Quality of remotely-collected gaze data in autistic and nonspectrum children

Rhiannon Luyster
Emerson College, Boston, MA

Taylor Boyd
New York University, New York, NY

Amelia Steele
Emerson College, Boston, MA

Thuy Buonocore
Emerson College, Boston, MA

Catherine Sancimino
Icahn School of Medicine at Mount Sinai, New York, NY

Sudha Arunachalam
New York University, New York, NY

Abstract: Many developmentalists have shifted to remote research. This project uses secondary data to evaluate the quality of eye-gaze data from 30 autistic children and a language-matched sample of 30 nonspectrum children (mean ages 48 and 27 months, respectively). All children completed an experimenter-moderated preferential looking paradigm via Zoom. Frequency of co-occurring child and household events, rates of missing data, and percent agreement between gaze coders were assessed. Results indicated that co-occurring events were minimal, with no diagnostic group differences. Missing data rates were low overall and were unrelated to diagnostic group, age, or language level of participants; however, higher rates of co-occurring child behaviors were associated with higher rates of missing data. Agreement between coders for eye gaze data was comparable to in-lab studies. Results affirm the usefulness of remote, experimenter-moderated gaze-based research with autistic and nonspectrum children.

Keywords: autism; preferential looking; online research; methodology

Corresponding author(s): Rhiannon J. Luyster, Communication Sciences and Disorders, Emerson College, 120 Boylston St., Boston, MA 02216. Email: rhiannon_luyster@emerson.edu

ORCID ID(s): Rhiannon Luyster: https://orcid.org/0000-0001-8311-4772; Taylor Boyd: https://orcid.org/0009-0004-3123-7550; Amelia Steele: https://orcid.org/0009-0003-1033-9841; Thuy Buonocore: https://orcid.org/0009-0007-3060-7048; Catherine Sancimino: https://orcid.org/0000-0003-3151-7734; Sudha Arunachalam: https://orcid.org/0000-0003-4394-3626

Citation: Luyster, R., Boyd, T., Steele, A., Sancimino, C., Buonocore, T., & Arunachalam, S. (2025). Quality of remotely-collected gaze data in autistic and nonspectrum children. *Language Development Research, 5*(3), 131–154. http://doi.org/10.34842/ldr2025-633

Introduction

With the COVID-19 pandemic, researchers devised new strategies for pursuing their work, including remote data collection via videoconferencing platforms (Tsuji et al., 2022). This approach appeared to yield similar results to in-person paradigms (Bánki et al., 2022; Chuey et al., 2021; Steffan et al., 2023) and offered unanticipated enrollment benefits for sample diversity (Shields et al., 2021, Ozernov-Palchik et al., 2022), making it likely to persist. These new opportunities are exciting but also pose potential challenges. When participants are at home, the experimenter has less control over the environment (Gijbels et al., 2021), which could lead to poorer quality data. In the current study, we used secondary data analysis to examine the quality of eye-gaze data collected remotely with young children on the autism spectrum and with non-spectrum children. Both groups completed an experimenter-moderated language learning task using a variant of the preferential looking paradigm (Golinkoff et al., 1987) on Zoom.[1]

The measure of interest in the current study is children's eye gaze as they looked at their video screen and heard an auditory prompt directing their attention to a particular image. Gaze was recorded using a webcam and later coded offline by trained coders. This paradigm, sometimes referred to as "intermodal preferential looking" or "looking while listening" (e.g., Fernald et al., 2008; Golinkoff et al., 1987) has been successfully used with autistic children in lab settings (e.g., Bebko et al, 2006; Ellis Weismer et al., 2016; Horvath et al., 2018; Venker et al., 2013) and in the home with experimenters bringing a portable setup (e.g., Goodwin et al., 2012; Naigles & Tovar, 2012; Swensen et al., 2007).

The central construct of this investigation—the quality of eye-gaze data collected remotely—requires consideration of different metrics of "eye-gaze data quality". In terms of eye-gaze quality, one important metric is missing data; that is, those moments when direction of gaze cannot be determined or when the child is looking off-screen. Missing data are inevitable, because blinking results in missing data. However, it can also occur because, for example, child participants may lean forward to look more closely at the screen, leaving their eyes outside the camera's range, or they may turn their heads to look at a caregiver. Some of these behaviors may be influenced by setting (i.e., lab-based vs. remote home-based) and diagnosis. For instance, Lapidow and colleagues (2021) noted that caregivers were more inclined to interact

[1] The terms autism, autism spectrum and autism spectrum disorder (ASD) will be used interchangeably. Moreover, in light of recent dialogue (e.g., Botha et al., 2021) around diverse preferences for person-first versus identity-first language, the terms "on the autism spectrum" and "autistic" will both be used to refer to individuals with a confirmed diagnosis of ASD per the DSM-5 (APA, 2013). Finally, rather than referring to the comparison sample as "typically developing," we will use the term "nonspectrum".

with their children during online (vs. lab-based) administrations. Somewhat surprisingly, then, collecting remote rather than lab-based data from children in manually coded gaze-tracking paradigms has not consistently been shown to substantially influence rates of missing data, at least for nonspectrum children (e.g., Scott & Schulz, 2017). For example, Morini and Blair (2021) enrolled nonspectrum preschoolers and reported that the number of analyzable trials was comparable across face-to-face and virtual settings (ranging from a mean difference of .1 to 1.6 trials across ages and trial types). Similarly, Bacon and colleagues (2021) used a looking-while-listening virtual platform with nonspectrum toddlers; they reported that data integrity was robust against internet quality and that the percentage of includable trials (88%) was comparable to previous lab-based rates (e.g., 66% to 78% in Venker et al., 2020).

We might expect that missing data might be more common in remote paradigms for autistic (vs. nonspectrum) children, however. Consider the fact, for instance, that missing data can result from movement, and autistic children may be particularly prone to movement-related data loss (e.g., Venker et al., 2020). Moreover, given suggestions that autistic children may, on one hand, find gaze-tracking paradigms particularly challenging due to the need to remain relatively still (Venker & Kover, 2015) but, on the other hand, may participate more easily in the predictable environment of a home-based study (Gijbels et al., 2021), it is particularly important to see if remotely collected data quality differs for autistic and nonspectrum children. Most previous studies with autistic children using preferential looking paradigms in the home have had experimenters physically present with the child during the study; this allowed study staff to ensure a consistent setup, monitor the environment for distractions, and support the child and caregiver in following directions (e.g., Jyotishi et al., 2017; Potrzeba et al., 2015; Tovar et al., 2015; but see Arunachalam et al., 2024 for a recent example of a study with fully remote task administration). With the fully remote task administration required during the pandemic, the experimenter has less knowledge of what is occurring in the home environment, including the details of the setup as well as what unrelated stimuli might be co-present. Thus, a new look at data quality with this population is warranted; in this study, we explore the rate of missing data in our remote paradigm, as well as whether this differs by diagnostic group (autistic vs. nonspectrum).

Another important consideration in evaluating the quality of eye-gaze data is interrater reliability. Manual coding of gaze from video generally yields lower track-loss rates compared to automatic eye-tracking, including for children on the autism spectrum (Haviland et al., 1996; Venker & Kover, 2015; Venker et al., 2020). When using manual coding, it is important to have multiple coders and to quantify their agreement (e.g., Fernald et al., 2008). In the home setting, where we have limited control over the precise visual angle between the child and the screen, as well as (in the current study) over the exact dimensions of the screen being used, it is likely that coders will have more difficulty determining whether a child is looking, for example, to the

right side of the screen or to an object to the right of the device. Prior findings suggest that remote data may present specific difficulties for agreement: Morini and Blair (2021) reported that inter-rater reliability was lower for remote data vs. a face-to-face setting. We therefore expect that remote data collection may result in lower agreement than lab-based paradigms or in-home studies where the experimenter is present and brings their own setup. Inter-rater reliability can be quantified by measuring the percent of frames on which coders agree on a particular code for direction of gaze, as well as by Cohen's kappa; Cohen (1960) suggested that kappa values of .81 or higher indicate excellent agreement. In our lab's training process, we require coders to achieve a kappa of >.9 with the training standard before they can code independently. Nevertheless, because percent agreement is more commonly reported in studies using this paradigm, we report here on percent agreement.

Therefore, in this study we use secondary data analysis to examine the quality of manually coded gaze data gathered from autistic and nonspectrum preschool-aged children via a remote platform by reporting on (1) missing data and (2) percent agreement among gaze coders. Additionally, we reviewed the videos to determine the frequency of co-occurring events that might affect the quality of the gaze data and asked whether these were associated with missing data or percent agreement among coders.

Method

All recruitment and testing procedures were approved by the Biomedical Research Alliance of New York (BRANY), which provides IRB services for multi-site studies.

Participants

Participants contributing data to this secondary data analysis are a subset of those in a larger study. A US national sample of families was recruited for a remote study focusing on language learning in children on the autism spectrum. Families of autistic children were recruited through online advertisements, a specialized recruitment service, and the SPARK national autism research registry (Feliciano et al., 2018). Families of nonspectrum children were recruited through online advertisements, parent organization emails, and our own research participant databases.

Children on the autism spectrum were eligible to participate in the larger study if they were 36.0 to 71.9 months old, had a previous medical or educational diagnosis of autism spectrum disorder (ASD), and scored 12 points or higher on the Social Communication Questionnaire (SCQ; Rutter et al., 2003), originally published as the Autism Screening Questionnaire (ASQ, Berument et al., 1999). Nonspectrum children were eligible if they were aged 24 to 48 months (younger than the autistic group in order to match groups on language, see below), had no previous diagnoses that would affect language or cognition, had no immediate family members diagnosed with autism,

and if they scored less than 12 points on the SCQ. Across both groups, children were not eligible to participate if their caregiver reported they (a) were born before 37 weeks of pregnancy, (b) had uncorrected vision or hearing impairments, (c) were colorblind or (d) heard English less than 70% of the time. With regards to the latter requirement, parents reported on their child's language exposure. The 70% cutoff is based on Cattani et al. (2014), who found that bilingual preschoolers perform as well as monolingual English-learning preschoolers on standardized language assessments if they have at least 60% exposure to English; here, a stricter criterion of 70% is used because the word learning tasks tap into processing abilities that go beyond the offline performance measured in standardized assessments.

Participant diagnostic status was confirmed using a multi-step process that was developed to be suitable for remote data collection. As mentioned above, SCQ scores were used as a screening tool for eligibility. SCQ validity studies indicate an optimum cutoff score of 15 for children 4 years and older (Berument et al., 1999; Rutter et al., 2003), however, subsequent research has identified a lower cutoff score of ≥12 to yield best sensitivity and specificity for children younger than age 4 years (Allen et al., 2007; Corsello et al., 2007; Wiggins et al., 2007). Thus, because of the younger age of many of the children in the present sample, the current study utilized a cutoff score of 12, requiring that children in the autistic group score 12 or higher and that nonspectrum children score below 12. This SCQ cutoff score also allowed the researchers to cast a wider net for recruitment of autistic children (given that we also had a licensed clinical psychologist, the fourth author, confirm diagnosis using all available data, as noted below).

Next, we gathered caregiver report information about intervention and diagnostic history using a questionnaire, including services provided in the school and the community (see data on OSF). Caregivers reported on their child's current autism-related symptoms using the Gilliam Autism Rating Scale–Third Edition (Gilliam, 2014). Caregivers also completed the Vineland Adaptive Behavior Scales, Third Edition (Vineland-3; Sparrow et al., 2016) to provide information about adaptive functioning skills; because autism is commonly associated with impairments in adaptive functioning, results of the Vineland-3 were reviewed to determine whether the Communication, Daily Living, and Socialization scales were consistent with autism. Finally, caregivers and children completed a 15-minute guided, semi-structured and video recorded interaction based on an adaptation of the Childhood Autism Rating Scale–Second edition (CARS-2; Schopler et al., 2010). Previous research published near the onset of the COVID-19 global pandemic demonstrated that the CARS-2 can be effectively adapted to a brief observation entitled "CARS-2-obs", with the examiner providing prompts to the caregiver while observing their interaction to identify child behaviors indicative of autism (Sanchez & Constantino, 2020). A licensed psychologist with advanced training and expertise in autism diagnosis (spanning research and clinical settings) reviewed all available clinical materials to confirm diagnostic status. Nine participants

from the larger study (out of 156) were recruited for the autistic group but did not receive diagnostic confirmation of autism based on clinical judgment after review of all available data.

Participant Matching

For the present analyses, we identified two subgroups of children (drawn from the larger study) who were matched by gender and language ability. To ensure comparability with similar in-lab studies, we selected a sample size based on previous research, which typically used groups of 30 or fewer participants per group (e.g., Goodwin et al., 2012; Venker, 2019; Venker et al., 2013). Our two goals in identifying subgroups for the present analyses were to include children with a wide range of language abilities, given a previous finding that children with lower language abilities were more likely to look away from the screen in a similar paradigm (Bebko et al., 2006), and within that, to match children on gender and expressive vocabulary.

Therefore, we first aimed to identify—from the 147 participants with confirmed autism diagnoses in the larger study—a subset of 30 children on the autism spectrum to include for the present analyses. To ensure a wide range of language abilities, we first binned all children in the larger study based on total number of words produced on the MacArthur Bates Communicative Development Inventory (MCDI) Words and Sentences Long Form (which contains 680 vocabulary words; Fenson et al., 2006). Six bins were used: fewer than 100 words, 100-199 words, 200-299 words, 300-399 words, 400-499 words and 500 or more. From each bin, we selected at least one child (when more than one child was available, we used random selection) while maintaining the same approximate MCDI distribution in the subsample that we had in the larger sample. This process yielded a subgroup of 30 children on the autism spectrum (18 males, 12 females) aged 36 to 67 months (M = 48.73, SD = 8.93), all of whom had their diagnosis confirmed by the licensed psychologist according to the process outlined earlier.

Next, we matched each autistic participant to a nonspectrum participant from the larger study based on gender and vocabulary scores from the MCDI (within +/- 20 words). Although +/- 20 on the MCDI is a relatively large spread, it has been previously used for language-based matching (e.g., Naigles et al., 2016) and allows the inclusion of children with lower vocabulary scores. A stricter matching protocol would have excluded autistic children with lower vocabulary scores due to difficulty in finding exact matches. When more than one nonspectrum child who was a vocabulary- and gender-match was available, a single one was selected at random. This process was successful for all autistic participants except for two, for whom a vocabulary matched participant within 20 words could not be identified; we therefore selected the closest available same-gender nonspectrum match for these children (one pair was matched within 21 words and the other was matched within 44 words; see similar approach in Luyster & Lord, 2009). This matching process resulted in our second subgroup,

comprising 30 nonspectrum children (18 males, 12 females) aged 24 to 34 months (M = 27.37, SD = 3.10).

As is common when using language-matched sampling for young autistic and nonautistic children (see Charman, 2004 for a discussion), the autistic and nonspectrum groups differed significantly on age (t = 12.38, p < .001) and SCQ (t = 12.31, p < .001) but not on MCDI (t = .094, p = .93). See Table 1. Caregivers reported on children's race and ethnicity using NIH categories as follows: 3 Black/African-American, 3 Asian, 43 White, 3 More than one race, 2 Prefer not to answer; 8 Hispanic or Latine, 52 Not Hispanic or Latine. We did not explicitly ask for information pertaining to socioeconomic status, but we note that to participate, families had to have a sufficiently strong internet connection to engage in a Zoom call and watch streaming videos on an appropriate device.[2]

On the day of the study, the caregiver and child logged onto a Zoom meeting with the experimenter. Children sat in front of a desktop, laptop, or tablet with a screen at least 5.5 inches by 8.5 inches. The child usually sat independently, but some children sat in their caregiver's lap. Parents were coached on an appropriate distance to have the child sit from the screen, but we did not require them to measure it. We discouraged parents from having the child hold anything during the study, but if the parent believed the child would be better able to sit still and participate while holding a toy or eating we allowed it. The experimenter conducted a warm-up, followed by a word-learning experiment (described below), and then a 15-minute guided play-based observation between the child and caregiver. The session lasted about 45 minutes and was recorded using Zoom. For this paper, we only report details of the procedure relevant to the current analyses of eye-gaze data quality during the word-learning experiment.

Word-Learning Experiment

The experiment comprised two word-learning trials on which the child was introduced to a new word (e.g., "modi") and then tested on the novel word's meaning. For the current paper, we focus on the "test phase" of each of the two trials, which were structured identically. Test phases consisted of 8-second videos. The 8-second video comprised three phases: Baseline (3 seconds), during which children viewed the images on the screen (as depicted in Figure 1) and heard a prompt designed to direct their attention to the screen ("Whoa, look!"), Query (2 seconds), during which the images disappeared, replaced by a large central fixation image, and children heard an auditory prompt to find the target (e.g., "Where's the modi?"), and Response (3 seconds) during which the images reappeared in the same locations as during Baseline

[2] From the larger study, there were 24 families who attended an orientation call but did not show up for the study visit.

and children heard an additional prompt (e.g., "Find the modi!"). We analyzed their gaze during Baseline and Response, as would typically be done in word learning experiments; Baseline provides a measure of children's a priori preference for the images and Response provides a measure of their preferences after being asked to find the referent. We did not analyze gaze during the Query phase for two reasons. First, this phase did not have images in the target locations, and second, in our experience, this phase is often when children are likely to look away from the screen (e.g., to share attention with their parent).

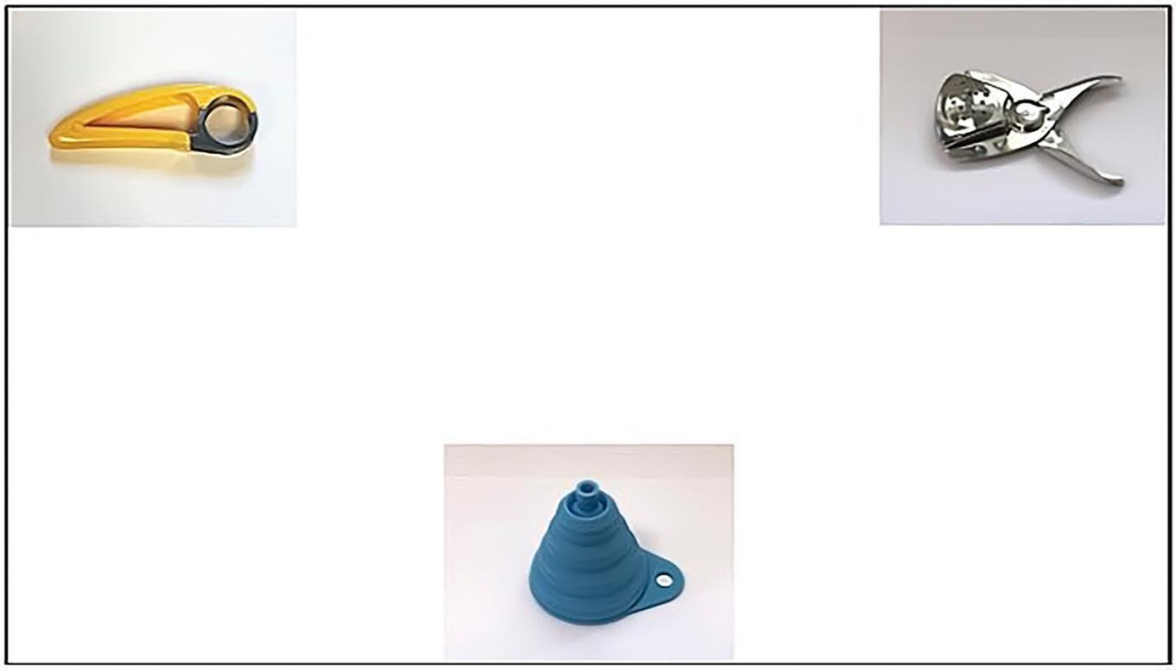

Figure 1. ***Example of visual stimuli during Baseline and Response phases***

Coding

We coded two types of variables: (1) co-occurring events, including both child behaviors and external household events, that might be expected to disrupt performance in an experimental task. (2) children's gaze behavior as the videos played, to assess both rates of missing data and inter-rater agreement.

Co-occurring events

We coded for child behaviors and external, household events (see Table 2 for definitions and guidelines used by coders). With respect to child behaviors, we coded for child vocalization or physically interacting with an object, and additionally, given Venker and Kover's (2015) suggestion that child behaviors associated with autism

might lower the quality of eye-gaze data, we also coded for repetitive child sensorimotor movements that are characteristic of autism, such as rocking or hand flapping (e.g., Rutter, Le Couteur, & Lord, 2003).

Table 2. ***Coding definitions for co-occurring events***

Event Category	Coding Definition
Child sensorimotor movements	Includes whole body, torso, or arm/hand/finger movements. For example: rocking, hand flapping, peering through hands/fingers. These movements must include *voluntary* repetitive movements. Nail biting, hair twisting, thumb sucking and the like are all excluded. Chewing and drinking will not be coded; oral motor movements (e.g., popping lips, sticking out tongue, sucking thumb) in isolation (i.e., in the absence of other sensorimotor movements as listed) will not be coded. If this co-occurs with another category, code both.
Child vocalizations	Nonword vocalizations (e.g., laughter, jargoning) or speech (e.g., talking to caregiver, repeating audio from experimental stimuli). Making noises while chewing and drinking or breathing will not be coded. Do not count yawning, grunting (unless communicative), lip popping, sighing, raspberries. If this co-occurs with another category, code both.
Child physical distractions	Resulted from something the child was doing. Active involvement of/with physical object or agent resulting from child's behavior; for instance, child playing with a toy in hand or touching a computer keyboard. If the child is holding something or has something in their lap but they are not actively involved with it (meaning, they are not playing with it, looking with it, moving it around etc.), do not code. If this co-occurs with another category, code both.
External physical distractions	Sudden appearance/interruption by agent/physical object that (1) enters the child's visual field (e.g., sibling running in front of child) or (2) makes physical contact with the child (e.g., cat jumping on child's lap). These are not due to the child's behavior and do not include ongoing physical contact from the parent, who may be holding the child during the session. If this co-occurs with another category, code both.

With respect to external, household events, we coded for intrusions (e.g., from caregivers, pets or siblings) that entered the child's field of view or made physical contact with the child; this final category was only observed for one child in the data set, and so we did not analyze it quantitatively. We initially intended to include auditory distractions such as caregiver vocalizations or external noises (e.g., phone ringing, baby crying) in this last category, but Zoom recordings varied in how much of this external noise was filtered out by the software, and so we could not reliably determine whether these noises were present in the home for all videos.

Coding was done in 1-second bins for each of the 3 seconds in the Baseline period and 3 seconds in the Response period; each second was binary coded (i.e., presence or absence) for each of the four categories of co-occurring events. Videos were played in Adobe Premiere Pro and codes were recorded on a spreadsheet. All videos were coded for co-occurring events by two research assistants; inter-rater reliability was calculated for a randomly selected 20% of the sample. Percent agreement between the two raters was high for all four event categories: child sensorimotor movements (100%), child vocalizations (97.2%), child physical distractions (100%), external physical distractions (100%). After calculating inter-rater reliability metrics, disagreements were resolved via consensus between the same two coders.

Gaze Coding

Using standard procedures (e.g., Fernald et al., 2008), three trained research assistants who were naive to diagnosis independently coded the direction of children's gaze on the screen (top left, top right, center) from video recordings of the test phase at a rate of 30 frames/second.[3] The coders had to be able to hear the audio to determine the onset of Baseline and Response phases, but they did not know which object was the intended target or where it was located on the screen, both of which were counterbalanced across participants. Each video was coded by two of the three coders, who viewed videos using Adobe Premiere Pro software and recorded gaze codes on a spreadsheet. (Note that while most studies involve multiple coders on only a subset of trials to check reliability, we enlisted two coders for coding gaze on all videos because we were specifically interested in evaluating inter-rater agreement.) Missing data consisted of frames on which the eyes were closed (blinks), the child was looking outside of the areas of interest (e.g., looking off-screen, turning to look at a caregiver), or the child's eyes were not visible to the coder (e.g., blurry video, child was out of frame). The proportion of codes for these events, out of all of the 90 coded frames

[3] As in our prior work using this remote data collection (Arunachalam et al., 2024), we coded the videos using audio cues indicating the start of each phase within each trial to minimize any lag that might have accumulated during video playback. None of the included videos had an accumulated lag of more than 2 frames (66 ms) using this approach.

during each of the Baseline and Response phases of each test video, was calculated. Percent agreement to determine inter-rater reliability between the two coders were calculated for each trial using the "irr" package (Gamer, Lemon, Fellows, & Singh, 2019) in R version 4.1.2 (R Core Team, 2020).

Results

Co-occurring events

As described above, co-occurring events were categorized as child sensorimotor movements, child vocalizations, child physical distractions, or external physical distractions (see Figure 2).

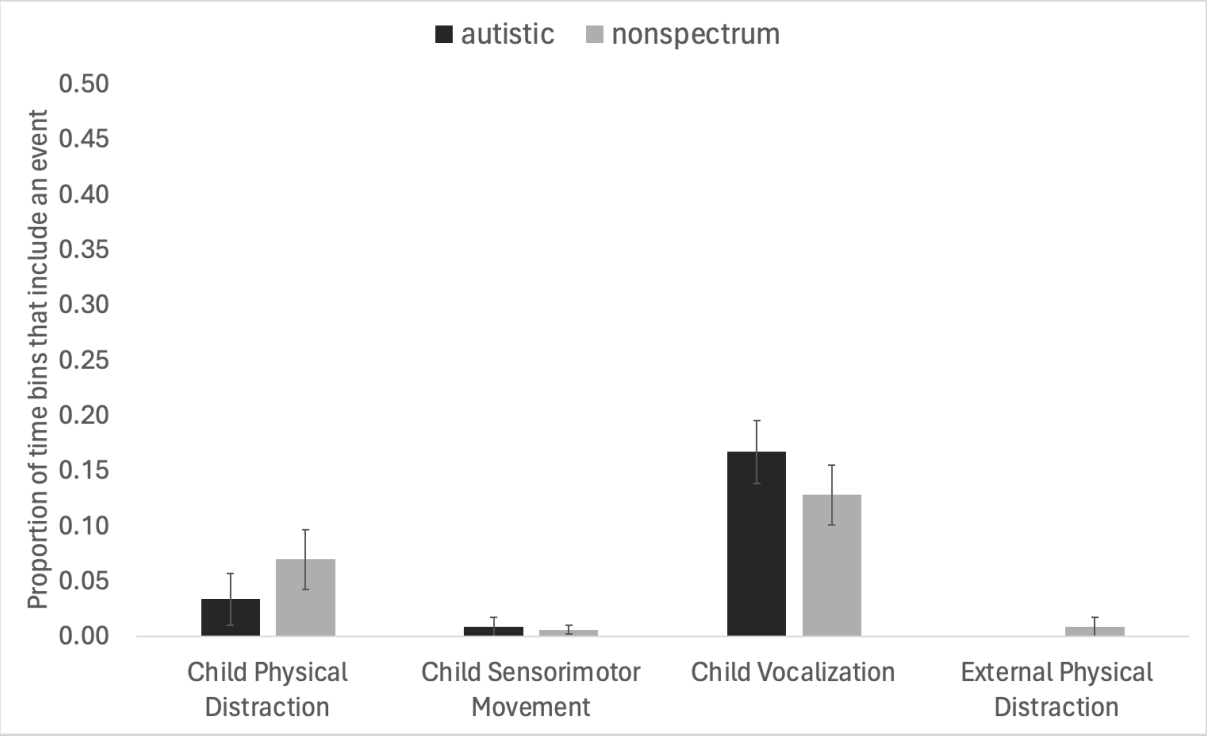

Figure 2. ***Proportion of 1-second time bins during which child behaviors or external physical distractions were observed across groups (autistic children, nonspectrum children) during the two coded phases of the trial (Baseline, Response).*** *Note:* The y-axis range is only depicted from 0 to 0.5 for readability. These events were coded across 6 seconds per trial (3 seconds of the Baseline phase, 3 seconds of the Response phase) over 60 trials per group (2 trials per child, 30 children per group).

Child vocalizations were the most common behaviors in both groups, followed by child physical distractions. Child sensorimotor movements were rare in both groups:

For 10 children in the autistic group and 9 children in the nonspectrum group, there were no child sensorimotor movements at all. As mentioned above, there was only one child who experienced an external event in the dataset, and so we did not analyze this category quantitatively. On half of the trials (n = 29 for the autistic group, n = 31 for the nonspectrum group), no events were coded at all. Given the small numbers of event occurrences, we used non-parametric two-tailed Mann-Whitney U tests for group differences; we found none (child sensorimotor movements: Mann-Whitney U = 450.5, $n_1 = n_2 = 30$, p = 1.00; child vocalizations: Mann-Whitney U = 530.5, $n_1 = n_2 = 30$, p = 0.22; child physical distractions: Mann-Whitney U = 406.0, $n_1 = n_2 = 30$, p = 0.25).

Missing Data

The mean proportion of frames with missing data during the Baseline phase was 0.097 (*SD* = 0.22, range 0-1) for the autistic group and 0.082 (*SD* = 0.20, range 0-1) for the nonspectrum group, and during the Response phase, it was 0.096 (*SD* = 0.20, range 0-1) for the autistic group and 0.059 (*SD*= 0.19, range 0-1) for the nonspectrum group. Missing data rates were severely right-skewed, and standard approaches to transform the data did not address this non-linearity; therefore, nonparametric approaches are more appropriate. We used quasibinomial nonparametric regression, or generalized additive models, with the "mgcv" package in R version 4.1.2 (R Core Team, 2020). (Note that the same results were obtained with parametric regression for all analyses, which are reported in our data repository.) We ran two generalized additive models, one for the Baseline phase and one for the Response phase, with missing data as the dependent variable and a fixed effect of diagnostic group (sum coded with the autistic group as +1 and the nonspectrum group as -1). For this and other regressions, we only report significant parameters of interest; full models are available at the OSF repository (https://osf.io/w9vmk/?view_only=dd1288d5ca014a0cbd0d472455c81c77). These analyses yielded no significant effects of diagnostic group (Baseline: $ß$ = 0.094, p = 0.70, deviance explained = 0.20%; Response: $ß$ = 0.26, p = 0.32, deviance explained = 1.49%).We then added age and MCDI scores (centered around their means); these were not highly correlated (R = 0.24) because of the heterogeneity of language abilities among the (chronologically older) autistic group. These analyses also yielded no significant effects for either the Baseline phase (Diagnostic group: $ß$ = -0.12, p = 0.82; Age: $ß$ = 0.019, p = 0.63; MCDI: $ß$ = -0.00094, p = 0.49; deviance explained = 0.85%) or the Response phase (Diagnostic group: $ß$ = 0.51, p = 0.31; Age: $ß$ = -0.026, p = 0.57; MCDI: $ß$ = -0.0016, p = 0.25; deviance explained = 5.66%), and we did not include these factors in subsequent missing data analyses.

To see if missing data rates were predicted by co-occurring child behaviors, we added to the simple models a fixed effect of the sum of the number of seconds (out of 3 seconds) during each phase (Baseline, Response) of each trial on which each of these behaviors were present (because 3 types of behaviors were measured during each of the 3 seconds, the range of values was 0-9), and its interaction with diagnostic group.

In the Baseline phase, this analysis yielded a significant simple effect of child behaviors ($ß = 0.67$, $p = 0.00014$), with a greater number of behaviors associated with higher rates of missing data, but no significant effect of group ($ß = 0.85$, $p = 0.17$) and no significant interaction ($ß = -0.51$, $p = 0.13$) (deviance explained = 18.3%). The same pattern obtained for the Response phase: a significant simple effect of child behaviors indicating that more behaviors was associated with higher rates of missing data ($ß = 0.45$, $p = 0.0072$), but no significant effect of group ($ß = 0.22$, $p = 0.73$) and no significant interaction ($ß = 0.28$, $p = 0.40$) (deviance explained = 10.4%).

Percent agreement among coders for gaze coding

We calculated percent agreement for gaze coding for each phase of each trial and participant separately and included those in similar analyses as for missing data. The mean percent agreement between gaze coders for the Baseline phase was 96.6% (*SD* = 8.1%) for the autistic group and 98.1% (*SD* = 8.1%) for the nonspectrum group; for the Response phase, it was slightly lower: 93.5% (*SD* = 16.0%) for the autistic group and 96.9% (*SD* = 8.3%) for the nonspectrum group. Percent agreement was left-skewed, so we used generalized additive models as above.

For both phases, this analysis did not yield a significant main effect of diagnostic group (Baseline: $ß = -0.61$, $p = 0.33$, deviance explained = 2.08%; Response: $ß = -0.76$, $p = .14$, deviance explained = 3.28%). We then added age and MCDI scores (centered around their means). For Baseline, this analysis yielded no significant main effects (group $ß = 0.38$, $p = 0.76$; age $ß = -0.044$, $p = 0.34$; MCDI $ß = 0.00046$, $p = 0.79$; deviance explained = 4.08%). For Response, there was still no significant main effect of group ($ß = 0.84$, $p = 0.37$), but there were significant effects of age ($ß = -0.074$, $p = 0.035$) and MCDI ($ß = 0.0036$, $p = 0.0042$) (deviance explained = 13.8%), indicating that there was higher agreement for younger children and children with higher MCDI scores.

Finally, we asked whether percent agreement was predicted by co-occurring child behaviors; we added to the simple model a fixed effect of the sum of the total number of seconds on each trial (out of 3 seconds) on which each of these behaviors was present for each trial (because three types of behaviors were measured during each of the 3 seconds, the range of values was 0-9), and its interaction with diagnostic group. This analysis yielded no significant simple effects and no significant interaction during either Baseline (behaviors $ß = 0.089$, $p = 0.80$; group $ß = -0.49$, $p = 0.48$; interaction $ß = -0.27$, $p = 0.71$; deviance explained = 2.43%) or Response (behaviors $ß = -0.099$, $p = 0.63$; group $ß = -0.78$, $p = 0.21$; interaction $ß = 0.032$, $p = 0.94$; deviance explained = 3.61%).

Discussion

The goal of this study was to explore the quality of remotely-collected eye-gaze data gathered from autistic and nonspectrum preschoolers. We quantified co-occurring events (both child and external/household) during brief Baseline and Responses phases of a word-learning task, and we tested the associations of co-occurring events with two common quality metrics (missing data and inter-rater reliability).

In our sample, sporadic co-occurring events—both child and external—were observed for many participants. As in lab-based settings, interruptions are inevitable during experimental sessions. However, these events were relatively infrequent and for half of the trials, these events did not occur at all. Moreover, children on the autism spectrum and nonspectrum children did not differ in rates of either child or external events. This is somewhat unexpected given that autistic children might be more prone to distraction and movement, and that the sensorimotor movements we coded for are characteristic of autism (Venker & Kover, 2015). It suggests that in a home-based remote testing condition, autistic children are not more likely to experience these interruptions compared to their nonspectrum peers. Moreover, our tallies indicated that external/household distractions were extremely rare, occurring for only one child (and affecting roughly .01 of time bins for nonspectrum children). This finding does not support previous suggestions that remote research may be particularly vulnerable to family interruptions and child attrition (Lapidow et al. 2021; Steffan et al., 2023). In our study, we believe that the pre-visit orientation video call that we provided may have helped caregivers create a focused environment for their child (Gijbels et al., 2021). Overall, then, these results attest to the suitability of curated remote-testing conditions.

Next, we explored missing data. We found that rates of missing data were relatively low and did not differ for children on the autism spectrum and nonspectrum children; this is in contrast to previous findings that autistic children look away from stimuli significantly more often than nonspectrum children (e.g., Tenenbaum et al., 2017). Moreover, given that many studies with autistic children of this age apply a criterion of >50% missing data when deciding whether to exclude children (e.g., Horvath et al., 2018; Venker et al., 2013; Venker, 2019; Venker et al., 2020), our exclusion rate on this basis would be just 4%—which is comparable to those studies for which exclusion rates range from approximately 5% (Horvath et al., 2018) to 16% (Venker et al., 2020). For subsequent analyses, we included even those children with high rates of missing data; these children would typically be excluded from analyses, but our reports of data quality are contingent on understanding how missing data is related to co-occurring events. Indeed, there was a significant association between co-occurring events and missing data across both the Baseline and Response phases. Therefore, even though overall rates of co-occurring events were quite low across groups, the frequency with which they occurred was associated with data loss. There

was, however, no main or interaction effect of group, age or language level. This finding affirms the importance of the quality control measures that many researchers take in order to minimize distractions, whether in laboratory or remote settings. Our finding indicates that by minimizing both child-based or environmental artifacts, we can reduce data loss.

The percent agreement between coders who manually coded children's eye gaze (autistic = 94-97%; nonspectrum = 97-98%) was slightly lower than but similar to what has been reported for children on the spectrum or with other developmental conditions or language delays in lab-based settings or at-home studies in which the experimenter brings a portable setup: e.g., 98% in Venker et al. (2013) and Venker et al. (2021); 97% in Ellis Weismer et al. (2016); 93-99% for pre-term and full-term toddlers in Loi et al. (2017); 98% in Naigles et al. (2011). This was a somewhat surprising outcome for us given that we had substantially less control over factors that would influence coders' judgments, such as distance from the screen (which affects visual angle) and dimensions of the device's screen. Moreover, the agreement between raters was not detrimentally affected by child behaviors or household events.

We did find, however, that gaze coding agreement was higher for younger children and those with higher MCDI scores during the Response phase. In other words, raters were less reliable when coding the children who were older and/or had more pronounced developmental (or at least language) delays. This finding is particularly intriguing in light of the fact that—due to the language-matched nature of our sample—the autistic group was older than the nonautistic group. We are not certain of an explanation for this variability in reliability. Our results suggest that agreement was not related to child behaviors, so it is unlikely to be caused by differences in regulation or externalizing behaviors. An alternative explanation might be that—perhaps related to language delay—these children had less clearly defined gaze patterns when asked to identify an object, perhaps doing more exploratory scanning than directed gaze, leading to lower coder agreement. This interpretation is consistent with the fact that we found an effect of age and MCDI only in the Response phase, and not during the Baseline period. Although analyses addressing whether or not these children successfully learned the word, as measured by gaze during the Response phase, is beyond the scope of this paper, future analyses might help to support or disprove our hypothesized interpretation.

There are certainly benefits to remote research including increased familiarity (and perhaps comfort) of the home environment, reduced barriers for study visits, and broader inclusion of diverse samples, and our work here suggest that there are relatively few disruptions arising from co-occurring events in remote research designs. Nevertheless, there are other disadvantages of at-home studies that researchers should consider. One notable difference from lab-based studies is that in the lab, we can keep the surrounding environment free from material and visual distractions. In

the home, we did ask caregivers to try to identify a space without a lot of clutter, but there were likely other objects nearby that could have attracted children's attention. In coding eye gaze, coders may have inferred that the child's gaze was directed to one side of the screen when in fact it was directed to something just beside the screen. This lack of precision in coding is a disadvantage of at-home studies, although the relatively high percent agreement among coders somewhat mitigates this concern. Another potential challenge is the difficulty of verifying that the child is viewing the stimuli as intended (Tompkins, 2022). While we believe we substantially minimized this concern by (1) offering a pre-visit orientation video call, (2) having the experimenter present during data collection, (3) checking in frequently with the caregiver during transitions from one part of the procedure to the next and (4) asking caregivers to turn on the "do not disturb" function on their devices, it is certainly possible that for example, colors appeared differently than we intended or that distracting notifications popped up on participants' screens.

Several important limitations of our work should be noted. First, in recruiting for this study, we explained to caregivers that children would be asked to sit in front of their home computer for the duration of the task; families who agreed to enroll were likely self-selected based on the likelihood that their child could meet the study demands. Therefore, the children (both on and off the spectrum) enrolled do not necessarily exhibit the full range of developmental heterogeneity observed at these ages. Second, our word-learning task was conducted synchronously; before and after the task, the child was interacting directly with an experimenter. These results may not generalize to other types of experimental paradigms that are less interactive or are asynchronous/unattended (e.g., Scott & Schulz, 2017). Third, because Zoom software filtered out background noise, we were not able to assess the frequency of household auditory distractions such as a phone ringing or baby crying. Fourth, this study was limited to families who had access to computer/tablet with Zoom software and a stable internet connection. Although upwards of 90% of American families have access to these resources as of 2021 (US Census Bureau, 2024), it is certainly possible that the patterns reported here might differ in the remaining 10% of families and/or in families who feel distrust for technology use, particularly in a research context (Beaton et al., 2017). Finally, due to the COVID-19 pandemic and the cessation of in-person data collection in our labs, we did not compare these results to an in-person version of this same task. Instead, we drew inferences from published in-lab studies, and—in doing so—we also want to recognize some differences between our paradigm and those in-lab studies; for example, our task was modeled after an in-person study that presented three test objects, while eye-tracking tasks often present only two. We do not expect this choice to substantially affect our conclusions about overall quality of remotely collected data because our primary questions of interest in the current paper did not concern whether children looked at a target or distractor but rather how easy it was to assess whether and where they were looking. However, it may explain the slightly decreased percent agreement statistics as compared to in-lab studies because there were more

possible codes to choose from. Our study also included a relatively small number of trials; while word learning studies commonly present only one trial (e.g., Dautriche et al., 2014; Yuan et al., 2012) or 2-4 trials (e.g., Horvath et al., 2018; Gliga et al., 2012; Naigles et al., 2011), other studies using similar paradigms with familiar/known words often have many more trials, including some of the studies we cite above as reference points (e.g., Venker et al., 2013; Venker, 2019; Venker et al., 2020). Our results are most straightforwardly relevant for other studies with similar task demands and may not generalize to other paradigms.

In sum, our findings suggest that—for both autistic and nonspectrum children—the data gathered from a remote gaze-data paradigm are characterized by minimal missing data and adequate agreement between coders. Child and household factors were noted (and the former were more frequent than the latter), and although the quality of gaze data was reduced by co-occurring events during the session, these events were generally infrequent. In a broader sense, the current study allowed us to test whether autistic and nonspectrum children differ from each other in remote gaze-based studies, which fills a crucial gap missing from prior work. We conclude that experimenter-moderated remote data collection offers a promising alternative to lab-based settings for manually-coded gaze paradigms for both autistic and nonspectrum children.

References

Allen, C.W., Silove, N., Williams, K. et al. (2007). Validity of the Social Communication Questionnaire in assessing risk of autism in preschool children with developmental problems. *Journal of Autism and Developmental Disorders, 37*, 1272–1278. https://doi.org/10.1007/s10803-006-0279-7

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). https://doi.org/10.1176/appi.books.9780890425596

Arunachalam, S., Steele, A., Pelletier, T., & Luyster, R. (2024). Do focused interests support word learning? A study with autistic and nonautistic children. *Autism Research, 17*(5), 955-971. https://doi.org/10.1002/aur.3121

Bacon, D., Weaver, H., & Saffran, J. (2021). A framework for online experimenter-moderated looking-time studies assessing infants' linguistic knowledge. *Frontiers in Psychology*, 4078. https://doi.org/10.3389/fpsyg.2021.703839

Bánki, A., de Eccher, M., Falschlehner, L., Hoehl, S., & Markova, G. (2022). Comparing online webcam-and laboratory-based eye-tracking for the assessment of infants' audio-visual synchrony perception. *Frontiers in Psychology, 12*, 733933. https://doi.org/10.3389/fpsyg.2021.733933

Beaton, B., Perley, D., George, C., & O'Donnell, S. (2017). Engaging remote marginalized communities using appropriate online research methods. *The Sage Handbook of Online Research Methods*, 563-577.

Bebko, J. M., Weiss, J. A., Demark, J. L., & Gomez, P. (2006). Discrimination of temporal synchrony in intermodal events by children with autism and children with developmental disabilities without autism. *Journal of Child Psychology and Psychiatry, 47*(1), 88-98. https://doi.org/10.1111/j.1469-7610.2005.01443.x

Berument, S. K., Rutter, M., Lord, C., Pickles, A., & Bailey, A. (1999). Autism screening questionnaire: Diagnostic validity. *British Journal of Psychiatry, 175,* 444-51. https://doi.org/10.1192/bjp.175.5.444

Botha, M., Hanlon, J., & Williams, G. L. (2021). Does language matter? Identity-first versus person-first language use in autism research: A response to Vivanti. *Journal of Autism and Developmental Disorders, 53*(2), 870-878. https://doi.org/10.1007/s10803-020-04858-w

Cattani, A., Abbot-Smith, K., Farag, R., Krott, A., Arreckx, F., Dennis, I., & Floccia, C. (2014). How much exposure to English is necessary for a bilingual toddler to perform like a monolingual peer in language tests? *International journal of language & communication disorders, 49*(6), 649-671. https://doi.org/10.1111/1460-6984.12082

Charman, T. (2004). Matching preschool children with autism spectrum disorders and comparison children for language ability: Methodological challenges. *Journal of Autism and Developmental Disorders, 34,* 59-64. https://doi.org/10.1023/B:JADD.0000018075.77941.60

Chuey, A., Asaba, M., Bridgers, S., Carrillo, B., Dietz, G., Garcia, T., Leonard, J. A., Liu, S., Merrick, M., Radwan, S., Stegall, J., Velez, N., Woo, B., Wu, Y., Zhou, X. J., Frank, M. C., & Gweon, H. (2021). Moderated online data-collection for developmental research: Methods and replications. *Frontiers in Psychology, 12*. https://doi.org/10.3389/fpsyg.2021.734398

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Corsello, C., Hus, V., Pickles, A., Risi, S., Cook, E. H., Leventhal, B. L., & Lord, C. (2007). Between a ROC and a hard place: Decision making and making decisions about using the SCQ. *Journal of Child Psychology and Psychiatry, 48*(9), 932–940. https://doi.org/10.1111/j.1469-7610.2007.01762.x

[dataset] Arunachalam, S., & Luyster, R. J.; 2022; Quality of remotely-collected gaze data in children with and without autism; Open Science Framework https://osf.io/w9vmk/?view_only=dd1288d5ca014a0cbd0d472455c81c77

Dautriche, I., Cristia, A., Brusini, P., Yuan, S., Fisher, C., & Christophe, A. (2014). Toddlers default to canonical surface-to-meaning mapping when learning verbs. *Child Development, 85*(3), 1168-1180. https://doi.org/10.1111/cdev.12164

Ellis Weismer, S., Haebig, E., Edwards, J., Saffran, J., & Venker, C. E. (2016). Lexical processing in toddlers with ASD: Does weak central coherence play a role? *Journal of Autism and Developmental Disorders, 46*(12), 3755–3769. https://doi.org/10.1007/s10803-016-2926-y

Feliciano, P., Daniels, A. M., Snyder, L. G., Beaumont, A., Camba, A., Esler, A., ... & Brewster, S. J. (2018). SPARK: a US cohort of 50,000 families to accelerate autism research. *Neuron, 97*(3), 488-493. https://doi.org/10.1016/j.neuron.2018.01.015

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S., & Bates, E. (2006). MacArthur-Bates Communicative Development Inventories, Second edition. *PsycTESTS Dataset*. https://doi.org/10.1037/t11538-000

Fernald, A. E., Zangl, R., Portillo, A. L., & Marchman, V. A. (2008). Looking while listening. *Language Acquisition and Language Disorders*, 97–135. https://doi.org/10.1075/lald.44.06fer

Gamer, M., Lemon, J., & Fellows, I., & Singh, P. (2019). *Irr: Various Coefficients of Interrater Reliability and Agreement*. R package version 0.84.1. https://CRAN.R-project.org/package=irr

Gijbels, L., Cai, R., Donnelly, P. M., & Kuhl, P. K. (2021). Designing virtual, moderated studies of early childhood development. *Frontiers in Psychology, 12*. https://doi.org/10.3389/fpsyg.2021.740290

Gilliam, J. E. (2014). Gilliam autism rating scale–Third edition (GARS-3). *Austin, TX: Pro-Ed*.

Gliga, T., Elsabbagh, M., Hudry, K., Charman, T., Johnson, M. H., & BASIS team. (2012). Gaze following, gaze reading, and word learning in children at risk for autism. *Child Development, 83*(3), 926-938. https://doi.org/10.1111/j.1467-8624.2012.01750.x

Golinkoff, R. M., Hirsh-Pasek, K., Cauley, K. M., & Gordon, L. (1987). The eyes have it: Lexical and syntactic comprehension in a new paradigm. *Journal of Child*

Language, 14(1), 23-45. https://doi.org/10.1017/S030500090001271X

Goodwin, A., Fein, D., & Naigles, L. R. (2012). Comprehension of wh-questions precedes their production in typical development and autism spectrum disorders. *Autism Research, 5*(2), 109–123. https://doi.org/10.1002/aur.1220

Haviland, J. M., Walker-Andrews, A. S., Huffman, L. R., Toci, L., & Alton, K. (1996). Intermodal perception of emotional expressions by children with autism. *Journal of Developmental and Physical Disabilities, 8*(1), 77–88. https://doi.org/10.1007/bf02578441

Horvath, S., McDermott, E., Reilly, K., & Arunachalam, S. (2018). Acquisition of verb meaning from syntactic distribution in preschoolers with autism spectrum disorder. *Language, Speech, and Hearing Services in Schools, 49*(3S), 668–680. https://doi.org/10.1044/2018_lshss-stlt1-17-0126

Jyotishi, M., Fein, D., & Naigles, L. (2017). Investigating the grammatical and pragmatic origins of wh-questions in children with autism spectrum disorders. *Frontiers in Psychology, 8*, 319. https://doi.org/10.3389/fpsyg.2017.00319

Lapidow, E., Tandon, T., Goddu, M., & Walker, C. M. (2021). A tale of three platforms: Investigating preschoolers' second-order inferences using in-person, zoom, and Lookit methodologies. *Frontiers in Psychology, 12,* 731404. https://doi.org/10.3389/fpsyg.2021.731404

Loi, E. C., Marchman, V. A., Fernald, A., & Feldman, H. M. (2017). Using eye movements to assess language comprehension in toddlers born preterm and full term. *The Journal of Pediatrics, 180,* 124-129. https://doi.org/10.1016/j.jpeds.2016.10.004

Lord, C., Rutter, M., DiLavore, P. C., Risi, S., Gotham, K., & Bishop, S. (2012). Autism Diagnostic Observation Schedule, Second Edition. *Torrance, CA: Western Psychological Services.*

Luyster, R., & Lord, C. (2009). Word learning in children with autism spectrum disorders. *Developmental Psychology, 45*(6), 1774. https://doi.org/10.1037/a0016223

Morini, G., & Blair, M. (2021). Webcams, songs, and vocabulary learning: A comparison of in-person and remote data collection as a way of moving forward with child-language research. *Frontiers in Psychology, 12*. https://doi.org/10.3389/fpsyg.2021.702819

Naigles, L. R., Cheng, M., Rattanasone, N. X., Tek, S., Khetrapal, N., Fein, D., & Demuth, K. (2016). "You're telling me!" The prevalence and predictors of pronoun

reversals in children with autism spectrum disorders and typical development. *Research in Autism Spectrum Disorders*, *27*, 11-20. https://doi.org/10.1016/j.rasd.2016.03.008

Naigles, L. R., Kelty, E., Jaffery, R., & Fein, D. (2011). Abstractness and continuity in the syntactic development of young children with autism. *Autism Research, 4*(6), 422-437. https://doi.org/10.1002/aur.223

Naigles, L. R., & Tovar, A. T. (2012). Portable intermodal preferential looking (IPL): Investigating language comprehension in typically developing toddlers and young children with autism. *JoVE (Journal of Visualized Experiments)*, *70*, e4331. https://doi.org/10.3791/4331

Ozernov-Palchik, O., Olson, H. A., Arechiga, X. M., Kentala, H., Solorio-Fielder, J. L., Wang, K. L., ... & Gabrieli, J. D. (2022). Implementing remote developmental research: A case study of a randomized controlled trial language intervention during COVID-19. *Frontiers in Psychology*, *12*, 734375. https://doi.org/10.3389/fpsyg.2021.734375

Potrzeba, E. R., Fein, D., & Naigles, L. (2015). Investigating the shape bias in typically developing children and children with autism spectrum disorders. *Frontiers in Psychology*, *6*, 446. https://doi.org/10.3389/fpsyg.2015.00446

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org/

Rutter, M., Bailey, A., & Lord, C. (2003). The Social Communication Questionnaire (SCQ). *Torrance, CA: Western Psychological Services.*

Rutter, M., Le Couteur, A., & Lord, C. (2003). Autism Diagnostic Interview-Revised (*ADI-R*). *Torrance, CA: Western Psychological Services.*

Sanchez, M. J., & Constantino, J. N. (2020). Expediting clinician assessment in the diagnosis of autism spectrum disorder. *Developmental Medicine & Child Neurology*, *62*(7), 806–812. https://doi.org/10.1111/dmcn.14530

Schopler, E., Van Bourgondien, M., Wellman, G., & Love, S. (2010). The Childhood Autism Rating Scale 2nd edition (CARS-2). *Torrance, CA: Western Psychological Services.*

Scott, K., & Schulz, L. (2017). Lookit (part 1): A new online platform for developmental research. *Open Mind*, *1*(1), 4-14. https://doi.org/10.1162/OPMI_a_00002

Shields, M. M., McGinnis, M. N., & Selmeczy, D. (2021). Remote research methods: Considerations for work with children. *Frontiers in Psychology, 12,* 703706. https://doi.org/10.3389/fpsyg.2021.703706

Sparrow, S. S., Cicchetti, D. V., & Saulnier, C. A. (2016). Vineland Adaptive Behavior Scales, Third Edition (Vineland-3). San Antonio, TX: Pearson.

Steffan, A., Zimmer, L., Arias-Trejo, N., Bohn, M., Ben, R. D., Flores-Coronado, M. A., et al. (2023). Validation of an open source, remote web-based eye-tracking method (WebGazer) for research in early childhood. *PsyArXiv*. https://doi.org/10.1111/infa.12564

Swensen, L. D., Kelley, E., Fein, D., & Naigles, L. R. (2007). Processes of language acquisition in children with autism: Evidence from preferential looking. *Child Development, 78*(2), 542–557. https://doi.org/10.1111/j.1467-8624.2007.01022.x

Tenenbaum, E. J., Amso, D., Righi, G., & Sheinkopf, S. J. (2017). Attempting to "increase intake from the input": Attention and word learning in children with autism. *Journal of Autism and Developmental Disorders, 47*(6), 1791–1805. https://doi.org/10.1007/s10803-017-3098-0

Tompkins, D. (2022). What do our participants really see during unmoderated remote studies? *International Congress of Infant Studies (ICIS)*. https://infantstudies.org/what-do-our-participants-really-see-during-unmoderated-remote-studies/

Tovar, A. T., Fein, D., & Naigles, L. R. (2015). Grammatical aspect is a strength in the language comprehension of young children with autism spectrum disorder. *Journal of Speech, Language, and Hearing Research, 58*(2), 301-310. https://doi.org/10.1044/2014_JSLHR-L-13-0257

Tsuji, S., Amso, D., Cusack, R., Kirkham, N., & Oakes, L. M. (2022). Editorial: Empirical research at a distance: New Methods for Developmental Science. *Frontiers in Psychology, 13,* 1-5. https://doi.org/10.3389/fpsyg.2022.938995

US Census Bureau. (2024). *Computer and Internet Use in the United States: 2021*. United States Census Bureau. https://www.census.gov/newsroom/press-releases/2024/computer-internet-use-2021.html

Venker, C. E. (2019). Cross-situational and ostensive word learning in children with and without autism spectrum disorder. *Cognition, 183,* 181–191. https://doi.org/10.1016/j.cognition.2018.10.025

Venker, C. E., Eernisse, E. R., Saffran, J. R., & Weismer, S. E. (2013). Individual

differences in the real-time comprehension of children with ASD. *Autism Research, 6*(5), 417–432. https://doi.org/10.1002/aur.1304

Venker, C. E., & Kover, S. T. (2015). An open conversation on using eye-gaze methods in studies of neurodevelopmental disorders. *Journal of Speech, Language, and Hearing Research, 58*(6), 1719–1732. https://doi.org/10.1044/2015_jslhr-l-14-0304

Venker, C. E., Mathée, J., Neumann, D., Edwards, J., Saffran, J., & Ellis Weismer, S. (2021). Competing perceptual salience in a visual word recognition task differentially affects children with and without autism spectrum disorder. *Autism Research, 14*(6), 1147-1162. DOI: https://doi.org/10.1002/aur.2457

Venker, C. E., Pomper, R., Mahr, T., Edwards, J., Saffran, J., & Ellis Weismer, S. (2020). Comparing automatic eye tracking and manual gaze coding methods in young children with autism spectrum disorder. *Autism Research, 13*(2), 271–283. https://doi.org/10.1002/aur.2225

Wiggins, L. D., Bakeman, R., Adamson, L. B., & Robins, D. L. (2007). The utility of the Social Communication Questionnaire in screening for autism in children referred for early intervention. *Focus on Autism and Other Developmental Disabilities, 22*(1), 33-38. https://doi.org/10.1177/10883576070220010401

Yuan, S., Fisher, C., & Snedeker, J. (2012). Counting the nouns: Simple structural cues to verb meaning. *Child Development, 83*(4), 1382-1399. https://doi.org/10.1111/j.1467-8624.2012.01783.x

Data, Code and Materials Availability Statement

The data and code reported on here are available in Open Science Framework at: https://osf.io/w9vmk/?view_only=dd1288d5ca014a0cbd0d472455c81c77. The methods and analyses were not preregistered.

Ethics Statement

Ethics approval was obtained from the Biomedical Research Alliance of New York (BRANY). All participants gave informed written consent before taking part in the study.

Authorship and Contributorship Statement

Rhiannon Luyster and **Sudha Arunachalam** conceived of the study. **Taylor Boyd**, **Amelia Steele**, **Thuy Buonocore**, and **Catherine Sancimino** made substantial contributions to data collection, coding, analysis, and interpretation. **Rhiannon Luyster**

and **Sudha Arunachalam** drafted the manuscript. All authors contributed to and approved the final version.

Acknowledgements

This work was supported by the National Institutes of Health (NIH R01DC017131) to Sudha Arunachalam and Rhiannon Luyster. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We would also like to express our gratitude to María Cobo Nieto and Taina Hernandez McShane for their help with data collection and coding. We would also like to thank the many children and families who shared their time with us as part of this study, including all of the families at the participating Simons Simplex Collection (SSC) sites, as well as the principal investigators (A. Beaudet, R. Bernier, J. Constantino, E. Cook, E. Fombonne, D. Geschwind, R. Goin-Kochel, E. Hanson, D. Grice, A. Klin, D. Ledbetter, C. Lord, C. Martin, D. Martin, R. Maxim, J. Miles, O. Ousley, K. Pelphrey, B. Peterson, J. Piggot, C. Saulnier, M. State, W. Stone, J. Sutcliffe, C. Walsh, Z. Warren, E. Wijsman). We appreciate obtaining access to phenotypic data on SFARI Base. Approved researchers can obtain the SSC population dataset by applying at https://base.sfari.org. We are also grateful to our Autistic Advisory Board for their insights into the autistic experience.