes (e.g., in the trial depicting “tickle” and “kiss,” one scene depicted a girl tickling a teddy bear, while the other depicted the same girl kissing the same teddy bear). Videos were looped to provide continuous depictions of the events; Some events were durative and therefore continuous (e.g., “run”), whereas others occurred punctually between two (e.g., “kick”) and five (e.g., “break”) times. Filler trials targeting nouns featured two static images side-by-side.

Auditory stimuli

A female American English speaker recorded the auditory stimuli in a sound-attenuated booth. Children heard attention-grabbing phrases (e.g., “Wow!”) and directives to find the target. For trials including both an actor and object, verbs were targeted using transitive syntax (e.g., “Where is she tickling the bear?”), whereas those including only an actor were targeted using intransitive syntax (e.g., “Where is she clapping?”). Children also heard prompts in neutral syntax (e.g., “Find clapping!”).

Design

Each trial included an Inspection Phase, a Baseline Phase, a Prompt Phase, and a Test Phase (see Fig. 1). Verb trials and noun filler trials were structured identically; however, the Inspection and Baseline Phases were shorter for noun trials than for verb trials because static images do not change over time and we did not want children to tire of looking at them.

In the Inspection Phase (8 s for verb trials; 4 s for noun trials), children previewed each visual stimulus individually, one on the left and the other on the right. Side (left or right first) and order (target or distractor first) were counterbalanced. The Baseline Phase (6 s for verb trials; 3 s for noun trials) depicted both visual stimuli simultaneously in the same locations they had appeared in during the Inspection Phase. The Inspection and Baseline Phases included attention-grabbing phrases to direct children’s attention to the screen (e.g., “Look!”, “Wow!”).

In the Prompt Phase (4 s for verb and noun trials), children heard a prompt to find the target scene or image. Scenes featuring only an actor were queried in intransitive syntax (e.g., “Where is she clapping?”), whereas scenes featuring an actor and agent were queried in transitive syntax (e.g., “Where is she throwing the balloon?”). Because pairs of scenes always featured the same actor and objects, nouns and pronouns were not a cue for the target versus distractor. A centrally positioned star directed children’s attention to the center of the screen. In the Test Phase (6 s for verb and noun trials), the visual stimuli reappeared in their original positions. Children heard an additional prompt in neutral syntax (e.g., “Find clapping!”) after two seconds.

Procedure overview

Participation was part of a two-visit protocol approved by Boston University’s Institutional Review Board. At the first visit, parents provided written consent and completed a demographics questionnaire, the MBCDI and M-CHAT-R. The first author, a licensed speech-language pathologist, administered the PLS. Children also participated in an unrelated experimental task. At the second visit, approximately 2 weeks later, children took part in

two additional experimental tasks, of which this study was the second. The MSEL-VR was also administered during the second visit.

Exclusionary criteria

All trials with more than 50% track loss (e.g., blinks) during the Test Phase were removed from analysis. After these removals, on average, 9 of 10 verb trials ($SD=1$, range=5–10) were included for typically developing children, while 7.5 of 10 ($SD=2$, range=4–10) were included for late talkers; this difference was significant ($t(43)=3.35$, $p=0.002$). Differences in the number of included trials is unsurprising given that late talkers show differences in attention during experimental tasks [75]. However, some of this inattentiveness may also be driven by task difficulty; for example, late talkers may look toward a parent or examiner for cues because they are unsure of the target word’s meaning.

Analysis

Our analyses considered children’s (1) accuracy and (2) processing speed. For each, we conducted a mixed-effects regression to determine whether there were group differences. This included the outcome variable of eye gaze behavior (accuracy or processing); random effects of participant and trial; and fixed effects of age, gender, and group (late talker or typically developing). Regressions were run using the lme4 package (Version 1.1–12; [76]) in R [77] with model comparisons made using the drop1() function with chi-square tests.

Accuracy

Following Reznick [7], we calculated accuracy as an increase of 15% in target looking between Baseline (before children are prompted to find the target) and Test (after the auditory prompt). To identify at what point in time during the 6-s test window we should make this calculation, we applied a bootstrapped cluster-based permutation analysis [66] using the eyetrackingR Package [78]. We hypothesized that late talkers might require a later time window for demonstrating vocabulary knowledge than typically developing toddlers, so we ran separate analyses for each group. The cluster analysis compared children’s gaze behaviors between Baseline and Test to identify if and when children preferred the target in the Test Phase above and beyond Baseline looking rates. For the Baseline Phase, we averaged proportion of looks to the target scene versus elsewhere across all time points and trials to obtain a single measure of each group’s overall preference for the target scene during this Phase. This is because we were not interested in the dynamics of their attention to the target scene during Baseline, but rather how much they preferred to look at

it overall. For the Test Phase, in which we were interested in the dynamics of children's attention over time, we calculated children's average proportion of looks to the target scene versus elsewhere in each 50-ms window. In both cases, when calculating the proportion of looking to the target, we included looks to neither the target nor the distractor (e.g., looking in between the two scenes) and track loss in the denominator of the proportion; these looks may reflect children's uncertainty and we did not want to remove these data points.

Our planned model for identifying clusters was a mixed-effects regression with the dependent variable of proportion of looks to the target scene versus elsewhere, the predictor variable of phase (Baseline or Test, dummy coded as "0" and "1"), and random effects of trial and participant. We applied a threshold of $p=0.05$, meaning the time bin had to reach this level of statistical significance in order to be included in a cluster. Adjacent clusters and those separated by only 50 ms were combined into larger clusters. We then ran the permutation analysis with 1000 permutations to confirm that these windows emerge even when the data is scrambled. Two paired t tests (following, e.g., [79]), one including all the children in the LT group and another including all children in the TD group, compared, for each trial for each child, the average proportion of looks to the target scene between each trial's overall Baseline looking and the identified cluster.

The earliest statistically significant cluster was used to identify the response window for the accuracy analysis. Response windows—separate for each group—began at the start of the earliest significant cluster wherein children looked more to the target scene in the Test Phase than in Baseline. We standardized the duration of the response windows to 1500 ms, as has been done in receptive noun vocabulary tasks (e.g., [8]).

Accuracy was then calculated, by-child by-trial, by comparing the average proportion of looks during the whole 6 s of the Baseline Phase and the response window of the Test Phase. A child was credited with knowing the meaning of the target verb if their looks increased at least 15% from Baseline to Test.

Processing speed

Processing speed was operationalized as latency, i.e., the earliest time point within the Test Phase of each trial in which the child looked toward the target scene. As in Valteau et al. [17], children who did not look to the target scene during the Test Phase at all were given a latency of 6000 ms. Also following Valteau et al. [17], we excluded looks in the first 50 ms of the Test Phase as being too early to be attributable to hearing the auditory stimuli; it takes approximately 300 ms to program and launch a saccade (e.g., [80]).

Results

De-identified gaze data are available on the Open Science Framework (<https://osf.io/ghp7q>). Figure 2 depicts children's preference for the target scene over time as the Test Phase unfolded; target preference is calculated as the proportion of frames in which children looked to the target scene versus all other locations. Baseline looking preference is indicated by the dashed lines. Late talkers averaged a smaller proportion of looking to the target scene than typically developing children during the Test Phase ($M(LTs)=0.39$, $SD(LTs)=0.08$; $M(TDs)=0.49$, $SD(TDs)=0.09$; $t(43)=3.8$, $p<0.001$). Conversely, late talkers averaged a higher track loss than typically developing children on included trials ($M(LTs)=0.28$, $SD(LTs)=0.09$; $M(TDs)=0.22$, $SD(TDs)=0.09$; $t(43)=2.3$, $p=0.024$). However, there were no between-group differences in average proportion of looks to the distractor scene ($M(LTs)=0.33$, $SD(LTs)=0.06$; $M(TDs)=0.29$, $SD(TDs)=0.07$; $t(43)=3.8$, $p=0.16$, *n.s.*). We observed from visual inspection of the graph that both groups preferred the target during the Test Phase above Baseline looking rates. This suggests that, overall, children know at least some of the target verbs queried.

Accuracy

For late talkers, the bootstrapped cluster-based permutation analysis revealed three clusters in which proportion of looking to the target scene differed between Baseline looking rates ($p=0.40$) and the Test Phase. The first cluster lasted from 0 to 600 ms: Here, late talkers looked less to the target scene in Test than they had in Baseline ($t(104)=15$, $p<0.001$). This is unsurprising given the trial structure: Recall that children begin the Test Phase looking at the center of the screen, as they have just seen a central fixation star. The second cluster began at 1550 ms and lasted to 3100 ms ($t(104)=-2.5$, $p=0.01$); here, late talkers looked more to the target scene in test than in Baseline. The third cluster, from 4850 to 6000 ms, was not statistically significant ($t(104)=-1.3$, $p=-0.19$, *n.s.*) after the permutation analysis and t test. Given the results of this analysis, late talkers were given the response window of 1550 to 3050 ms for the accuracy analysis (because we standardized windows to a duration of 1500 ms).

For typically developing children, two significant clusters emerged in which the proportion of looking to the target differed between Baseline ($p=0.43$) and Test. The first cluster lasted from 0 to 600 ms. As with late talkers, typically developing children began the Test Phase looking less to the target scene than they had in Baseline ($t(276)=19$, $p<0.001$). The second cluster lasted from 900 to 6000 ms. Here, typically developing children looked at the target scene significantly more during test

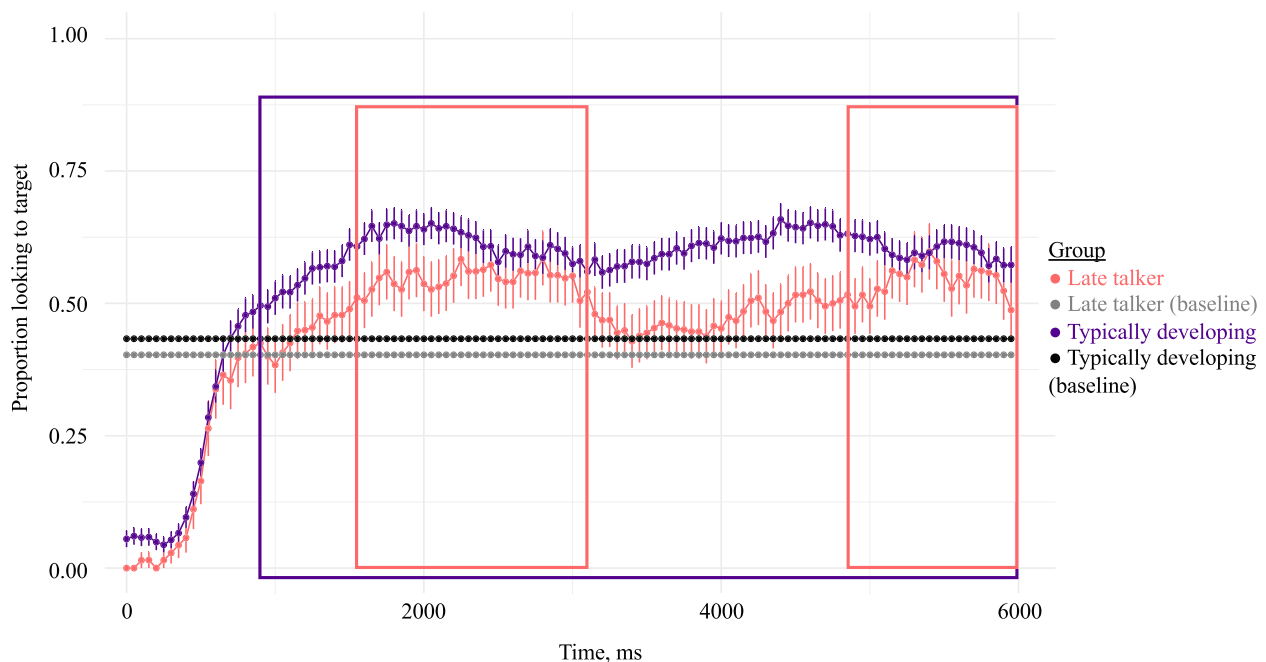


Fig. 2 Timecourse of children's gaze to the target scene during the test phase by group (experiment 1). The x-axis represents time, in ms, from the onset of the test phase, and the y-axis represents the proportion of looks to the target scene versus elsewhere. Error bars indicate standard error of participant means. Dashed lines indicate group baseline averages. The boxes indicate times in which proportion of looking to the target was significantly greater in the Test Phase over the Baseline Phase for LTs (red) and TDs (purple), per cluster-based permutation analysis

than they had during baseline ($t(276) = -5.7, p < 0.001$). We therefore used a response window for the accuracy analysis of 900 to 2400 ms for typically developing children to standardize the duration, but we note that it is interesting that the original cluster for typically developing children was much longer (5100 ms) than for late talkers (1600 ms), suggesting that once typically developing children had settled on the target they sustained their attention on it for much longer.

Using the threshold of 15% increase between Baseline and response window, late talkers knew 51% of the target verbs ($SD = 0.22$, range = 0.125–1) for trials they contributed. Typically developing children knew 49% of the target verbs ($SD = 0.19$; range = 0.0–0.9) for the trials they contributed. The regression model indicated no significant relationship between children's accuracy and any of the fixed effects included ($b_{\text{group}} = 0.02, t_{\text{group}} = 0.34, p_{\text{group}} = 0.70, n.s.$; $b_{\text{age}} = 0.02, t_{\text{age}} = 1.88, p_{\text{age}} = 0.054, n.s.$; $b_{\text{gender}} = -0.09, t_{\text{gender}} = -1.51, p_{\text{gender}} = 0.11, n.s.$). This indicates that, when provided enough time to demonstrate knowledge of target items, there are no significant differences in the number of verbs late talkers and typically developing children know.

At the recommendation of one reviewer, we conducted a post-hoc analysis wherein late talkers were given the same response window as typically developing children (900 to 2400 ms). Given this window, late talkers knew

only 35% of the target verbs ($SD = 0.20$; range = 0.0–0.75). Here, the analysis yielded between-group differences, wherein late talkers knew significantly fewer verbs than typically developing children ($(b_{\text{group}} = 0.15, t_{\text{group}} = 2.3, p_{\text{group}} = 0.016)$); no other factors were significant ($b_{\text{age}} = 0.02, t_{\text{age}} = 1.86, p_{\text{age}} = 0.063, n.s.$; $b_{\text{gender}} = -0.09, t_{\text{gender}} = -1.79, p_{\text{gender}} = 0.079, n.s.$). These findings indicate that late talkers perform poorer than typically developing children when assessment measures do not account for differences in overall response time.

We also included exploratory analyses in which language was treated as a continuous variable. However, we found no significant effect of the language variables (MBCDI raw score $b = -0.001, t = -0.19, p = 0.84, n.s.$; PLS-AC standard score $b = 0.002, t = 0.916, p = 0.31, n.s.$).

As an illustration of which verbs children in both groups tended to know, Table 2 shows the rank order of most to least accurate trials, by group. We note considerable variability between groups; late talkers had the highest proportion accuracy on the trial targeting “jump” (with the distractor “run”), whereas typically developing children had the highest proportion of accuracy on trials targeting “lick” (with the distractor “break”). Interestingly, “lick” was the second most-difficult verb for late talkers. Although intriguing, we note here that our goal was to explore methodological preliminaries surrounding our abilities to use eye-tracking to collect such information.

Table 2 The proportion of each verb known by late talkers and typically developing children, ranked (experiment 1). Numbers in parentheses indicate the proportion of participants who knew the target verb, by group

Rank	LTs	TDs
1	Jump (0.72)	Lick (0.60)
2	Squeeze (0.67)	Wash (0.58)
3	Tickle (0.63)	Squeeze (0.53)
4	Open (0.62)	Tie (0.53)
5	Roll (0.55)	Open (0.52)
6	Clap (0.38)	Roll (0.48)
7	Wash (0.38)	Clap (0.46)
8	Throw (0.33)	Tickle (0.44)
9	Lick (0.29)	Jump (0.38)
10	Tie (0.2)	Throw (0.36)

With this in mind, we propose further research using this methodology but specifically designed to explore whether, as with expressive verb vocabulary [57, 58], late talkers and typically developing children differ in their receptive verb vocabulary compositions.

Processing speed

Participants' latency to look to the target scene averaged 1500 ms ($SD=502$ ms). Surprisingly, late talkers ($M=1551$ ms, $SD=477$ ms) did not average longer latencies than typically developing children ($M=1477$ ms, $SD=519$ ms; $t(43)=0.66$, $p=0.45$, *n.s.*). The regression analysis indicated that age significantly predicted latency ($b=-72$, $t=-4.9$, $p=0.004$), but group ($b=-119$, $t=-0.75$, $p=0.44$, *n.s.*) and gender ($b=36$, $t=0.25$, $p=0.79$, *n.s.*) did not.

This finding is perhaps striking given that there were between group differences in the start of the response window for accuracy analysis. We note that, although both relate to processing in some way, they are distinct measures. Latency is children's first look to the target; it is calculated by-trial and independently of children's performance in the Baseline Phase. By contrast, the response window represents patterns of sustained looking across all trials, and it is calculated relative to Baseline looking rates. In so doing, we are capturing how quickly children demonstrate a sufficiently robust representation of the target verb above and beyond chance looking rates. For example, a child may look first toward the target as their initial guess, but their representation may not be sufficient to feel confident in this choice; they would therefore scan back and forth between the two scenes before settling back onto the target with certainty (see [81] for a similar pattern in autistic children's sentence processing). What we therefore interpret from these two measures

together is that LTs and TDs were equally quick to first look to the target scene (latency), but that LTs took longer than TDs to settle on the target for a sustained period of time (response window), perhaps indicating less robust lexical entries.

We again included exploratory analyses in which language was treated as a continuous variable. We observed that broader receptive language abilities, but not vocabulary, predicted performance, such that children with higher standard scores on the Auditory Comprehension subtest of the PLS-5 averaged faster latencies; here, too, we found no significant effect of the language variables (MBCDI raw score $b=1.17$, $t=0.245$, $p=0.79$, *n.s.*; PLS-AC standard score $b=-15.04$, $t=-1.88$, $p=0.049$).

Discussion

In experiment 1, we explored children's eye gaze during a receptive verb vocabulary task with 2-year-old late talkers and typically developing children. We considered children's overall accuracy and processing speed.

In calculating children's accuracy, prior research has suggested that the response window that has been typically used with static images and noun stimuli (300 to 1800 ms) is inappropriate for dynamic scene targets [15, 17]. Instead, we identified a response window using bootstrapped cluster-based permutation analyses [66]. Given that late talkers are slower lexical processors than typically developing toddlers [11], it is perhaps unsurprising that they required a later window than their typically developing peers to demonstrate verb knowledge. While typically developing children preferred the target scene above Baseline looking rates beginning at 900 ms in the Test Phase, late talkers did not do so until 1550 ms. These findings echo research on older children with developmental language disorder, who show delayed responses during receptive language tasks (e.g., [82]). However, when provided additional time, late talkers knew as many verbs as did typically developing children. By contrast, when late talkers were held to the same expectations as typically developing toddlers (i.e., the 900 to 2400 ms window), there was a significant group difference. This discrepancy highlights the importance of adapting assessment measures to the population being studied, and accounting for differences between toddlers who are typically developing and those with language delay or disorder.

While not what we had hypothesized, the finding that late talkers and typically developing children showed receptive knowledge of the same proportion of the tested verbs is not altogether unsurprising. Late talkers are defined by the size of their expressive vocabularies; prior research indicates that although some late talkers also have smaller receptive vocabularies, others do not

show receptive language deficits [51]. We also acknowledge that although late talkers and typically developing children knew on average the same number of verbs, it is not necessarily the case that they have equally robust representations of those verbs. Indeed, we observed in our bootstrapped cluster-based permutation analysis that, unlike typically developing toddlers, late talkers did not sustain a preference for the target scene for as long a duration once they identified it. One possibility is that this reflects late talkers' confidence in their responses. In support of this hypothesis, we note that late talkers lost significantly more trials due to track loss than typically developing children, indicating more looks away from the screen (and possibly to a parent or researcher for cues or confirmation). Alternatively, the difference in sustained attention may be an indication that late talkers' representations are more fragile than typically developing toddlers' representations. It remains an open question in the field of how best to operationalize robustness of a lexical entry. One possibility is that overall looking time to the target indicates robustness of the lexical entry [83], but it is also possible that children with robust entries look quickly and then scan as they become bored with the task [84, 85]. We advocate for continued research into how best to operationalize robustness of representation, and whether this may vary as a factor of age, language ability, or population.

It is also worth noting that, although overall rates of accuracy did not differ between LTs and TDs, there were group differences in which verbs children were most likely to identify correctly (Table 2). This is consistent with prior research demonstrating that LTs and TDs show differences in the composition of their vocabularies (e.g., [54–58]). While most of these studies have focused on children's expressive vocabularies, we offer evidence for possible differences in receptive vocabularies as well. Given the small number of trials in the current experiment and that they were not balanced across different types of verbs to be able to make systematic comparisons, we do not provide an interpretation of the differences in the lists in Table 2, but we leave this topic for future work, which should consider the intersection of verb knowledge, verb learning, and subsequent grammatical development. We tentatively hypothesize that a nuanced understanding of late talkers' emerging verb vocabularies—both in the number and type of verbs acquired—may support endeavors to identify which late talkers are at greatest risk for developmental language disorder (similar to 19, which included outcomes for autistic toddlers).

Latency is a well-established eye gaze measure for processing speed given static images and noun targets, but research with dynamic scene stimuli has drawn mixed conclusions [16, 17, 28]. Although late talkers average

slower latencies than typically developing children given noun targets and static images [11], we found no group differences in average latency to verb targets and dynamic scenes. Instead, children's age significantly predicted performance, with older children faster to orient to the target than younger children. These results may provide insight into the discrepancies of prior findings. Golinkoff et al. [16] and Valteau et al. [17], who found no relationship between language ability and latency, both studied children who were younger than 2 years of age. However, Koenig et al. [28] did find that language predicted latency in 3-year-olds. We hypothesize that children are refining their processing abilities during the third year of life, improving their incremental language processing skills as well as their ability to focus on task demands over and above the ways in which dynamic scenes draw their attention. This in turn results in processing speed better reflecting other aspects of language knowledge. We would therefore expect that among older children, latencies reflect the difficulty of identifying the word's referent, which should relate to their performance on other language assessments.

Experiment 2: autistic preschoolers

Participants

The final sample comprised 20 children (3 female, 17 male) on the autism spectrum who were recruited from the greater Boston area. Children averaged 41.9 months old at the time of participation ($SD=10.2$, range=26.5 to 64.5 months). Data collection for this project began in 2012 and ran for several years. Participants in the early part of data collection had been initially diagnosed under DSM-IV, and participants in the latter part of data collection were all diagnosed under DSM-V. Per parent report, all children had a diagnosis of autism spectrum disorder, autism, or PDD-NOS; diagnosis was confirmed in the lab with the ADOS-2 (Autism Diagnostic Observation Schedule, Modules 1–3 or Toddler Module; [86, 87]). Parents reported that their child was exposed to English at least 70% of the time and had no history of hearing loss or comorbid developmental disorders. Participants were White (90%) or mixed ethnicity (10%). The majority of mothers (65%) had at least a college degree; one family did not provide maternal education information. An additional 12 children participated but were excluded from analysis because they contributed insufficient data (see below).

Expressive vocabulary was assessed using the MBCDI Level 2 Short Form A [71]. Parents reported that their children produced, on average, 62 of the 100 words on the checklist ($SD=32$, range=0–99). Parents of two children did not complete the MBCDI. Three subtests from the Mullen Scales of Early Learning (MSEL: [73]), widely

Table 3 Participant characteristics (experiment 2). Participants V02 and V03 and participants V10 and V19 are sets of twins. Testing took place within one month of participation in the experimental task

ID	Age	Sex	List	MacArthur Bates communicative development inventories: raw scores	Mullen scales of early learning: raw score (T-score)		
					Visual reception	Receptive language	Expressive language
V01	26.5	M	1	0	12 (20)	10 (<20)	10 (<20)
V02	26.8	M	1	20	21 (34)	11 (20)	14 (28)
V03	26.8	M	1	0	24 (43)	16 (27)	14 (28)
V04	33.6	M	1	85	39 (66)	34 (62)	35 (65)
V05	40.3	M	1	31	31 (35)	27 (34)	16 (<20)
V06	42.9	M	1	74	32 (47)	31 (45)	40 (53)
V07	44.6	M	1	85	32 (28)	34 (44)	32 (39)
V08	44.7	M	1	80	45 (61)	41 (59)	30 (34)
V09	48.1	M	1	75	45 (56)	31 (32)	28 (25)
V10	56.1	F	1	54	29 (<20)	27 (<20)	27 (<20)
V11	33.7	M	2	77	30 (48)	36 (67)	31 (56)
V12	35.5	F	2	89	21 (20)	34 (56)	36 (61)
V13	36.8	M	2	63	40 (63)	30 (45)	27 (40)
V14	40.4	M	2	29			
V15	41.5	M	2		28 (21)	26 (25)	18 (<20)
V16	42.4	M	2	67	31 (30)	28 (32)	29 (36)
V17	45.3	M	2	97	44 (63)	40 (61)	40 (59)
V18	51.1	M	2		26 (<20)	30 (26)	31 (29)
V19	56.1	F	2	87	34 (20)	32 (24)	30 (20)
V20	64.6	M	2	99	50 (56)	48 (54)	48 (51)

used with autistic children (e.g., [56]), were administered. We report these scores as raw numbers with age equivalents because standardized scores may fail to capture variability within a narrowed range [88, 89]. On the Expressive Language subscale (MSEL-EL), the average raw score was 27 ($SD=9.4$, range=10–48, age equivalent=29 months); on the Receptive Language subscale (MSEL-RL), 30 ($SD=9.5$, range=10–48, age equivalent=33 months); and on the Visual Reception subscale (MSEL-VR), 32 ($SD=9.4$, range=12–50, age equivalent=31 months). More than half the children's t -scores were more than one standard deviation below the mean for their chronological age on the MSEL-RL, and the MSEL-EL, indicating that they had delayed language development (see Table 3). Children were randomly assigned to one of two stimuli lists described below; no group differences existed between lists with respect to age, MBCDI, or MSEL scores.

Apparatus

Identical to experiment 1.

Stimuli

Whereas experiment 1 used a subset of the stimuli developed by Konishi et al. [74] and Valteau et al. [17]; in

experiment 2, we used all 36 verb trials and 14 filler noun trials. Otherwise, the stimuli and trial structure were identical to experiment 1.

Children were randomly assigned to one of two trial lists ($N=10$ children in each). Each list included the same visual stimuli, but the verb queried from the pair differed between the two lists. For example, on the trial depicting the events “clap” and “stretch,” children assigned to list 1 were asked to find “clap” while children assigned to list 2 were asked to find “stretch.” The order of trials varied between the lists. All children saw 18 verb trials and 7 filler noun trials, once each. Fourteen verb trials featured dynamic scenes with an actor and an object (e.g., “shaking” and “opening” a present) and four trials featured dynamic scenes with just an actor (e.g., “clapping” and “stretching”). See Additional file 2: Appendix B for the full list of trials.

Procedure overview

Participation was part of a two-visit protocol approved by Boston University's Institutional Review Board. At the first visit, parents provided written consent and completed a demographics questionnaire and the MBCDI. Children were assessed on the three relevant subscales of the MSEL and participated in an unrelated experimental

task. At the second visit, approximately 1 month later, children participated in this experimental task. The ADOS-2 was completed at the second visit or within a 6-month period of the child's participation in this study.

Exclusionary criteria

Given that autistic children, including those in our sample, vary widely in language abilities, we ran bootstrapped cluster-based permutation analyses separately for each child. This necessitated more stringent exclusionary criteria; we removed from analysis all trials in which track loss was greater than 33% during the Test Phase. We also removed from the sample all participants who lost more than half of their trials to this criterion ($n=12$ of the original 32 participants). There were no significant differences in average age or in MBCDI or MSEL scores between excluded participants and those in the final sample. From the final sample of participants ($N=20$), 17% of trials were excluded from analyses due to track loss.¹

Analysis

As with experiment 1, we were interested in two different measures: accuracy and processing speed. To analyze whether individual factors predicted eye gaze behaviors, we ran two mixed-effects regressions to explore the contributions of (1) vocabulary (MBCDI) and (2) receptive language (MSEL-RL). We elected for two models rather than one because MBCDI and MSEL-RL were highly correlated ($r=0.84$). Each trial for each participant was included as a separate data point. Models included gaze behavior as the dependent variable (accuracy or processing speed), the random effects of participant and trial, and the fixed effects of language measures (MBCDI or MSEL-VR), age, gender, and MSEL-VR score.

Accuracy

Whereas in experiment 1 we divided children by group to create response windows, in this study we created response windows for each child separately. Prior research has found wide variability in autistic children's language abilities, making a uniform response window across the sample unlikely. We therefore created

individualized response windows for each child using a procedure similar to experiment 1. But while in experiment 1, we ran two cluster-based analyses, one for the LT group and one for the TD group, in experiment 2 we ran a cluster-based analysis for each child separately using the eyetrackingR package [78] in R (Version 3.3.1; 78). Clusters were identified using a linear mixed-effects regression, which compared the child's Baseline proportion of looks to the target scene, collapsed over the 6-s Baseline Phase, to their proportion of looks to the target scene in each 50-ms bin of the Test Phase. The model included the dependent variable of proportion looks to the target scene versus elsewhere, the random effect of trial, and the predictor variable of phase (Baseline or Test). Time bins with a p value of less than 0.05 were included in a cluster.

In experiment 1, we then ran a t test for each group to identify significance of the clusters. In experiment 2 we wanted to evaluate each child individually, but, given the small number of trials each child saw, we did not have enough power to run a separate t test for each child; no child had more than 18 trials, and many had fewer. We acknowledge this as a limitation of this approach that will need to be addressed in subsequent research on the feasibility of individualizing response windows. Instead, we performed a paired t test for all participants together, comparing, by-trial and by-child, the proportion of looks to the target during Baseline to the proportion of looks to the target scene during that child's cluster as identified by the cluster analysis. This step was therefore parallel to the group analysis in experiment 1, even though the clusters were initially determined for each child individually. This allowed us to provide additional descriptive statistics on gaze behavior differences between Baseline and Test.

As with experiment 1, adjacent time bins in which gaze behaviors differed significantly between Baseline and Test, and those separated by only 50 ms, were combined into a single cluster. The first cluster lasting at least 500 ms was termed the child's "sustained preference window." We required that the cluster lasted at least 500 ms to eliminate short clusters (e.g., 50–100 ms) that may have been spurious differences as a result of scanning behavior rather than an indication of vocabulary knowledge. Four children did not have any clusters lasting 500 ms, so their "sustained preference window" was instead their longest cluster.² This sustained preference window was the basis for each child's individual response window. Individual response windows began at the start of each child's sustained preference window and were standardized to 1500 ms in duration (consistent with

¹ One possible limitation of our procedure that may have amplified track loss rates is that we did not include standardized breaks throughout the task, although we did pause the procedure and take breaks if we (or the child's parent) thought that children needed to. As other researchers have noted, some autistic children may need frequent breaks, while others may prefer to view multiple videos in one sitting (e.g., [94]). While all participants had participated in an unrelated eye-tracking task a month before—evidence for their ability to watch videos on an eye tracker—this task was several minutes longer than that one. As a design feature for future iterations of this study, we plan to explore how many and what types of breaks best support children's participation without requiring them to engage in too many transitions, which can also be challenging for autistic children.

² Three children did not have any significant 50-ms time bins (V01, V05, V16). These participants were excluded from further accuracy analyses.

experiment 1). Three children's sustained preference windows began within the last 1500 ms of the Test Phase, so they were assigned the individual response window of 4500–6000 ms.

Accuracy was calculated by-child by-trial by comparing the proportion of looks to the target scene during the whole of the Baseline Phase and during the child's individual response window of the Test Phase. Children were credited with knowing the verb if they demonstrated a 15% increase in looks between Baseline and their individual response window [7]. Each child's overall accuracy was calculated as a proportion of the number of trials correct over the number of trials contributed to account for the trials removed due to track loss.

Processing speed

As with experiment 1 (and 17), latency was operationalized as the first look to the target scene, by-child by-trial. Looks during the first 50 ms were excluded as chance looking based on trial design, and participants who did not look to the target scene during the Test Phase were given a latency of 6000 ms.

Results

De-identified gaze data are available on the Open Science Framework (<https://osf.io/egwya/>). We began by graphing each participant's individual performance. See examples in Fig. 3. Here, children's average proportion of looks to the target scene versus elsewhere (including looks to neither scene and track loss) during the Test Phase is depicted. The bar indicates the average proportion of looks to the target scene during all time points at Baseline. We observe that all three participants shown in Fig. 3 preferred the target scene during Test above and beyond Baseline looking; however, each child did so at a different point. This is particularly striking for participants V04 and V07, who had identical MBCDI and MSEL-RL scores. Participant V07 demonstrated a preference for the target scene early in the Test Phase, whereas V04 did not do so until the latter half. V20, who had the highest scores of any participant on the MBCDI and MSEL-RL, looked most consistently to the target scene during the middle of the Test Phase. We picked these three examples because they also demonstrate

broad patterns observed in our sample: Some of the children ($n=7$) tended to peak in the beginning third of the response window (before 2000 ms), but the majority had either a middle peak ($n=3$; 2000–4000 ms) or a late peak ($n=5$; after 4000 ms). (Notably, not all participants could be easily classified into “early” “middle” or “late” responders, instead spanning across these arbitrary divisions or because it was not clear that they distinguished their looks from Baseline at any point $n=5$). This division is noteworthy given that children received a second verbal prompt beginning at 2000 ms to, e.g., “Find clapping!” Some participants, and most likely those with late peaks, may have benefitted from this second verbal directive to find the target.

Accuracy

Using cluster-based permutation analyses, we identified each child's individualized response window (see Table 4). As a group, the paired t test indicated that average proportion of looks to the target scene significantly differed between the Baseline Phase ($M=0.41$, $SD=0.21$) and children's individualized response windows in the Test Phase ($M=0.64$, $SD=0.35$; $t(246)=9.71$, $p<0.001$). However, there was considerable within-group variability for the average change in proportion looking between Baseline and the individualized response window ($M=0.22$, $SD=0.12$, range = -0.04 – 0.37).

Children knew, on average, 60% of the verbs queried ($SD=0.14$, range = 0.40 – 0.83). In Table 5, we list the trials in which children were the most accurate in their responses. We observed that the top four trials in which children were most accurate were the those of run versus jump; dance versus cry; kiss versus tickle; and stretch versus clap. Interestingly, three of these four featured only actors and no objects in the visual stimuli.

In our vocabulary model, MBCDI scores had a marginal but non-significant relationship with children's accuracy ($b=0.02$, $t=1.86$, $p=0.06$, *n.s.*); no other factors were significant ($b_{\text{age}}=-0.004$, $t_{\text{age}}=-0.89$, $p_{\text{age}}=0.37$, *n.s.*; $b_{\text{gender}}=0.03$, $t_{\text{gender}}=0.27$, $p_{\text{gender}}=0.78$, *n.s.*; $b_{\text{MSEL-VR}}=0.008$, $t_{\text{MSEL-VR}}=0.56$, $p_{\text{MSEL-VR}}=0.57$, *n.s.*). In our receptive language model, MSEL-RL scores predicted performance: Children with higher MSEL-RL scores performed better ($b=0.02$, $t=2.18$, $p=0.03$). No

(See figure on next page.)

Fig. 3 A–C Timecourse graphs for three individual participants in experiment 2. The timecourse displays each child's average proportion of looks to the target scene versus elsewhere across the testphase. Error bars represent standard error of trial means. The boxes indicate times in which proportion of looking to the target was significantly greater in the Test Phase over the Baseline Phase. The horizontal bar represents children's individual baseline average, calculated as the proportion of looks to the target versus elsewhere across the entire baseline phase, collapsing across all trials. Participants V04 (A) and V07 (B) had identical MBCDI and MSEL-RL raw scores. Participant V20 (C) had the highest MBCDI and MSEL raw scores

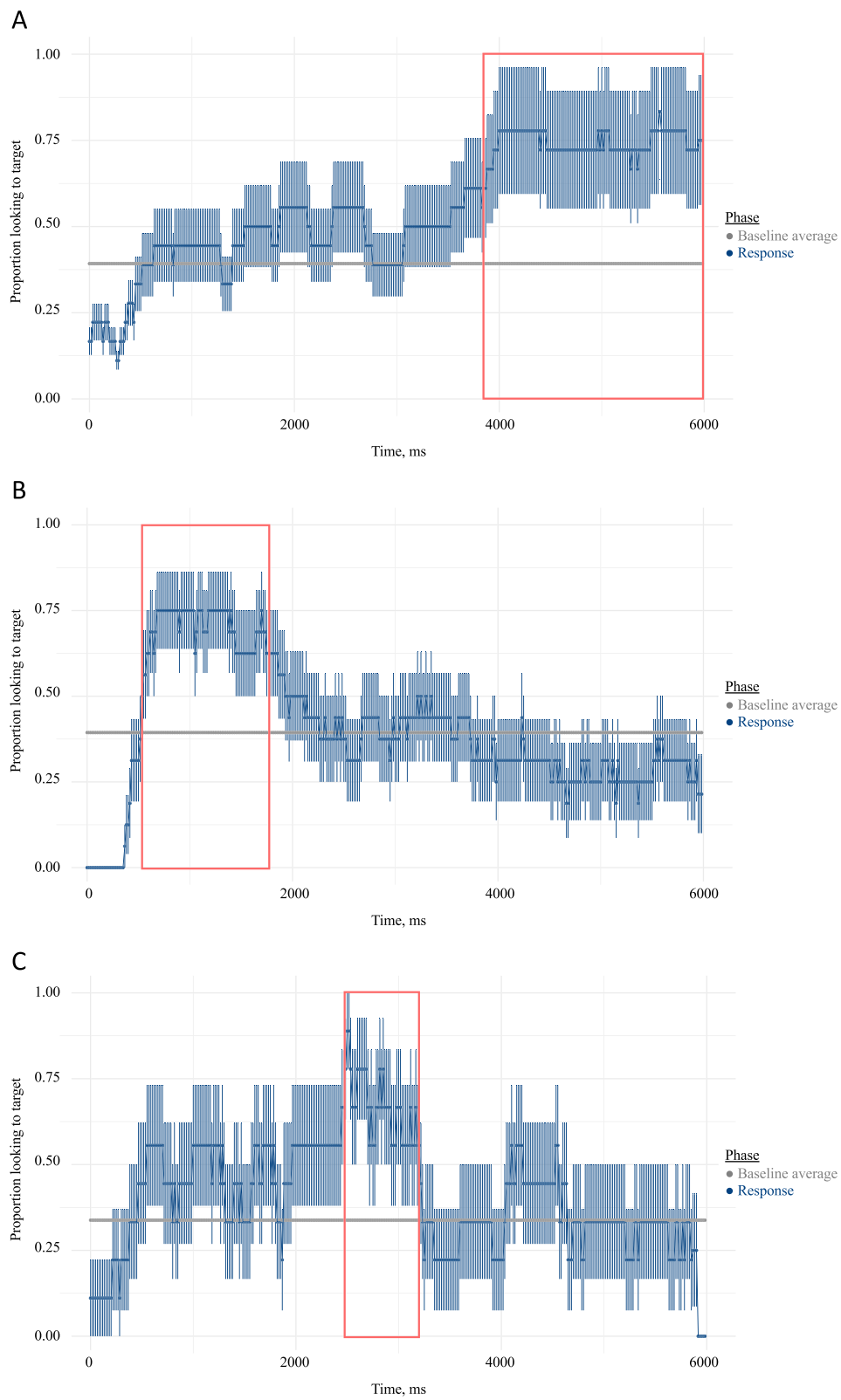


Fig. 3 (See legend on previous page.)

Table 4 Children’s sustained preference window, individual response window, and accuracy proportions (experiment 2)

ID	Sustained preference window	Individual response window (for accuracy analysis)	Accuracy proportion (from individual response window)
V01			
V02	1600–1900 ms	1600–3100 ms	0.44
V03	3300–3900 ms	3300–4800 ms	0.40
V04	3850–6000 ms	3850–5350 ms	0.83
V05			
V06	800–2300 ms	800–2300 ms	0.67
V07	550–1750 ms	550–2050 ms	0.75
V08	1150–3000 ms	1150–2650 ms	0.69
V09	4700–5100 ms	4500–6000 ms	0.50
V10	3900–4700 ms	2900–5400 ms	0.50
V11	1400–2750 ms	1400–2900 ms	0.75
V12	4700–6000 ms	4500–6000 ms	0.67
V13	400–3050 ms	400–2100 ms	0.67
V14	5800–5950 ms	4500–6000 ms	0.46
V15	1600–1650 ms	1600–3100 ms	0.42
V16			
V17	800–2600 ms	800–2300 ms	0.82
V18	500–1400 ms	500–2000 ms	0.63
V19	2450–3850 ms	2450–3950 ms	0.64
V20	2450–3200 ms	2450–3950 ms	0.44

other factors were significant ($b_{age} = -0.004$, $t_{age} = -0.89$, $p_{age} = 0.37$, *n.s.*; $b_{gender} = 0.03$, $t_{gender} = 0.27$, $p_{gender} = 0.78$, *n.s.*; $b_{MSEL-VR} = 0.008$, $t_{MSEL-VR} = 0.56$, $p_{MSEL-VR} = 0.57$, *n.s.*)³.

Table 5 The rank order of trials with greatest to least accuracy (experiment 2)

Rank order	Trial	Proportion
1	Run-Jump	0.82
2	Dance-Cry	0.80
3	Kiss-Tickle	0.79
3	Stretch-Clap	0.79
5	Kick-Throw	0.71
6	Cut-Tie	0.69
7	Roll-Bounce	0.67
7	Squeeze-Blow	0.67
9	Lift-Pull	0.64
10	March-Spin	0.62
10	Pour-Drink	0.62
12	Wash-Rock	0.57
13	Shake-Open	0.56
14	Lick-Break	0.53
14	Read-Rip	0.53
16	Feed-Hug	0.45
17	Eat-Push	0.43
18	Drop-Bite	0.40

Processing speed

Children averaged a latency of 1030 ms before their first look to the target scene ($SD = 378$, range 452–1761). In our vocabulary model, no fixed effect significantly predicted children’s latency to the target scene ($b_{MBCDI} = -28$, $t_{MBCDI} = -0.79$, $p_{MBCDI} = 0.37$, *n.s.*; $b_{age} = -0.61$, $t_{age} = -0.04$, $p_{age} = 0.94$, *n.s.*; $b_{gender} = 101$, $t_{gender} = 0.22$, $p_{gender} = 0.81$, *n.s.*; $b_{MSEL-VR} = -43$, $t_{MSEL-VR} = -0.90$, $p_{MSEL-VR} = 0.31$, *n.s.*). However, in our receptive language model, MSEL-RL scores significantly

³ As a post-hoc analysis, we explored whether the start or duration of children’s sustained preference windows correlated with individual factors. We found that the duration of the sustained preference window correlated with MBCDI scores ($r = 0.60$, $t = 2.7$, $p = .02$), but not MSEL-RL scores ($r = 0.32$, $t = 1.3$, $p = .23$, *n.s.*) or age ($r = -0.02$, $t = -0.08$, $p = .94$, *n.s.*). This suggests that children with larger expressive vocabularies can direct more sustained attention to a labeled scene. The starting time point of children’s sustained preference windows did not correlate with language abilities ($r_{MBCDI} = -0.25$, $t_{MBCDI} = -0.97$, $p_{MBCDI} = .35$, *n.s.*; $r_{MSEL-RL} = 0.03$, $t_{MSEL-RL} = 0.10$, $p_{MSEL-RL} = .92$, *n.s.*) or age ($r = -0.03$, $t = -0.13$, $p = .90$, *n.s.*), indicating that how early children settle on the target scene is independent of language or broader developmental factors.

predicted latency ($b = -109$, $t = -2.5$, $p = 0.008$). No other factors were significant ($b_{\text{age}} = 0.11$, $t_{\text{age}} = 0.08$, $p_{\text{age}} = 0.97$, *n.s.*; $b_{\text{gender}} = 13$, $t_{\text{gender}} = 0.04$, $p_{\text{gender}} = 0.99$, *n.s.*; $b_{\text{MSEL-VR}} = 6.4$, $t_{\text{MSEL-VR}} = 0.15$, $p_{\text{MSEL-VR}} = 0.85$, *n.s.*).

We also explored possible correlations between latency and sustained preference window. These are distinct measures: Latency indicates the speed of children's first look on any given trial to the target scene, but a sustained attention window reflects prolonged fixations on the target and is representative of children's gaze behaviors across the whole of the experimental session. However, both may be indicative of children's processing abilities. We found that children's average latency across trials correlated with the start of the sustained preference window ($r = 0.60$, $t = 2.92$, $p = 0.01$), indicating that children with shorter average latencies were also faster to demonstrate a sustained preference to the target scene. Average latency also significantly negatively correlated with the length of the sustained preference window ($r = -0.74$, $t = -4.33$, $p < 0.001$), indicating that children with shorter average latencies also had longer sustained attention windows, perhaps suggesting more robust lexical representations.

Discussion

In experiment 2, we assessed the receptive verb vocabularies of autistic children, considering both the number of words they know and how quickly they process language. We found that both children's accuracy and processing speed were predicted by concurrent receptive language skills but not expressive vocabulary. This is likely due to the complex relationship between receptive and expressive language among autistic children. For example, our sample included two participants who were nonspeaking (V01 and V03), one of whom (V03) knew nearly half of all verbs presented in the task. We argue that this underscores the importance of full and accurate assessment of receptive language abilities.

One notable contribution of experiment 2 is that we established response windows during the Test Phase that were individualized to each child's gaze behaviors. This is an approach worth future research: Given the heterogeneity of profiles of autistic children [59–61], researchers and clinicians would be well served to have a systematic method of identifying the correct response window at the individual level. But, we also acknowledge major limitations to the approach we have presented. First, our study included too few trials overall, such that we were unable to determine whether the windows identified by the bootstrapped cluster-based permutation analysis for each child significantly differed from chance levels; our paired t test was run at the group, rather than individual, level, even though our clusters were calculated by-child.

A longer study with more trials, however, would likely have exceeded children's attention spans. Thus, there is no easy resolution to this limitation. Second, we applied stringent exclusionary criteria—which was necessary to have enough data to run this bootstrap analysis—but in so doing excluded 12 of the initial 32 participants recruited for this study. This is sizeable. Although we engaged more children in our task than may be able to complete standardized assessments—for example, Brady and colleagues found that nearly half of minimally verbal 4- to 7-year-old autistic children are unable to pass screening items for the Peabody Picture Vocabulary Test [22]—this task with minimal task demands may still leave out many children.

General discussion

Although prior research has demonstrated the feasibility of using eye tracking to assess receptive vocabulary of a variety of word types (e.g., 8, [15–17]), several gaps in the current literature must be addressed before this technology can be used in clinical settings. First, prior research has primarily focused on noun vocabulary depicted by static images. However, it is important to also assess verb vocabulary, which is more appropriately depicted with dynamic scenes. Second, most prior research has involved typically developing children, although there are notable exceptions (e.g., [3, 11, 41, 44–46]). However, children with language delays and disorders, including late talkers (experiment 1) and autistic children (experiment 2), may exhibit different patterns of gaze behavior during eye-tracking assessments. It is important that research address both of these gaps, while using appropriate measures of eye gaze for these stimulus types and populations.

Experiments 1 and 2 demonstrated one approach to measuring accuracy and processing speed, with considerations for both the stimuli and populations included. Accuracy—taken as evidence that a child knows a word—has been operationalized in previous work as a 15% increase in looking to the target from a baseline period to a “response window” that lasts from 300 to 1800 ms after the auditory prompt to look to the target. However, prior research has made clear that for dynamic scenes, a different response window is needed. We therefore applied bootstrapped cluster-based permutation analyses [66] to determine which portion of the Test Phase would be most appropriate.

In experiment 1, we used bootstrapped cluster-based permutation analysis at the group level, hypothesizing that late talkers would require more time than their typically developing peers to demonstrate knowledge of target words. This hypothesis was confirmed: Late talkers required approximately one half-second more to orient

their gaze to the target scene. Given this extra time, however, we observed no differences in the average number of verbs late talkers and typically developing children knew.

In experiment 2, we applied bootstrapped cluster-based permutation analysis individually, yielding an individualized response window for each child in the final sample. This can be interpreted as evidence for the feasibility of the approach, but more research is needed to determine whether this is reliable and valid. Given the small number of trials included, our statistical power was limited and we were forced to run a group-level paired *t* test when individual *t* tests would have been more appropriate. This is further exacerbated by the fact that permutation analyses assume exchangeability of data, but there is a higher interdependence of performance within a single participant, potentially leading to elevated type 1 errors. Finally, we adopted conservative exclusionary criteria, losing a sizeable portion of the initial sample in the process. Future research must balance the needs for high quality data with the goal of including as many participants as possible.

A second commonly used eye-gaze measure, latency, is well established for assessing lexical processing given noun targets and static images, but it is unclear whether it is similarly informative given verb targets and dynamic scenes (e.g., [11, 17]). The current experiments provide new insight into prior, discrepant findings. In experiment 1, we found that age but not language predicted processing speed for late talkers and typically developing children. Our tentative hypothesis from this result was that, between 2 and 3 years of age, there is a maturation of processing ability that results in a tighter relation between children's language knowledge and their ability to quickly look at a named target, even when the scenes are dynamic. Experiment 2 also supports this hypothesis: For older, autistic preschoolers, the majority of whom ($n=14$ of 20) were greater than 3 years of age, processing speed was reliably predicted by concurrent receptive language abilities. In further support for our hypothesis, we noted that in experiment 1, 2-year-olds' average latency to look to the target was 1500 ms, while in experiment 2, older autistic children averaged a much shorter latency of 1030 ms, even though the majority of these children had delayed language. While the data from these two experiments supports our hypothesis, it is important to note that this is but one possible explanation for the discrepant findings across the literature. It is important that researchers find a reliable way to operationalize verb processing, and to that point we can only state that latency does not appear to be always the best measure. Future work is necessary to determine when latency is

appropriate, when it is not, and what other measures may be used to capture processing abilities.

Limitations and cautionary tales

In addition to those already discussed, we note several additional limitations to this study. First, the samples in both experiments are not representative of contemporary U.S. demographics: Participants were disproportionately white, and most came from higher socio-economic backgrounds. Both limit the generalizability of our findings. For example, we acknowledge that children from higher socioeconomic backgrounds may have different experiences with technology, which may impact their engagement with the task.

Second, we interpret our accuracy measurement with a word of caution. Results indicate that late talkers (experiment 1) and autistic children (experiment 2) have some knowledge of the target verbs. However, we cannot claim from this measure that our children with language delays and disorders have equally robust representations to those of their typically developing peers' (experiment 1) or even to one another (experiment 2). Indeed, it is likely that there are differences across and within groups robustness, as has been observed in older children with developmental language disorder (e.g., [90, 91]). Some studies have attempted to operationalize other eye-gaze behaviors to measure robustness (e.g., Yu and Smith [83], who observed differences in overall looking time between "strong" and "weak" word learners), but these measures are not well-established. We see some evidence for other between-group differences in gaze behavior, including overall proportion looking to the target, but further research in this area is required.

Third, it is clear from the participants' highly varied performance in experiment 2 that researchers and clinicians must be able to individualize response windows to each child's unique patterns of gaze behavior (see Fig. 3A–C). However, we caution readers against settling too quickly on bootstrapped cluster-based permutation analyses as the best approach to take. Although we demonstrated the feasibility of using this method to identify individualized response windows, there remain many questions and concerns regarding this application. For example, the statistical approach for identifying individualized response windows must be reproducible, but we are unable to assess reproducibility given the current data. Second, as we have discussed, using this method requires a delicate balancing act between a high number of trials and participant demands and between inclusionary/exclusionary criteria that is stringent enough to run the analysis but relaxed enough to include the broad spectrum of abilities. Of course, we recognize that in a

continually evolving field, new and better methods may soon be developed.

Fourth, automated eye-tracking currently carries limitations that almost certainly affected the current study. Despite the widespread use of automated eye-tracking with young autistic children, some studies have suggested that manual coding of eye gaze yields more data than automated eye-tracking for autistic children [5, 12], and the relatively higher rates of track loss within experiment 2 are consistent with the possibility that autistic children could have contributed more trials had we used manual coding instead. Looking toward the ultimate goal of clinical use of eye-tracking assessments of vocabulary, we cannot suggest that manual coding is an appropriate solution; however, we note that eye-tracking technology is continually improving and it is likely that advances will reduce track loss rates to an acceptable level even with automated measures.

Finally, it is important to raise the possibility that it may be difficult to ever garner robust and clinically valuable item-level data from an eye-tracking assessment. If each word is only assessed once, as in the current design, there is the risk that children did not happen to be attending on that trial. If multiple times, there is the risk that children are learning throughout the course of the task, or alternatively, habituating to the target (if they know the target word) and focusing more on the distractor. We suggest that one possible use of an eye-tracking assessment will be to capture receptive language abilities that may be otherwise difficult to measure for certain populations who have challenges participating in standardized assessments. Understanding receptive language skills, and their relation to expressive abilities, is critical for clinical decision-making. For example, the outcome of an eye-tracking assessment could reveal whether a minimally speaking autistic child needs more support with communicating verb concepts for which they already have receptive knowledge or whether they require intervention to learn those verbs in the first place.

Future directions

One benefit of assessing receptive vocabulary using eye gaze is that performance on noun trials predicts later language and developmental outcomes (e.g., [10]); in theory, such information may have a role in clinical decision-making. Although verbs may be particularly powerful predictors of later language skills [18, 19], whether eye gaze measures for receptive verb vocabulary can predict later language abilities is yet unstudied. We propose future work considering whether verb processing (latency, for older children) and broader processing abilities (start time of individualized response window) predict later outcomes. We are particularly interested in

the latter: Although we found no concurrent relationship between children's sustained attention window (from which we derived individual response windows) and language abilities, it may be that the confluence of factors that result in these individualized differences may impact long-term vocabulary acquisition. For example, if the starting time of a sustained attention window is reflective of children's abilities to integrate visual and linguistic information, we might expect that this would impact vocabulary acquisition: Children who are slower to process words they know may miss opportunities to acquire the meanings of unfamiliar words in the same sentence (e.g., [10, 92]).

We would also like to see further research exploring individualized response windows. Beyond the questions we have already raised about the reliability and reproducibility of the method, it would also be beneficial to explore this approach in a variety of populations, including possibly revisiting this idea with late talkers and typically developing children.

Finally, we acknowledge that there are significant barriers to be addressed before eye-tracking can be reliably used in clinical settings. While there are numerable benefits to be had for using eye-tracking in research to answer questions about language development and disorder, one goal of our research is for its ultimate clinical application. Substantial research and practice changes would need to be made before this could happen. One limitation is undoubtedly that eye-tracking is cost-prohibitive for most clinical settings, meaning that most clinicians do not have access to it. With recent advances in technology, including pandemic-driven research into using computer cameras for eye-tracking data [93], we hope that this will present less of a barrier in the future. Additionally, extensive research would be required to consider how best to translate this method to a standardized assessment, considering such factors as task demands, duration, and target selection. This, too, will take time. However, we hope that this work is a first step into developing tools so that speech-language pathologists may accurately assess a broad spectrum of language abilities in clinic.

Conclusion

This study addresses two notable gaps in prior literature on assessing receptive vocabulary in children with language delays using eye tracking. First, an assessment task should include many types of words, including verbs; however, most prior studies have included only noun targets depicted by static images (e.g., [8, 27]; although, see [9–11]). Second, although some research has been done on assessing receptive vocabulary in children with language delays and disorders [3, 27, 35, 38–40], more research is warranted on how best

to interpret eye gaze behaviors as an indication of linguistic knowledge. We demonstrated one way to measure children's accuracy and lexical processing of verbs, given considerations for the type of stimuli (dynamic scenes rather than static images) and populations (late talkers, autistic children). Our findings highlight the importance of adapting these measures at the group and even at the individual level to account for variation in performance.

In considering accuracy, we have demonstrated the feasibility of applying a bootstrapped cluster-based permutation analysis [61] to identify response windows for calculating accuracy. Using this approach, we found in experiment 1 no differences in the average proportion of verbs that late talkers and typically developing toddlers knew, although late talkers required more time to demonstrate this knowledge. In experiment 2, we found that the number of verbs children knew in the eye tracking task was predicted by their receptive language abilities. However, we interpret these results with caution: As highlighted in our discussion, we were underpowered to run *t*-tests for significance given individual windows, limiting our interpretations. We strongly encourage future research on this or alternative approaches to individualizing response windows based on gaze behavior.

In considering lexical processing, results from experiments 1 and 2 shed light on prior, discrepant conclusions about the appropriateness of using latency as a measure of lexical processing given dynamic scene stimuli. We argue that latency is not a reliable measure of lexical processing for children younger than 3 years of age in this context. In experiment 1, age but not language ability predicted latency for 2-year-old late talkers and typically developing children. In experiment 2, however, receptive language skills predicted latency for older autistic preschoolers. Future research is warranted on how best to operationalize lexical processing for dynamic scene stimuli in younger children.

Using eye gaze may allow us to assess receptive vocabulary skills in children who might otherwise be unable to participate in standardized assessments [1–7]. The findings of this study support this long-term goal, but more research is needed in how best to adapt assessments to such populations.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s11689-023-09512-x>.

Additional file 1: Appendix A. List of trials for Experiment 1.

Additional file 2: Appendix B. List of trials for Experiment 2.

Acknowledgements

Thank you to Mathew Gregoski, Ph.D., and Naomi Brownstein, Ph.D., for their statistical support on this project; and to Haruka Konishi, Aimee Stahl, Roberta Golinkoff, and Kathy Hirsh-Pasek for sharing their visual stimuli.

Authors' contributions

SA conceptualized the study; SH and SA both contributed to experimental design and the data analysis plan; analyses were carried out by SH. Both authors contributed to the writing of the manuscript and approved its final version.

Funding

This research was supported by NIH R01DC016592 and an ASHFoundation New Century Scholars Research Grant to SA; an ASHFoundation Student Research Grant in Early Childhood Language Development (supported by the Arlene M. and Noel D. Matkin Memorial Fund) to SH, and a Boston University Dudley Allen Sargent Research Fund grant to SH.

Availability of data and materials

De-identified data are publicly available on the Open Science Framework (Experiment 1: <https://osf.io/ghp7q/>; Experiment 2: <https://osf.io/egwya/>).

Declarations

Ethics approval and consent to participate

Experiments 1 and 2 were approved by the Boston University IRB. All parents provided consent on behalf of their children prior to participation.

Consent for publication

Individuals pictured provided consent for use of their image in publication.

Competing interests

The authors declare no competing interests.

Received: 21 February 2023 Accepted: 16 November 2023

Published online: 13 December 2023

References

1. Cauley KM, Golinkoff RM, Hirsh-Pasek K, Gordon L. Revealing hidden competencies: a new method for studying language comprehension in children with motor impairments. *Am J Ment Retard.* 1989;94(1):53–63.
2. Lampe R, Turova V, Blumenstein T, Alves-Pinto A. Eye movement during reading in young adults with cerebral palsy measured with eye tracking. *Postgrad Med.* 2014;126(5):146–58. <https://doi.org/10.3810/pgm.2014.09.2809>.
3. Brady NC, Anderson CJ, Hahn LJ, Obermeier SM, Kapa LL. Eye tracking as a measure of receptive vocabulary in children with autism spectrum disorders. *Augment Altern Commun.* 2014;30(2):147–59. <https://doi.org/10.3109/07434618.2014.904923>.
4. Venker CE, Eernisse ER, Saffran JR, Weismer SE. Individual differences in the real-time comprehension of children with ASD. *Autism Res.* 2013;6(5):417–32. <https://doi.org/10.1002/aur.1304>.
5. Venker CE, Kover ST. An open conversation on using eye-gaze methods in studies of neurodevelopmental disorders. *J Speech Lang Hear Res.* 2015;58(6):1719–32. https://doi.org/10.1044/2015_JSLHR-L-14-0304.
6. Bergelson E, Swingle D. At 6–9 months, human infants know the meanings of many common nouns. *Proc Natl Acad Sci.* 2012;109(9):3253–8. <https://doi.org/10.1073/pnas.1113380109>.
7. Reznick JS. Visual preference as a test of infant word comprehension. *Appl Psycholinguist.* 1990;11(2):145–66. <https://doi.org/10.1017/S0142716400008742>.
8. Fernald A, Perfors A, Marchman VA. Picking up speed in understanding: speech processing efficiency and vocabulary growth across the 2nd year. *Dev Psychol.* 2006;42(1):98. <https://doi.org/10.1037/0012-1649.42.1.98>.
9. Killing SE, Bishop DV. Move it! Visual feedback enhances validity of preferential looking as a measure of individual differences in vocabulary in

- toddlers. *Dev Sci.* 2008;11(4):525–30. <https://doi.org/10.1111/j.1467-7687.2008.00698.x>.
10. Marchman VA, Fernald A. Speed of word recognition and vocabulary knowledge in infancy predict cognitive and language outcomes in later childhood. *Dev Sci.* 2008;11(3):F9-16. <https://doi.org/10.1111/j.1467-7687.2008.00671.x>.
 11. Fernald A, Marchman VA. Individual differences in lexical processing at 18 months predict vocabulary growth in typically developing and late-talking toddlers. *Child Dev.* 2012;83(1):203–22. <https://doi.org/10.1111/j.1467-8624.2011.01692.x>.
 12. Venker CE, Edwards J, Saffran JR, Ellis WS. Thinking ahead: incremental language processing is associated with receptive language abilities in preschoolers with autism spectrum disorder. *J Autism Dev Disord.* 2019;49:1011–23.
 13. Zhou P, Zhan L, Ma H. Predictive language processing in preschool children with autism spectrum disorder: an eye-tracking study. *J Psycholinguist Res.* 2019;48:431–52.
 14. Lany J. Lexical-processing efficiency leverages novel word learning in infants and toddlers. *Dev Sci.* 2018;21(3):e12569. <https://doi.org/10.1111/desc.12569>.
 15. Goldfield BA, Gencarella C, Fornari K. Understanding and assessing word comprehension. *Appl Psycholinguist.* 2016;37(3):529–49. <https://doi.org/10.1017/S0142716415000107>.
 16. Golinkoff RM, Hirsh-Pasek K, Cauley KM, Gordon L. The eyes have it: lexical and syntactic comprehension in a new paradigm. *J Child Lang.* 1987;14(1):23–45. <https://doi.org/10.1017/S030500090001271X>.
 17. Valteau MJ, Konishi H, Golinkoff RM, Hirsh-Pasek K, Arunachalam S. An eye-tracking study of receptive verb knowledge in toddlers. *J Speech Lang Hear Res.* 2018;61(12):2917–33. https://doi.org/10.1044/2018_JSLHR-L-17-0363.
 18. Hadley PA, Rispoli M, Hsu N. Toddlers' verb lexicon diversity and grammatical outcomes. *Lang Speech Hear Serv Sch.* 2016;47(1):44–58. https://doi.org/10.1044/2015_LSHSS-15-0018.
 19. LeGrand KJ, Weil LW, Lord C, Luyster RJ. Identifying childhood expressive language features that best predict adult language and communication outcome in individuals with autism spectrum disorder. *J Speech Lang Hear Res.* 2021;64(6):1977–91. https://doi.org/10.1044/2021_JSLHR-20-00544.
 20. Ellis Weismer S, Evans JL. The role of processing limitations in early identification of specific language impairment. *Top Lang Disord.* 2002;22(3):15–29.
 21. Horvath S, Arunachalam S. Optimal contexts for verb learning. *Perspectives of the ASHA special interest groups.* 2019;4(6):1239–49.
 22. Dunn LM, Dunn DM, Lenhard A, Lenhard W, Suggate S. PPVT-4: Peabody picture vocabulary test. Ausgabe. Pearson Assessments. 2015.
 23. Cocking RR, McHale S. A comparative study of the use of pictures and objects in assessing children's receptive and productive language. *J Child Lang.* 1981;8(1):1–3. <https://doi.org/10.1017/S030500090000297X>.
 24. Friedman SL, Stevenson MB. Developmental changes in the understanding of implied motion in two-dimensional pictures. *Child Dev.* 1975;773–8. <https://doi.org/10.2307/1128578>
 25. Dorr M, Martinetz T, Gegenfurtner KR, Barth E. Variability of eye movements when viewing dynamic natural scenes. *J Vision.* 2010;10(10):28.
 26. Smith TJ, Mital PK. Attentional synchrony and the influence of viewing task on gaze behavior in static and dynamic scenes. *J Vision.* 2013;13(8):16.
 27. Wass SV, Smith TJ. Individual differences in infant oculomotor behavior during the viewing of complex naturalistic scenes. *Infancy.* 2014;19(4):352–84.
 28. Koenig A, Arunachalam S, Saudino KJ. Lexical processing of nouns and verbs at 36 months of age predicts concurrent and later vocabulary and school readiness. *J Cogn Dev.* 2020;21(5):670–89. <https://doi.org/10.1080/15248372.2020.1802277>.
 29. Naigles L. Are English-speaking one-year-olds verb learners, too. In: *Proceedings of the 28th annual child language research forum 1997.* Center for the Study of Language and Information. pp. 199–212
 30. Andreu L, Sanz-Torrent M, Guàrdia-Olmos J. Auditory word recognition of nouns and verbs in children with Specific Language Impairment (SLI). *J Commun Disord.* 2012;45(1):20–34. <https://doi.org/10.1016/j.jcomdis.2011.09.003>.
 31. Bavin EL, Kidd E, Prendergast L, Baker E, Dissanayake C, Prior M. Severity of autism is related to children's language processing. *Autism Res.* 2014;7(6):687–94.
 32. Brock J, Norbury C, Einav S, Nation K. Do individuals with autism process words in context? Evidence from language-mediated eye-movements. *Cognition.* 2008;108(3):896–904.
 33. Ellis Weismer S, Haebig E, Edwards J, Saffran J, Venker CE. Lexical processing in toddlers with ASD: does weak central coherence play a role? *J Autism Dev Disord.* 2016;46:3755–69.
 34. Hahn N, Snedeker J, Rabagliati H. Rapid linguistic ambiguity resolution in young children with autism spectrum disorder: eye tracking evidence for the limits of weak central coherence. *Autism Res.* 2015;8(6):717–26.
 35. Kamio Y, Robins D, Kelley E, Swainson B, Fein D. Atypical lexical/semantic processing in high-functioning autism spectrum disorders without early language delay. *J Autism Dev Disord.* 2007;37:1116–22. <https://doi.org/10.1007/s10803-006-0254-3>.
 36. LaTourrette A, Waxman S, Wakschlag LS, Norton ES, Weisleder A. From recognizing known words to learning new ones: comparing online speech processing in typically developing and late-talking 2-year-olds. *J Speech Lang Hear Res.* 2023;66(5):1658–77.
 37. Naigles LR, Fein D. Looking through their eyes: tracking early language comprehension in ASD. In: *Innovative investigations of language in autism spectrum disorder.* 2017;49–69. <https://doi.org/10.1037/15964-004>
 38. Naigles LR, Hoff E. 13 verbs at the very beginning: parallels between comprehension and input. In: *Action meets word: How children learn verbs.* 2006:336.
 39. Prescott KE, Mathée-Scott J, Reuter T, Edwards J, Saffran J, Ellis Weismer S. Predictive language processing in young autistic children. *Autism Res.* 2022;15(5):892–903.
 40. Swensen LD, Kelley E, Fein D, Naigles LR. Processes of language acquisition in children with autism: evidence from preferential looking. *Child Dev.* 2007;78(2):542–57.
 41. Barone R, Spampinato C, Pino C, Palermo F, Scuderi A, Zavattieri A, Gulisano M, Giordano D, Rizzo R. Online comprehension across different semantic categories in preschool children with autism spectrum disorder. *PLoS ONE.* 2019;14(2):e0211802. <https://doi.org/10.1371/journal.pone.0211802>.
 42. Chita-Tegmark M, Arunachalam S, Nelson CA, Tager-Flusberg H. Eye-tracking measurements of language processing: developmental differences in children at high risk for ASD. *J Autism Dev Disord.* 2015;45:3327–38. <https://doi.org/10.1007/s10803-015-2495-5>.
 43. Frazier TW, Hauschild KM, Klingemier E, Strauss MS, Hardan AY, Youngstrom EA. Rapid eye-tracking evaluation of language in children and adolescents referred for assessment of neurodevelopmental disorders. *J Intellect Dev Disabil.* 2020;45(3):222–35. <https://doi.org/10.3109/13668250.2019.1698287>.
 44. Muller K, Brady N. Assessing early receptive language skills in children with ASD. *Perspect ASHA Special Interest Groups.* 2016;1(1):12–9. <https://doi.org/10.1044/persp1.SIG1.12>.
 45. Plesa Skwerer D, Jordan SE, Brukilacchio BH, Tager-Flusberg H. Comparing methods for assessing receptive language skills in minimally verbal children and adolescents with autism spectrum disorders. *Autism.* 2016;20(5):591–604. <https://doi.org/10.1177/1362361315600146>.
 46. Venker CE. Spoken word recognition in children with autism spectrum disorder: the role of visual disengagement. *Autism.* 2017;21(7):821–9. <https://doi.org/10.1177/1362361316653230>.
 47. Rescorla L. The language development survey: a screening tool for delayed language in toddlers. *J Speech Hearing Disord.* 1989;54(4):587–99. <https://doi.org/10.1044/jshd.5404.587>.
 48. Dale PS, Price TS, Bishop DV, Plomin R. Outcomes of early language delay. *JSLHR.* 2003;46:544–60. [https://doi.org/10.1044/1092-4388\(2003/044\)](https://doi.org/10.1044/1092-4388(2003/044)).
 49. Hammer CS, Morgan P, Farkas G, Hillemeier M, Bitetti D, Maczuga S. Late talkers: a population-based study of risk factors and school readiness consequences. *J Speech Lang Hear Res.* 2017;60(3):607–26.
 50. Paul R. Clinical implications of the natural history of slow expressive language development. *Am J Speech Lang Pathol.* 1996;5(2):5–21. <https://doi.org/10.1044/1058-0360.0502.05>.
 51. Fischel JE, Whitehurst GJ, Caulfield MB, DeBaryshe B. Language growth in children with expressive language delay. *Pediatrics.* 1989;83(2):218–27.

52. Paul R, Looney SS, Dahm PS. Communication and socialization skills at ages 2 and 3 in "late-talking" young children. *J Speech Lang Hear Res.* 1991;34(4):858–65. <https://doi.org/10.1044/jshr.3404.858>.
53. Beckage N, Smith L, Hills T. Small worlds and semantic network growth in typical and late talkers. *PLoS ONE.* 2011;6(5):e19348. <https://doi.org/10.1371/journal.pone.0019348>.
54. Colunga E, Sims CE. Not only size matters: early-talker and late-talker vocabularies support different word-learning biases in babies and networks. *Cogn Sci.* 2017;41:73–95. <https://doi.org/10.1111/cogs.12409>.
55. MacRoy-Higgins M, Shafer VL, Fahey KJ, Kaden ER. Vocabulary of toddlers who are late talkers. *J Early Interv.* 2016;38(2):118–29. <https://doi.org/10.1177/1053815116637620>.
56. Rescorla L, Alley A, Christine JB. Word frequencies in toddlers' lexicons. *J Speech Lang Hear Res.* 2001;44(3):598–609. [https://doi.org/10.1044/1092-4388\(2001\)049](https://doi.org/10.1044/1092-4388(2001)049).
57. Horvath S, Kueser JB, Kelly J, Borovsky A. Difference or delay? Syntax, semantics, and verb vocabulary development in typically developing and late-talking toddlers. *Lang Learn Dev.* 2022;18(3):352–76.
58. Horvath S, Rescorla L, Arunachalam S. The syntactic and semantic features of two-year-olds' verb vocabularies: a comparison of typically developing children and late talkers. *J Child Lang.* 2019;46(3):409–32. <https://doi.org/10.1017/S0305000918000508>.
59. Ellis Weismer S, Lord C, Esler A. Early language patterns of toddlers on the autism spectrum compared to toddlers with developmental delay. *J Autism Dev Disord.* 2010;40:1259–73. <https://doi.org/10.1007/s10803-010-0983-1>.
60. Kasari C, Brady N, Lord C, Tager-Flusberg H. Assessing the minimally verbal school-aged child with autism spectrum disorder. *Autism Res.* 2013;6(6):479–93. <https://doi.org/10.1002/aur.1334>.
61. Tager-Flusberg H. Defining language impairments in a subgroup of children with autism spectrum disorder. *Sci China Life Sci.* 2015;58:1044–52. <https://doi.org/10.1007/s11427-012-4297-8>.
62. Boucher J. Research review: structural language in autistic spectrum disorder—characteristics and causes. *J Child Psychol Psychiatry.* 2012;53(3):219–33. <https://doi.org/10.1111/j.1469-7610.2011.02508.x>.
63. Davidson MM, Ellis WS. A discrepancy in comprehension and production in early language development in ASD: is it clinically relevant? *J Autism Dev Disord.* 2017;47:2163–75. <https://doi.org/10.1007/s10803-017-3135-z>.
64. Reznick JS, Goldfield BA. Rapid change in lexical development in comprehension and production. *Dev Psychol.* 1992;28(3):406. <https://doi.org/10.1037/0012-1649.28.3.406>.
65. Fernald A, Swingle D, Pinto JP. When half a word is enough: infants can recognize spoken words using partial phonetic information. *Child Dev.* 2001;72(4):1003–15. <https://doi.org/10.1111/1467-8624.00331>.
66. Maris E, Oostenveld R. Nonparametric statistical testing of EEG-and MEG-data. *J Neurosci Methods.* 2007;164(1):177–90. <https://doi.org/10.1016/j.jneumeth.2007.03.024>.
67. Peter MS, Durrant S, Jessop A, Bidgood A, Pine JM, Rowland CF. Does speed of processing or vocabulary size predict later language growth in toddlers? *Cogn Psychol.* 2019;115:101238.
68. Hartley C, Bird LA, Monaghan P. Comparing cross-situational word learning, retention, and generalisation in children with autism and typical development. *Cognition.* 2020;200:104265. <https://doi.org/10.1016/j.cognition.2020.104265>.
69. He AX, Luyster RJ, Arunachalam S. Parental tuning of language input to autistic and nonspectrum children. *Front Psychol.* 2022;13:5945.
70. Robins DL, Fein D, Barton M. Modified checklist for autism in toddlers, revised, with follow-up (M-CHAT-R/F) TM. LineageN. 2009.
71. Lord C, Pethick S, Renda C, Cox JL, Dale PS, Reznick JS. Short-form versions of the MacArthur communicative development inventories. *Appl Psycholinguist.* 2000;21(1):95–116. <https://doi.org/10.1017/S0142716400001053>.
72. Zimmerman IL, Steiner VG, Pond RE. *Preschool language scales—fifth edition (PLS-5)*. Bloomington: Pearson; 2011.
73. Mullen EM. *Mullen scales of early learning*. Circle Pines: American Guidance Service; 1995.
74. Konishi H, Stahl AE, Golinkoff RM, Hirsh-Pasek K. Individual differences in nonlinguistic event categorization predict later motion verb comprehension. *J Exp Child Psychol.* 2016;151:18–32. <https://doi.org/10.1016/j.jecp.2016.03.012>.
75. Macroy-Higgins M, Montemarano EA. Attention and word learning in toddlers who are late talkers. *J Child Lang.* 2016;43(5):1020–37. <https://doi.org/10.1017/S0305000915000379>.
76. Bates D, Maechler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *J Stat Softw.* 2015;67(1):1–48. <https://doi.org/10.18637/jss.v067.i01>.
77. R Core Team. *A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing; 2016. (<https://www.R-project.org/>).
78. Dink JW, Ferguson B. *eyetrackingR: an R library for eye-tracking data analysis*. 2015. Retrieved from <http://www.eyetracking.com>
79. Rocchi L, Di Santo A, Brown K, Ibáñez J, Casula E, Rawji V, Di Lazzaro V, Koch G, Rothwell J. Disentangling EEG responses to TMS due to cortical and peripheral activations. *Brain Stimul.* 2021;14(1):4–18. <https://doi.org/10.1016/j.brs.2020.10.011>.
80. Haith MM, Wentworth N, Canfield RL. The formation of expectations in early infancy. *Adv Infancy Res.* 1993;8:251–97.
81. Su Y, Naigles LR. online processing of subject–verb–object order in a diverse sample of Mandarin-exposed preschool children with autism spectrum disorder. *Autism Res.* 2019;12(12):1829–44. <https://doi.org/10.1002/aur.2190>.
82. Pijnacker J, Davids N, van Weerdenburg M, Verhoeven L, Knoors H, van Alphen P. Semantic processing of sentences in preschoolers with specific language impairment: evidence from the N400 effect. *J Speech Lang Hear Res.* 2017;60(3):627–39. https://doi.org/10.1044/2016_JSLHR-L-15-0299.
83. Yu C, Smith LB. What you learn is what you see: using eye movements to study infant cross-situational word learning. *Dev Sci.* 2011;14(2):165–80. <https://doi.org/10.1111/j.1467-7687.2010.00958.x>.
84. Fernald A, Pinto JP, Swingle D, Weinberg A, McRoberts GW. Rapid gains in speed of verbal processing by infants in the 2nd year. *Psychol Sci.* 1998;9(3):228–31. <https://doi.org/10.1111/1467-9280.00044>.
85. Durrant S, Jessop A, Chang F, Bidgood A, Peter MS, Pine JM, Rowland CF. Does the understanding of complex dynamic events at 10 months predict vocabulary development? *Lang Cogn.* 2021;13(1):66–98. <https://doi.org/10.1017/langcog.2020.26>.
86. Lord C, Rutter M, DiLavore P, Risi S, Gotham K, Bishop S. *Autism diagnostic observation schedule—2nd edition (ADOS-2)*. Los Angeles: Western Psychological Corporation; 2012.
87. Luyster R, Gotham K, Guthrie W, Coffing M, Petrak R, Pierce K, Bishop S, Esler A, Hus V, Oti R, Richler J. The autism diagnostic observation schedule—toddler module: a new module of a standardized diagnostic measure for autism spectrum disorders. *J Autism Dev Disord.* 2009;39:1305–20. <https://doi.org/10.1007/s10803-009-0746-z>.
88. Bishop SL, Guthrie W, Coffing M, Lord C. Convergent validity of the Mullen scales of early learning and the differential ability scales in children with autism spectrum disorders. *Am J Intellect Dev Disabil.* 2011;116(5):331–43. <https://doi.org/10.1352/1944-7558-116.5.331>.
89. Bishop SL, Farmer C, Thurm A. Measurement of nonverbal IQ in autism spectrum disorder: scores in young adulthood compared to early childhood. *J Autism Dev Disord.* 2015;45:966–74. <https://doi.org/10.1007/s10803-014-2250-3>.
90. Brooks PJ, Maouene J, Sailor K, Seiger-Gardner L. Modeling the semantic networks of school-age children with specific language impairment and their typical peers. In: *Proceedings of the 41st Annual Boston University Conference on language development*. Cascadilla Press. 2017;114–127.
91. Sheng L, McGregor KK. Lexical–semantic organization in children with specific language impairment. *J Speech Lang Hear Res.* 2010;51(1):146–59. [https://doi.org/10.1044/1092-4388\(2009\)08-0160](https://doi.org/10.1044/1092-4388(2009)08-0160).
92. He AX, Kon M, Arunachalam S. Linguistic context in verb learning: less is sometimes more. *Lang Learn Dev.* 2020;16(1):22–42. <https://doi.org/10.1080/15475441.2019.1676751>.
93. Kandel M, Snedeker J. Assessing two methods of webcam-based eye-tracking for child language research. Preprint ahead of publication.
94. Naigles LR, Tovar AT. Portable intermodal preferential looking (IPL): investigating language comprehension in typically developing toddlers and young children with autism. *JoVE (Journal of Visualized Experiments)*. 2012;70:e4331.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.