Published in final edited form as:
Autism. 2016 July ; 20(5): 591–604. doi:10.1177/1362361315600146.

Comparing methods for assessing receptive language skills in minimally verbal children and adolescents with autism spectrum disorders

Daniela Plesa Skwerer[1], **Samantha E Jordan**[2], **Briana H Brukilacchio**[1], **Helen Tager-Flusberg**[1]

[1]Boston University, USA

[2]MGH Institute of Health Professions, USA

Abstract

This research addresses the challenges of assessing receptive language abilities in minimally verbal children with autism spectrum disorder by comparing several adapted measurement tools: a standardized direct assessment of receptive vocabulary (i.e. Peabody Picture Vocabulary Test-4); caregiver report measures including scores on the Vineland-II Communication domain and a vocabulary questionnaire consisting of a list of words ranging from simple, developmentally early, to more advanced words expected to be understood by at least some older children and adolescents; an eye-tracking test of word comprehension, using a word–image pair matching paradigm similar to that often used in studies of infant language acquisition; and a computerized assessment using a touch screen for directly measuring word comprehension with the same stimuli used in the eye-tracking experiment. Results of this multiple-method approach revealed significant heterogeneity in receptive language abilities across participants and across assessment methods. Our findings underscore the need to find individualized approaches for capturing the potential for language comprehension of minimally verbal children with autism spectrum disorder who remain otherwise untestable, using several types of assessment that may include methods based on eye-tracking or touch-screen responding.

Keywords

eye-tracking; minimally verbal; receptive language; standardized assessments

It is now well recognized that there is enormous heterogeneity within the autism spectrum disorder (ASD) population not only in core symptoms but also in co-morbid characteristics (American Psychiatric Association (APA), 2013). The most widely varying co-morbidities include IQ, which can range from profoundly disabled to superior levels of intelligence, and language, which varies from the complete absence of language to intact linguistic functions including vocabulary level and grammatical abilities (Tager-Flusberg et al., 2005). Co-morbid characteristics sometimes travel together, for example, low IQ is associated with

Corresponding author: Daniela Plesa Skwerer, Department of Psychological and Brain Sciences, Boston University, 100 Cummington Mall, Boston, MA 02215, USA. dplesas@bu.edu.

limited or absent language, more frequent atypical behaviors, and some medical conditions; however, they are recognized as independent features each of which impact diagnosis, classification, treatment approaches, and long-term prognosis (Doshi-Velez et al., 2014).

Recent reports have highlighted the fact that research over the past several decades has seriously neglected the minimally verbal end of the autism spectrum (Tager-Flusberg and Kasari, 2013). These children and adults with little or no spoken language, who comprise about 30% of the ASD population, have not been included in the majority of studies, so little is known about them. While it has been difficult to define this population, now referred to as "minimally verbal," it is generally agreed that the term covers school-aged children and older who have little to no spoken language that is used spontaneously for communication. Formal definitions of this group focus on expressive language only (Kasari et al., 2013). Some may be completely nonverbal with no spoken words; others may have a very limited vocabulary of up to a few dozen words or fixed phrases that are used communicatively. Some may have little spontaneous speech though they may echo the speech or songs heard from others. It is not known whether there are meaningful differences within this heterogeneity of spoken language skills among minimally verbal individuals (Tager-Flusberg and Kasari, 2013).

Along with variability in expressive abilities, there are also significant differences within the minimally verbal population in receptive language skills (Gernsbacher et al., 2005). Some may have relatively good comprehension that goes beyond single words; other seem not to understand more than a very limited number of words or phrases (Rapin et al., 2009). One major roadblock to investigations of receptive language skills in this population is the paucity of valid assessment tools (Kasari et al., 2013). Standardized tests may be especially challenging for a variety of reasons. These include difficulty engaging a child to attend; lack of understanding of the pragmatics of test situations; socially unresponsive to the examiner; distractions in the environment; unfamiliarity of the testing environment; absence of a pointing response; perseverative responding; frustrations with performance as test items get harder; anxiety about test performance (Tager-Flusberg, 1999). Any or all of these issues could affect the validity and reliability of the standardized test measure. Furthermore, given the need to establish a basal level of performance and the presence of floor effects for standard scores on most tests, often a minimally verbal individual is found to be untestable on even the most common standardized language tests (Kasari et al., 2013).

Despite these difficulties, it is important for the field to address the challenge of assessing receptive language abilities in minimally verbal children from the perspective of both advancing research and enhancing clinical practice. Without better methods for assessing receptive language, minimally verbal children will continue to be excluded from research studies and little progress will be made in broadening our knowledge about the full autism spectrum or in understanding the mechanisms that underlie their severe impairments. From a clinical perspective, knowing the limits of child's receptive language is important in guiding the choice and content of interventions that target the specific profile of that child. To begin to address this challenge, the study reported here was designed to compare a number of alternative approaches to evaluating receptive language in a group of minimally verbal children and adolescents with ASD with an emphasis on lexical comprehension.

Measures available for assessing receptive language range from traditional standardized tests to caregiver report questionnaires and interviews to on-line methods that rely on eye-tracking technology (Kasari et al., 2013; Tager-Flusberg and Kasari, 2013). Each of these methods has strengths and limitations, and in general, a multi-measure approach is advocated in the field of language assessment, especially for the ASD population (cf. Tager-Flusberg et al., 2009). For this study, we included measures drawn from each of these categories, some of which were adapted for the age range of the participants enrolled.

Among standardized tests that directly assess children's receptive language, the only one recommended by Kasari and her colleagues for use with minimally verbal children and adults was the Peabody Picture Vocabulary Test (PPVT; Dunn and Dunn, 2007). Their evaluation was based on the psychometric properties of the test and its coverage of a wide age range. The most widely used caregiver report measure of receptive vocabulary for typically developing infants and toddlers is the MacArthur–Bates Communicative Development Inventory (MCDI; Fenson et al., 2007). This questionnaire has been used in numerous studies of young children with ASD, and shown to be reliable for this population (Luyster et al., 2008), although Bruckner et al. (2007) found that some of the items were not appropriate for the ASD population. Caregivers can also provide useful information about their children's receptive (and other) language skills using the Vineland Adaptive Behavior Scales (Sparrow et al., 2005), which has the advantage of covering the full life span and going beyond just vocabulary knowledge. More recently, researchers have begun exploring alternative methods for directly assessing receptive language skills based on new technologies that do not require comprehension of test instructions or an overt motor response (Tager-Flusberg and Kasari, 2013). Among these new technologies, eye movements have been most extensively studied as a measure of language comprehension in typically developing infants and children (e.g. Yurovsky et al., 2013). Only a few language studies have explored the use of eye-gaze measures in children with ASD using automated eye-tracking technology (e.g. Bavin et al., 2014; Norbury et al., 2010; Venker et al., 2013). Venker et al. (2013) based their work on the seminal studies of infants by Fernald and Marchman (2012) using the "looking while listening" paradigm. In their study, older preschoolers with ASD heard words that were presented over loudspeakers as the children were looking at two pictures side-by-side on a screen. The key responses were speed and accuracy of the children's ability to switch and maintain their gaze to the matching picture. This study validated this measure of vocabulary comprehension for the 5-year olds in their study. Bavin et al. (2014) used a very similar eye-tracking task with children aged 5 to 7 years old. They found that severity of ASD was related to efficiency in lexical processing, which has implications for developing the ability to integrate auditory and contextual information. Thus, eye-gaze patterns appear to hold some promise for use as a measure of lexical comprehension; however, it is important to note that none of the earlier studies explored their use in older minimally verbal participants.

In order to identify the strengths and limitations of a wide range of methods for assessing receptive language in minimally verbal children and adolescents with ASD, we compared several adapted measurement tools. Given its popularity in research on ASD and other neurodevelopmental disorders (Kasari et al., 2013), we included the PPVT as our standardized direct assessment of receptive vocabulary. Caregiver report measures included the Vineland (Communication domain) and a vocabulary questionnaire that was a modified version of the MCDI. The questionnaire we developed expanded the list of words that caregivers were asked to report on to cover the broader range of vocabulary one might expect to be understood by at least some older children and adolescents. A subset of words that were included in the questionnaire, extending from simple to more advanced words, were incorporated into an eye-tracking test of word comprehension, using a paradigm similar to that used by Venker et al. (2013). This same set of words was adapted into a second method for directly measuring comprehension by presenting them to the participant via a computerized assessment using a touch screen.

Methods

Participants

Participants included 19 children and adolescents with a diagnosis of ASD (mean age = 12.5 years; range 5.75–21.1 years; 4 girls) and with very limited expressive language, as reported by caregivers. Only children older than 5 years were included because being minimally verbal implies the failure to develop fluent spoken language by school age. Participants were recruited from the greater Boston area via advertisements on the Autism Consortium website, at different autism-related events and via local caregiver advocacy groups. A study staff person contacted families who expressed interest in the research advertised, described the project in more detail, and obtained verbal consent to conduct a phone screening to determine eligibility for this study. During the screening interview, caregivers were asked to briefly describe how their child communicates, whether they use spoken language functionally, and whether they had received a clinical diagnosis of autism or ASD. Families of children and adolescents, who were described as “practically nonverbal” by their caregiver, were invited to participate in the study. Children and adolescents were included if they did not use phrase-speech spontaneously and meaningfully on a daily basis and/or produced fewer than about 30 words/phrases used communicatively (e.g. non-echoed use) as reported by caregivers. Participants were excluded if English was not the primary language spoken in the home, if they had been diagnosed with a known genetic disorder, or had a history of significant neurological/psychiatric illnesses. Research visits were scheduled as many times as necessary for the participant to complete the assessments, keeping a flexible schedule to allow time to manage any unexpected behavioral challenges that came up during testing.

The participants’ diagnoses were confirmed using the Autism Diagnostic Observation Schedule (ADOS-2; Lord et al., 2012) Module 1 or the Adapted ADOS (A-ADOS; Hus and Lord, in preparation). Table 1 presents the descriptive characteristics of the sample. All participants had normal or corrected-to-normal vision and none had any significant sensory or neurological impairment. Only participants from English-speaking homes were included in the study.

Not all participants provided reliable data on all assessments due to difficulties with the pragmatic requirements of the testing procedures (e.g. task compliance challenges, behavioral, and motivational difficulties), or due to recording problems, so the number of participants included in data analyses varied between 14 and 19 on each measure. The procedures used in this study were approved by the Boston University Institutional Review Board.

Measures

Diagnostic and behavioral assessments

ADOS.: To assess autism diagnostic status, the recently developed version of the Adapted ADOS (A-ADOS; Lord et al., 2012a, unpublished), modified for use with older individuals who do not have functional speech, was administered by trained examiners to 16 participants, and Module 1 of the ADOS-2 (Lord et al., 2012) was administered to the 3 participants younger than 10 years. Although the Adapted ADOS assessment is still in the validation phase and a diagnostic algorithm is not yet available, the items and coding system are similar enough to allow one to calculate a total score based on the validated ADOS-2 Module 1 algorithm.

Vineland.: Participants' adaptive functioning was assessed with Vineland Adaptive Behavior Scales-II, Caregiver Rating Scale form (Vineland-II; Sparrow et al., 2005), which was completed by the primary caregivers, usually the mother. Caregivers rated their child's adaptive behavior and related skills in the domains of Communication (Receptive, Expressive, Written communication skills), Daily Living Skills (Personal, Domestic, Community skills), and Socialization (Interpersonal Relationships, Play and Leisure Time, Coping Skills) from which domain, subdomain, and a total adaptive behavior scores were derived. Researchers have argued for employing Vineland special population norms as well as national norms when assessing children, adolescents, and adults with ASD who are minimally verbal (Carter et al., 1998).[1]

Ravens matrices.: Nonverbal cognitive ability was evaluated using the Raven Colored Progressive Matrices (Ravens; Raven et al., 1998), a test of nonverbal reasoning currently regarded as an appropriate IQ test for use with individuals with ASD. It is a measure of general and fluid intelligence, considered largely independent of verbal abilities (Dawson et al., 2007). The test consists of geometric analogy problems in which a matrix of geometric figures is presented with one entry missing, and the correct missing entry must be selected from a set of six choices. The Raven Colored Progressive Matrices test was developed to assess young children (5 to 10; 6 years) or individuals falling into lower IQ ranges (Raven et al., 1998). We modified the administration of the Ravens by using a magnet-board and magnet-based geometric figure pieces, to allow the participant to pick up and place the piece selected in the space indicated, as demonstrated by the experimenter in one or more teaching trials. This adaptation helped to minimize the need for verbal instructions. We added four teaching trials, two based on color matching and two based on fill-patterns to illustrate the type of solution expected. If needed, the experimenter used hand-over-hand to show the placement of the correct figure during a teaching trial. After participants "passed" two teaching trials without help, they were administered the test items.

[1] We present the Vineland-II results relative to both national and special population norms; we used the table of supplementary norm group percentile ranks corresponding to raw scores for ages 10 years and older—Autism Special Population: Mute, table available in Carter et al. (1998).

Language measures

Standardized test: PPVT-4.: The PPVT-4 (Dunn and Dunn, 2007) was administered to assess receptive lexical knowledge. Norms are available for children over the age of 2 years 6 months through adulthood. The administration of this test followed the conventional book format, and participants were asked to point to the correct picture labeled by the experimenter. To adapt the test for use with our minimally verbal participants, we increased the number of teaching trials (up to 6) and, if needed, the experimenter modeled a pointing response using a handover-hand procedure, on up to two teaching trials.

Caregiver Vocabulary Checklist.: To obtain a more detailed estimate of participants' single-word vocabulary comprehension outside the context of laboratory testing, we developed a checklist of words and asked the caregiver to check off which words were understood by the child. Caregivers reviewed a list of 275 words, which included all the words used as stimuli in the computer-based tasks (see below). The list consisted of nouns, verbs, and adjectives, selected from the MCDI, Words, and Gesture form (Fenson et al., 2007); the first six sets (up to age 7) of the PPVT-4; and additional words more appropriate for older children and adolescents. There was a 60% overlap with the MCDI (Words and Gestures) list and 30% overlap with words from the first six sets of the PPVT-4. The list of words was ordered alphabetically, and included a 3-point rating scale (Yes, Maybe, No) for "words understood," to provide an estimation of the caregiver's certainty about their child's comprehension of the listed words.

Vineland receptive communication.: The Vineland Communication domain provides information on receptive (and expressive) aspects of communication and includes items organized in order of developmentally appropriate behaviors or abilities. Because our participants were recruited based on their minimally verbal status, caregivers were asked to rate the items starting from the beginning, regardless of their child's age. This assessment goes beyond single-word knowledge, tapping a variety of communicative behaviors and providing meaningful data even for the participants found to be unscorable on other standardized tests of receptive language. The Vineland-II Caregiver Rating Form includes 20 items pertaining to receptive communication in the "Listening and Understanding" section, covering the age range 0 to 5 + years. The measure provides raw, scaled, and age-equivalent scores for this scale.

Experimental tasks.: The two computer-administered experimental tasks were designed to probe single-word comprehension of nouns, verbs, and adjectives (eye-tracking and touch-screen response). The basic design and the stimuli used in the two tasks were the same: two pictures appeared side by side on the screen at the beginning of each trial and were displayed for 2.5 s before an auditory word, which matched one of the two pictures (the target), was played over speakers. A total of 84 different target words (the same words for both tasks) were presented in 84 trials distributed over three developmentally ordered blocks each including 28 trials. The target words consisted of 54 nouns, 18 verbs, and 12 adjectives.

Stimuli and apparatus

Visual stimuli.: The visual stimuli were digital color photographs illustrating the target words. Images were selected from Internet image databases. To ensure that the images included in the task were easily recognizable examples of objects, actions, and qualities, we asked 16 adults to label a set of 200 images with one word, and to rate how prototypical each image was for the referent illustrated. The images with the highest agreement on the label used (i.e. the same word provided by all raters) and with the highest prototypicality ratings were selected for inclusion in the tasks. The selected images were informally matched for complexity and brightness when presented in pairs. Each of the pictures in the pair displayed was marked as an area of interest (AOI) measuring 397 × 397 pixels in size, subtending 14.5° of visual angle vertically and 12 horizontally when viewed at a distance of 60 cm. The paired pictures were presented against a black background, separated by a space subtending 5° visual angle. Figure 1 illustrates an example of image-pairings for each of the grammatical categories included nouns, verbs, and adjectives. To control for possible effects of image salience, within each block the images serving as target in one trial were used as foil in another trial, paired with a different target image. Within each block of 28 trials, image-pairs were presented in pseudo-random order with the following constraints: the target picture did not appear on the same side (left or right) more than three times in a row and target and foil images were counterbalanced for side presentation (left–right) across trials within each block.

Auditory stimuli.: Auditory stimuli consisted of digital voice recordings of the target words in one the following formats: “Look, [target word]!” or “See, [target word]!” played at 80 dB volume. These phrases were recorded by a female speaker with a standard American accent. The carrier phrase was 500 ± 30 ms in duration across all trials and the two formats were played in a quasi-random order, the same number of times within each block. The target-word audio stimuli were matched for duration, grammatical category, and number of syllables within blocks. The mean duration of the target words included in the first block was 540 ms, while the target words included in the second and third block of trials averaged 603 and 675 ms in duration, respectively (reflecting the fact that later acquired words tend to have more syllables).

Eye-tracking task.: Language-mediated eye movements were recorded using a Tobii T60 binocular eye-tracker run using Tobii Studio 2.0.3. (Tobii Technology AB, Danderyd, Sweden). This eye-tracking system is completely noninvasive, with no head-mounted apparatus and with the camera embedded in the flat panel computer display. The system tracked both eyes, to a rated accuracy of 0.5°, sampled at 60 Hz. It was calibrated for each participant using a 5-point calibration.

Stimuli were presented with E-prime Version 2.11 software (Psychology Software Tools, Inc., Sharpsburg, PA) on a 17-in monitor integrated with eye-tracker. Image-pairs were displayed for 5 s on a black screen and the target audio word was played 2.5 s after the visual stimulus onset. Trials were separated by a black screen showing a central cartoon character (*Thomas the train* and friends), lasting about 1 s, in order to capture attention and orient the participant’s gaze to the center of the screen before the display of the test image-pairs. Several brief cartoon movies lasting between 5 and 10 s and presented full-screen were randomly interspersed between test trials. This procedure was meant to interrupt the predictable succession of image-pairs and words and to help reorient the participants' fluctuating attention to the screen.

Participants were seated in an upright armchair, approximately 60 cm from the monitor, with eye-level approximately even with the center of the screen. The experiment was conducted in a dimly lit room to allow for optimal functioning of the eye-tracking equipment. The experimenter monitored data collection on a separate computer, using the Live-viewer mode to observe the participants' looking behavior in real time. If the participant's gaze was not displayed on the screen at all, the experimenter could stop the experiment between trials and try to reorient the participant toward the screen. Thus, although trials were always 5 s in duration, the inter-trial interval was not always 1 s long, but varied under the control of the experimenter, in an effort to minimize eye-movement data loss.

Touch-screen task.: The touch-screen task was always administered after the eye-tracking task on a visit scheduled at least 1 week later. We used a Magic Touch KTMT-1315 add-on touch-screen (Keytec, Inc.), mounted with Velcro straps over the Tobii T60 monitor. The design of the task was the same as the eye-tracking task, but the images remained on the screen until the participant touched one of the two AOIs after hearing the word. If the participant touched the area outside of the two AOI images, the experiment would not advance to the next trial; if the participant repeatedly touched the images before the word onset, that trial was discarded. Before the administration of the test trials, the participant was introduced to the touch-screen method through a set of six practice trials, which could be repeated, if needed, until the participant provided three correct responses in a row on the training trials. During the administration of both the eye-tracking and the touchscreen tasks, a research assistant was present in the room to help with behavior management if needed and to reorient the participant's attention to the screen as needed.

Visit procedures

Testing took place over several visits to the lab, each lasting approximately 2 h. The visits usually started with administering the PPVT-4 or Ravens, but if the participant was not compliant (e.g. getting up from the table, taking and throwing the testing materials, or not passing the training items), the experimenter would move on to administering one of two computer tasks or the ADOS. No more than three administration attempts were made for the standardized tests. Caregivers who accompanied the children and adolescents to the lab completed the set of questionnaires during the visits or at home.

Data processing

Eye-tracking task

Data screening.: Prior to analyses of language-mediated eye movements, we assessed an overall measure of attention allocation to the stimuli, by investigating the number of trials on which the participants attended to the screen-display after hearing the word. Specifically, the data were examined for fixations within the time window of 200 ms post-word onset to the end of the trial (*approx*. 2400 ms post-auditory word onset). This time window was chosen because research with older typically developing children and adults indicates that it takes around 200 ms to plan and launch an eye movement (Dahan et al., 2001). On average, participants contributed reliable eye-gaze data within the time window selected on 53 of the 84 trials, with individual contributions ranging from 30 to 82 trials. Because little is known about the attentional characteristics of minimally verbal individuals with ASD, and because we wanted to be able to determine to what extent, if at all, words heard would influence attention monitoring in this population, we included in the analyses participants who provided data on a relatively low number of trials. Figure 2(a) and (b) illustrates the proportion of trials with reliable data provided by each participant for each individual block of trials. Individual contributions of usable gaze data trials averaged over the three blocks ranged from 36% to 97%, with a group mean of 67%. More specifically, 4 participants provided reliable gaze data on over 80% of the trials (of these participants, 2 had over 95% reliable gaze data trials), 10 contributed between 50% and 80% (of which 5 had reliable gaze data on over 70% of the trials), while 4 participants contributed reliable gaze data on fewer than 50% of the trials. Only trials with at least one fixation in any AOI post-word onset were included in statistical analyses. Based on these criteria, about 48% of trials were lost or removed from analyses across the 18 participants who could be calibrated. The proportion of reliable trials contributed by participants did not differ significantly between blocks (all p > 0.70). Given the substantial variability in participants' attending to the AOIs, we present individual data in all figures and tables, underscoring that any inferences about lexical comprehension should take into account the heterogeneity in basic *attentional processes* in this population.

We calculated two looking-time measures of lexical comprehension:

1. *Proportion of total fixation duration*. Following the widely accepted approach to interpreting fixation patterns in a looking-while-listening (LWL) procedure (Fernald et al., 2008; Venker et al., 2013), we operationalized participants' accuracy as a proportion of looking time to the target picture in each trial, after the onset of the auditory stimulus. To control for individual variation in overall looking toward the screen, we calculated the proportional looking time spent in each AOI relative to the total fixation time on the entire screen for each trial. This variable is considered indicative of word comprehension if participants spend proportionally more time fixating on the matching image AOI over the foil image.

2. *Image-related fixation duration difference score*. We calculated differences between looking time at an image when presented as target and looking time at the same image when presented as foil on a different trial. The rationale was to verify whether the within-trial differences in the proportional looking time to the target AOI relative to the foil AOI were related to the process of matching the word with the corresponding image (i.e. comprehension of the word) rather than related to characteristics of the image itself. By calculating these difference scores, we were able to control for possible effects of image salience on attention deployment. Positive scores, indicating proportionally longer looking time at an

image when it matches a word than when it serves as foil, are considered indicative of word comprehension (Bergelson and Swingley, 2012).

Touch-screen task

Data screening.: Accuracy of responses on the touch screen was recorded for the trials in which the first touch response was made after the auditory stimulus onset (88.6% of trials). Even though we trained the participants to listen to the word before touching the screen, on some trials some of the participants started tapping one of the images before hearing the target word; these trials were removed from analyses. The distribution of removed trials was uneven across participants, ranging from 0 to 37 trials out of 84 trials. One participant (who received only the first block of 28 trials) had 1 trial removed, 4 participants had more than 20 trials (out of 84), while 11 participants had fewer than 5 trials removed (of those, 5 participants responded appropriately on all 84 trials); the remaining 2 participants had 12 and 18 trials removed, respectively. Given that a relatively high rate of touch responses unrelated to the words could indicate either a high degree of inattention and impatience or not understanding the task, we conducted analyses first including all valid trials from all participants tested, then with 15 participants after excluding 3 who had a consistently high rate of invalid trials across all blocks. There were no significant differences in group-level results with and without the 3 participants with high levels of invalid trials, t(14) = 0.91, p = 0.38. Table 2 presents the percentage of reliable trials contributed by individual participants in the eye-tracking and the touch-screen tasks.

Results

A: Standardized assessments of receptive vocabulary, nonverbal IQ, and caregiver report measures

The PPVT-4 was administered to 18 of the 19 participants, of whom 6 (32% of the sample) were unable to achieve a basal score, and 5 established a basal score in the first set, but obtained a standard score of 20, the floor on this measure. The remaining 7 participants (39% of the sample) scored between 1.5 and 4 standard deviations (SDs) below the mean, with one participant obtaining a standard score of 76 and 4 participants obtaining standard scores below 50. Table 2 illustrates which participants scored at floor on the PPVT-4. For the subgroup of 7 participants with scores above the floor, the mean (and SD) standard score was 43.3 (21.5). The same 18 participants were administered the Raven’s matrices to estimate nonverbal intellectual functioning. Twelve participants scored at floor on the Raven’s (standard score of 65), including the 11 participants who scored at floor on the PPVT (see Table 2). The group of six participants with scores above the floor on the Raven’s had a mean standard score of 90.8, and a mean age equivalent score of 92 months. Comparison of individual scores on the standardized assessments of verbal and nonverbal intellectual functioning indicate a diverse profile of abilities in our sample, largely unrelated to participants’ age. Figure 3(a) and (b) illustrates individual ratio scores on the PPVT-4 and Raven’s for each participant tested. The mean (and SD) of ratio nonverbal IQ scores for the entire sample (18 participants) was 62.7 (29) ranging from very low to above-age expectations.

On the Caregiver Vocabulary Checklist, no participant was considered to understand fewer than 10 words and only one caregiver marked fewer than 20 words understood, while 15 caregivers marked between 100 and 250 words on the list as being understood by their child. The group mean for words "understood always" was 153.3 (SD = 68) but individual ratings ranged from 17 to 258 words. Table 2 presents the percentage of listed words that caregivers marked as understood by their child, with and without total certainty.

The Vineland-II Communication domain was completed by 17 of the caregivers. For the Receptive Communication subdomain (Listening and Understanding), raw scores ranged between 7 and 32, with a group mean of 19.06 (SD = 8.1), and a mean v-scale score[2] of 5.23 (SD = 2.8), corresponding to the 0.1 percentile rank on this measure.[3] Analyses of individual items on the Vineland-II revealed significant heterogeneity in the language comprehension skills of the participants. For instance, caregivers of nine participants (53% of the sample) were certain their child could understand and follow one-step directions, while caregivers of three participants were certain their child could not. Similarly, six caregivers (35% of the sample) indicated that their child could always understand if–then statements, while five (30%) were certain their child could not. Some of the items on which caregivers responded more consistently were comprehension of idioms and figurative language, answering questions appropriately with words, understanding and following three-step instructions, and asking "wh" questions. The majority of children (over 82% of the sample) were rated as lacking these abilities.

B: Experimental tasks

Eye tracking.—All 19 participants were administered the eye-tracking task, but one participant failed calibration. Of the remaining 18, 16 completed all three blocks of the experiment for a total of 84 trials, while 2 participants completed two blocks (56 trials).

Proportion of total fixation duration.—An analysis of variance (ANOVA) with trial type (target, foil) and experiment-block (1 to 3) as repeated measures yielded a significant effect of trial type, F(1, 15) = 13.4, p=0.002, η^2=0.472, but no other effects or interaction. On average, across the three blocks of the experiment, participants looked proportionally longer at the target images (M = 59.1%, SD = 8.9) than at the foil images (M=41.3%, SD = 8.9), suggesting that eye movements provide evidence of word comprehension for minimally verbal children and adolescents with ASD.

Fixation difference score.—The difference-score measure provided a robust index of word comprehension, as suggested by the significant difference between the number of positive versus negative difference-scores, relative to the number of trials with valid data for both instances of the same image being displayed as target and as foil, t(17) = 3.35, p = 0.002 (mean=75% positive difference-scores compared to 25% negative or zero difference-scores across all relevant trials). Given that half the images appeared first as foils, the image novelty or salience could not explain these results. However, the number of participants

[2]. *V*-scale scores range from 1 to 24 with an average score of 15 and a standard deviation of 3.

[3]. When compared to special population norms, based on the table for "Autism Special Population: Mute, ages 10 years and older" (cf. Carter et al., 1998), participants in our sample scored between the 10th and 75th percentile rank.

providing valid data on both instances of the image displayed in different image-pairs varied from 6 to 11, resulting in 625 difference-scores (from a possible total of 1326). Thus, this pattern of results can be reliably interpreted as an indicator of word comprehension for about half the participants in this study.

Touch-screen task.—Of the 18 participants who were administered the touch-screen task, 17 completed all three blocks, while 1 finished only the first block. Accuracy rates varied widely across the participants: while six responded at chance level, seven participants obtained accuracy levels of 90% and higher, with the remaining five achieving accuracy levels between 59% and 78%. The group mean accuracy was 72.1% (SD = 20.4%) when including all participants, and 73.6% without the three participants with a large number of invalid trials. Although the words in the three consecutively administered blocks increased in difficulty, there were no significant differences between the three blocks of trials, F(2, 15) = 1.62, p = 0.23, η^2=0.17 (M=70.59%, SD = 21.6 for Block 1, M=72.41%, SD = 21.6 for Block 2, and M=69.41%, SD = 19.9 for Block 3).

Correlations between types of assessments and comparison of assessment measures.—We found robust significant correlations among all measures of single-word comprehension: PPVT-4, caregiver checklist, eye-tracking, and touch-screen accuracy (see Table 3), thus providing an index of validity for the two experimental tasks. The exception to this pattern of relations among language measures was the Vineland receptive communication score, which did not correlate with the other language measures. As shown in Table 3, scores on the PPVT-4, touch-screen accuracy, and the caregiver word checklist were significantly negatively correlated with the ADOS scores, suggesting an inverse relationship between autism symptom severity and lexical comprehension among minimally verbal children and adolescents.

Figure 4(a) and (b) show individual performance (i.e., percentage accuracy) on the three different types of assessment of receptive vocabulary used, for the 84 words overlapping across all these measures. For the one participant who completed only one block on the touch-screen task, we compared the 28 words from that block to the same smaller set of words on the caregiver checklist and eye-tracking tasks. Of the 18 participants, 11 showed the highest accuracy on the touch-screen task, 2 on the eye-tracking task, and for 5 participants the vocabulary checklist included the highest proportion of known words (out of the overlapping words), according to caregivers' report.

Discussion

In this study, we examined several methods for assessing language comprehension in minimally verbal children and adolescents with ASD based on the assumption that even with uniformly minimal expressive skills, receptive language may vary in this population, but that conventional standardized assessments often fail to capture their linguistic potential. While experimental methods using eye-tracking tasks have been used with children with ASD in several other studies, this is, to our knowledge, the first study to include a minimally verbal population, and comparing across multiple assessment measures.

We adopted an individual differences approach to examining profiles of receptive language abilities on each of our measures in an effort to determine which of the assessment methods might prove most effective for providing reliable estimates of receptive language abilities in minimally verbal individuals with ASD. But evaluating the effectiveness of alternative approaches to assessing receptive language depends crucially on comprehensive characterizations of the cognitive and behavioral phenotypes of this population. Therefore, we will discuss findings regarding individual differences in profiles of abilities among the minimally verbal children and adolescents included in our research in parallel with comparing and evaluating the effectiveness of the different methods of assessing receptive language used in our study.

Results of this multiple-method approach revealed significant heterogeneity in receptive language abilities across participants and across assessment methods. Performance on direct assessments can be substantially influenced by behavioral challenges, engagement with the tasks and motivation to succeed, profiles of attention deployment, and cognitive impairments, among other factors related to the testing situation. Reliable assessments are complicated by individual variation in such factors, as well as by children's varied histories of interventions and treatments. Our multiple-method approach to assessment was an attempt to address some of these challenges in several ways. We developed experimental tasks similar in design but which placed different behavioral demands on the participant; we varied the social-pragmatic requirements of the assessment situation by using computerized tasks as well as face to face, experimenter-administered tests; we modified the format and administration style of some tests to make the assessment situation less stressful and even enticing for the participant; and we used an experimenter-driven administration of the computerized tasks to minimize attention-related data loss. In sum, we used an individualized behavior management approach in testing on all the measures administered to the participants and have provided details of each of our modifications in order to guide future work in this area.

The adaptations did not necessarily have a significant impact on performance on the standardized tests: the PPVT-4 and the Ravens. On the PPVT-4, more than half of the sample over 60% (N = 11) were at floor or failed to establish a basal, despite our efforts to increase the number of teaching trials and to model a pointing response before administering the PPVT-4 test items. The 11 participants who scored at floor level on the PPVT-4 also performed at floor level on the Ravens, a result which underscores the often encountered comorbidity of language-related impairments and nonverbal cognitive deficits (Doshi-Velez et al., 2014; Joseph et al., 2002; Lincoln et al., 1995). However, two of our participants scored at age level (i.e. within one SD of the norm) and one scored above age level expectations on our measure of nonverbal IQ while no participant approached age expectations on the PPVT-4, on which the highest standard score was 76, obtained by the youngest child in the study. The presence of floor effects on the two standardized tests makes it difficult to appreciate whether the minimally verbal population with ASD includes mostly individuals with consistent verbal and nonverbal profiles of deficits, or whether the prevalent profile is that of higher nonverbal than verbal abilities, but with impairments in both areas relative to typical peers. Nevertheless, in our sample, every participant had higher ratio scores on the Ravens compared to the PPVT-4 (see Figure 3(a) and (b)), suggesting that nonverbal reasoning abilities are higher than verbal abilities at this end of autism spectrum.

Significant variability was also found in the caregivers' ratings of how many words their child understands, ranging from 6% (17 words) to 95% (260 words) from a list of 275 common nouns, verbs, and adjectives. Children who failed to obtain a basal on the PPVT-4 were also reported by their caregivers as understanding fewer than 50% of the words listed, suggesting that the social-pragmatic challenges of the standardized testing context can only partially explain why some children could not provide valid test data. It is important to note that caregiver ratings of their children's comprehension of specific words were highly correlated with child scores on the other assessments of vocabulary knowledge administered, including the experimental tasks providing cross-measure validity to the different measures of single-word vocabulary that we included in this study.

An exception to this pattern of correlations among language measures was found for the Vineland receptive communication subscale score, which was not correlated with any of the other language measures (see Table 3). This subscale includes questions about language that go beyond single-word vocabulary as well as various aspects of the communicative use of language, including pragmatic aspects of communication. Therefore, it is not too surprising that scores from this caregiver report instrument were not significantly correlated with other evaluations of receptive language in this study, given that both the standardized and the experimental measures administered tapped understanding of single words, in the absence of any contextual information.

The two computer-based experimental tasks minimized the participants' interaction with an examiner, though they were programmed to allow the examiner to control the succession of trials, by monitoring the participants' attention to the screen. Another difference between our eye-tracking task and those used in other studies focusing on single-word comprehension is the substantially higher number of trials included, distributed over three blocks, with the added flexibility of administering different blocks during different research visits. While other research using the LWL method probed comprehension of 6 to 18 words (e.g. Venker et al., 2013, tested eight target words; Bavin et al., 2014, included 18 target items), usually words acquired early in language development, our tasks included 84 target words, with 28 target words per block, which increased in difficulty from the first to the last block.

Results on both the experimental computerized tasks revealed a wide range of performance, measured by visual attention deployment in the eye-tracking version, and by a number of correct responses in the touch-screen version of the task. Performance on these tasks is highly dependent on attention monitoring and should be analyzed in relation to the proportion of usable data trials contributed by each participant. Given the cognitive and behavioral profile of our minimally verbal participants, we did not want to drop from analyses individuals who contributed a relatively small number of usable trials, because these individuals are most likely representative of this population. Although increasing the number of trials to 84 and using an experimenter-controlled administration lengthened the duration of the tasks, this approach enabled us to obtain sufficient reliable data from each participant to be able to interpret eye movement and touch-screen responses as indices of word comprehension. It is important to note that even with the experimental adaptations described above, the proportion of data loss was around 40% on the eye-tracking tasks and around 20% on touch-screen task. There was considerable variability in the percentage of data loss across participants. For the eye-tracking task in particular, the relatively high proportion of data loss is a significant limitation, because, for some individuals, knowledge of some of the words could not be reliably assessed if those stimuli were not attended to by the participant. By comparison, in the touch-screen task participants contributed on average 90% trials with reliable responses (correct or incorrect), with only three participants contributing reliable data on fewer than 70% of the trials. Accuracy, however, ranged from 43% to 100%, though 9 out of 18 participants responded correctly on over 75% of the trials, suggesting that 60 or more target words (out of 84) were understood by at least half of the participants. These are encouraging findings given that many of these participants were not testable on the PPVT, even though many of the words included in the experimental tasks were drawn from that measure.

One interpretation of the findings discussed is that significant variability in performance is found across all our measures, and that several alternative approaches to assessing receptive language including new technologies could provide more reliable assessment of language comprehension than the commonly used but more limited standardized tests. However, clear advantages of one method over another did not emerge from this study. The significant heterogeneity in performance found on the different approaches to language assessment compared in our study may reflect true differences in abilities, or differences related to behavioral and motivational factors, influential even in a simple passive viewing paradigm.

This study is a first attempt to systematically compare different types of receptive language assessments, including methods based on innovative technologies, with a minimally verbal population of individuals with ASD. Our results do not demonstrate that any one of the alternative approaches to assessment discussed here may be ready for integration into clinical practice at this time. More research is needed before direct clinical applications could be derived from the experimental use of these innovative, technology-based assessment methods of receptive language. Nevertheless, our findings suggest that an important avenue for capturing the true potential for language comprehension of minimally verbal children who remain otherwise untestable is to find individualized approaches to testing, using several types of assessment, including methods based on eye-tracking or touch-screen responding. These new technologies provide reliable alternative approaches to assessment that can be more easily used with the minimally verbal population. It is now time to develop a range of measures that can capture the receptive language abilities of minimally verbal children and adults that go beyond single-word vocabulary and to begin implementing these measures in clinical practice. Ultimately, interventions depend crucially on knowing how much language and in which contexts a person can understand.

Acknowledgements

We offer special thanks to the participants and their families for their involvement in our research.

Funding

This research was supported by grants from Autism Speaks and NIH (RO1 DC 10290-S1; P50 DC 13027).

References

American Psychiatric Association (APA) (2013) Diagnostic and Statistical Manual of Mental Disorders. 5th ed Washington, DC: APA.

Bavin E, Kidd E, Predergast L, et al. (2014) Severity of autism is related to children's language processing. Autism Research 7: 687–694. [PubMed: 25262588]

Bergelson E and Swingley D (2012) At 6–9 months, human infants know the meanings of many common nouns. Proceedings of the National Academy of Sciences A 109: 3253–3258.

Bruckner C, Yoder P, Stone W, et al. (2007) Construct validity of the MCDI-I receptive vocabulary scale can be improved: differential item functioning between toddlers with autism spectrum disorders and typically developing infants. Journal of Speech, Language, and Hearing Research 50: 1631–1638.

Carter AS, Volkmar FR, Sparrow SS, et al. (1998) The Vineland Adaptive Behavior Scales: supplementary norms for individuals with autism. Journal of Autism and Developmental Disorders 28(4): 287–302. [PubMed: 9711485]

Dahan D, Magnusson JS and Tanenhaus MK (2001) The course of frequency effects in spoken-word recognition: evidence from eye-movements. Cognitive Psychology 42: 317–367. [PubMed: 11368527]

Dawson M, Soulières I, Gernsbacher MA, et al. (2007) The level and nature of autistic intelligence. Psychological Science 18: 657–662. [PubMed: 17680932]

Doshi-Velez F, Ge Y and Kohane I (2014) Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis. Pediatrics 133: e54–e63. [PubMed: 24323995]

Dunn LM and Dunn LM (2007) Peabody Picture Vocabulary Test. 4th ed Circle Pines, MN: American Guidance Service.

Fenson L, Marchman VA, Thal DJ, et al. (2007) MacArthur-Bates Communicative Development Inventories User's Guide and Technical Manual – Second Edition Baltimore: Brookes Publishing.

Fernald A and Marchman VA (2012) Individual differences in lexical processing at 18 months predict vocabulary growth in typically developing and late-talking toddlers. Child Development 83: 203–222. [PubMed: 22172209]

Fernald A, Zangl R, Portillo AL, et al. (2008) Looking while listening: using eye movements to monitor language comprehension by infants and young children In: Sekerina IA, Fernandez EM and Clahsen H (eds) Developmental Psycholinguistics: On-Line Methods in Children's Language Processing. Amsterdam: John Benjamins, pp.113–135.

Gernsbacher MA, Geye H and Ellis Wesimer S (2005) The role of language and communication impairments within autism In: Fletcher P and Miller JC (eds) Language Disorders and Developmental Theory. Philadelphia, PA: John Benjamins, pp.73–93.

Hus V and Lord C. The ADOS adapted modules: An autism diagnostic observation measure for adolescents and adults with limited language. (In preparation).

Joseph RM, Tager-Flusberg H and Lord C (2002) Cognitive profiles and social-communicative functioning in children with autism spectrum disorder. Journal of Child Psychology and Psychiatry 43(6): 807–821. [PubMed: 12236615]

Kasari C, Brady N, Lord C, et al. (2013) Assessing the minimally verbal school-aged child with autism spectrum disorder. Autism Research 6: 479–493. [PubMed: 24353165]

Lincoln AJ, Allen M and Kilman A (1995) The assessment and interpretation of intellectual abilities in people with autism In: Schopler E and Mesibov G (eds) Learning and Cognition in Autism. New York: Plenum, pp. 89–117.

Lord C, Rutter M, DiLavore PC, et al. (2012) Autism Diagnostic Observation Schedule (ADOS), 2nd Edition Manual: Modules 1–4. Torrance, CA: Western Psychological Services.

Luyster RJ, Kadlec MB, Carter A, et al. (2008) Language assessment and development in toddlers with autism spectrum disorders. Journal of Autism and Developmental Disorders 38: 1426–1438. [PubMed: 18188685]

Norbury CF, Griffiths H and Nation K (2010) Sound before meaning: word learning in autistic disorders. Neuropsychologia 48: 4012–4019. [PubMed: 20951710]

Rapin I, Dunn MA, Allen DA, et al. (2009) Subtypes of language disorders in school-age children with autism. Developmental Neuropsychology 34: 66–84. [PubMed: 19142767]

Raven J, Raven JC and Court JH (1998) Manual for Raven's Progressive Matrices and Vocabulary Scales Section 2: The Coloured Progressive Matrices. San Antonio, TX: Harcourt Assessment.

Sparrow S, Cicchetti D and Balla D (2005) Vineland Adaptive Behavior Scales. 2nd ed Circle Pines, MN: AGS.

Tager-Flusberg H (1999) The challenge of studying language development in autism In: Menn L and Bernstein Ratner N (eds) Methods for Studying Language Production. Mahwah, NJ: Lawrence Erlbaum Associates, pp.313–332.

Tager-Flusberg H and Kasari C (2013) Minimally verbal school-aged children with autism spectrum disorder: the neglected end of the spectrum. Autism Research 6: 468–478. [PubMed: 24124067]

Tager-Flusberg H, Paul R and Lord CE (2005) Language and communication in autism In: Volkmar F, Paul R, Klin A, et al. (eds) Handbook of Autism and Pervasive Developmental Disorder. 3rd ed, vol. 1 New York: Wiley, pp.335–364.

Tager-Flusberg H, Rogers S, Cooper J, et al. (2009) Defining spoken language benchmarks and selecting measures of expressive language development for young children with autism spectrum disorders. Journal of Speech, Language and Hearing Research 52: 643.

Venker CE, Eernisse ER, Saffran JR, et al. (2013) Individual differences in the real-time comprehension of children with ASD. Autism Research 6: 417–432. [PubMed: 23696214]

Yurovsky D, Wade A and Frank MC (2013) On-line processing of speech and social information in early word learning In: Knauff M, Pauen M, Sebanz N, et al. (eds) Proceedings of the 35th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society, pp.1641–1646.

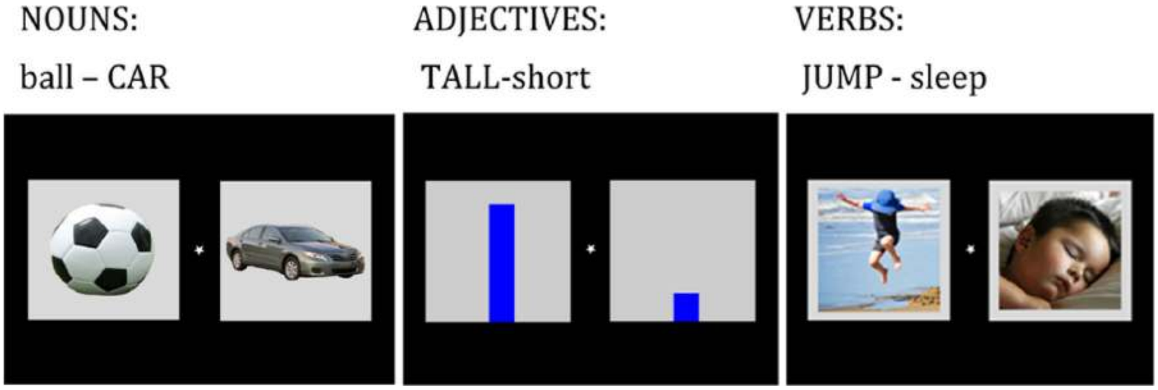

Figure 1.

Sample image-pairs (target word in capital letters).

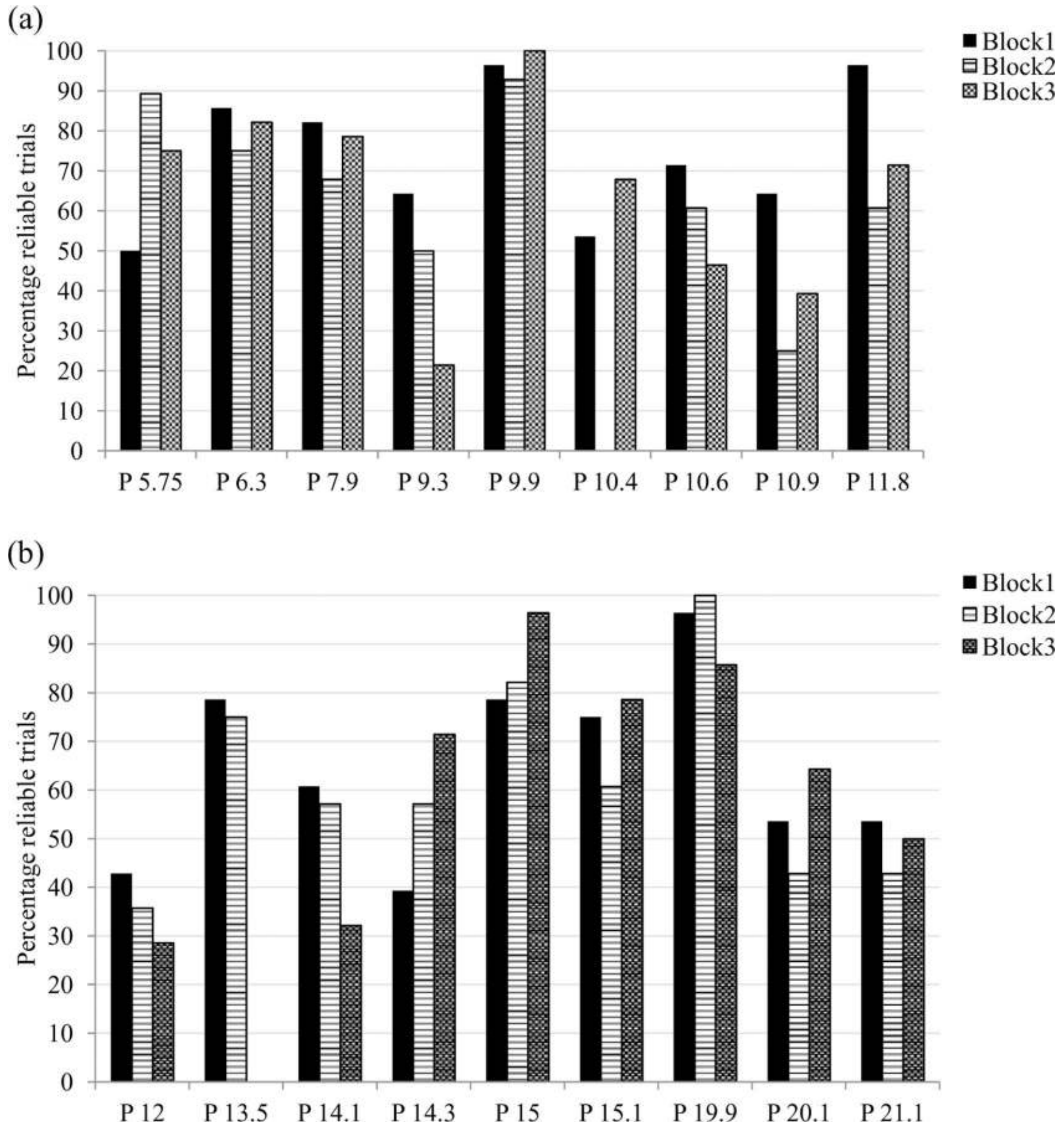

Figure 2.

Percentage of reliable gaze data trials contributed by each participant. Individual participants (P) are labeled by their age: (a) younger, including participants between 5.75 and 11.8 years and (b) older, including participants between 12 and 21.1 years.

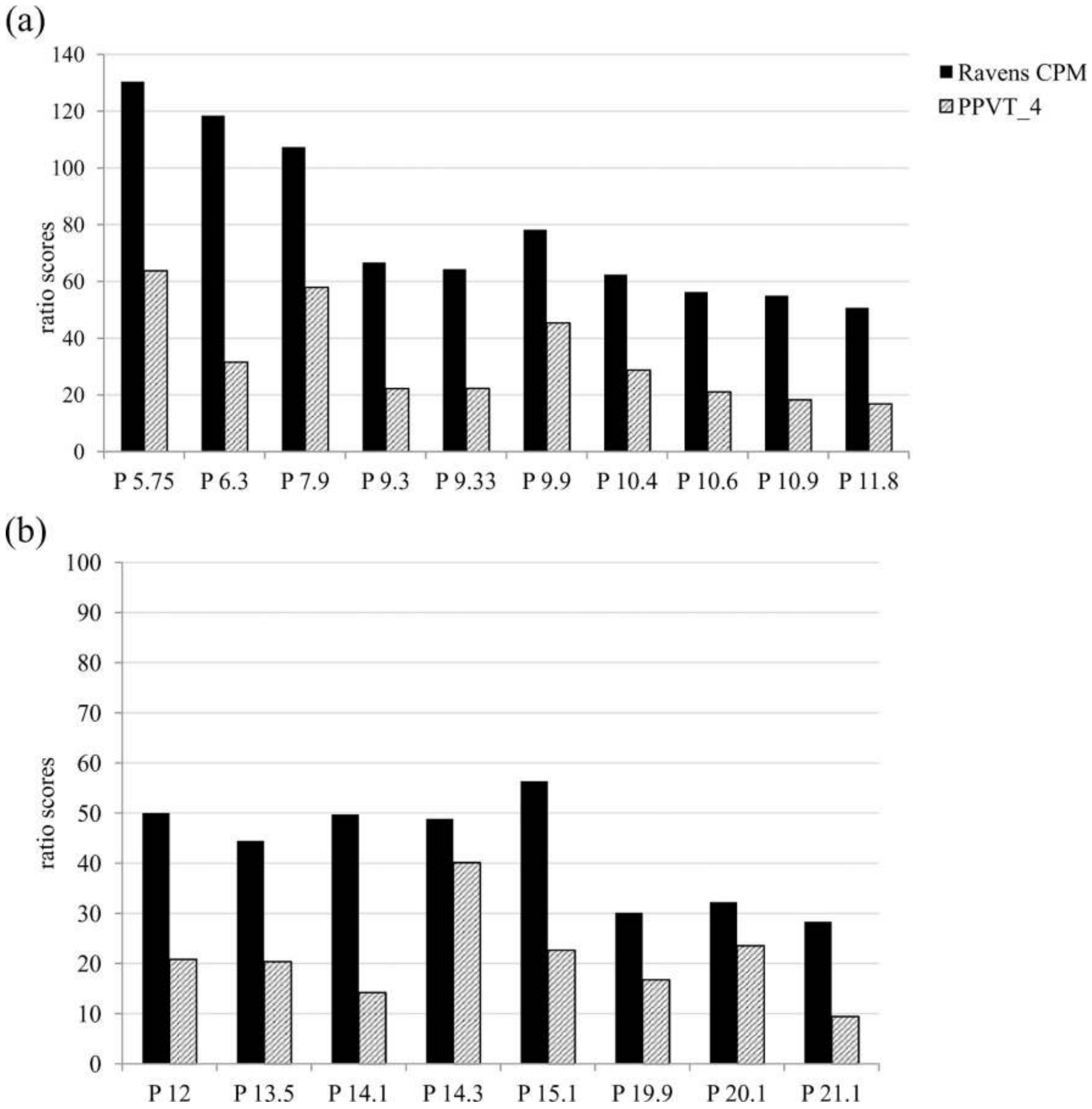

Figure 3.

Individual performance on standardized measures of nonverbal IQ and vocabulary (ratio scores) (a): Younger group and (b): Older group.

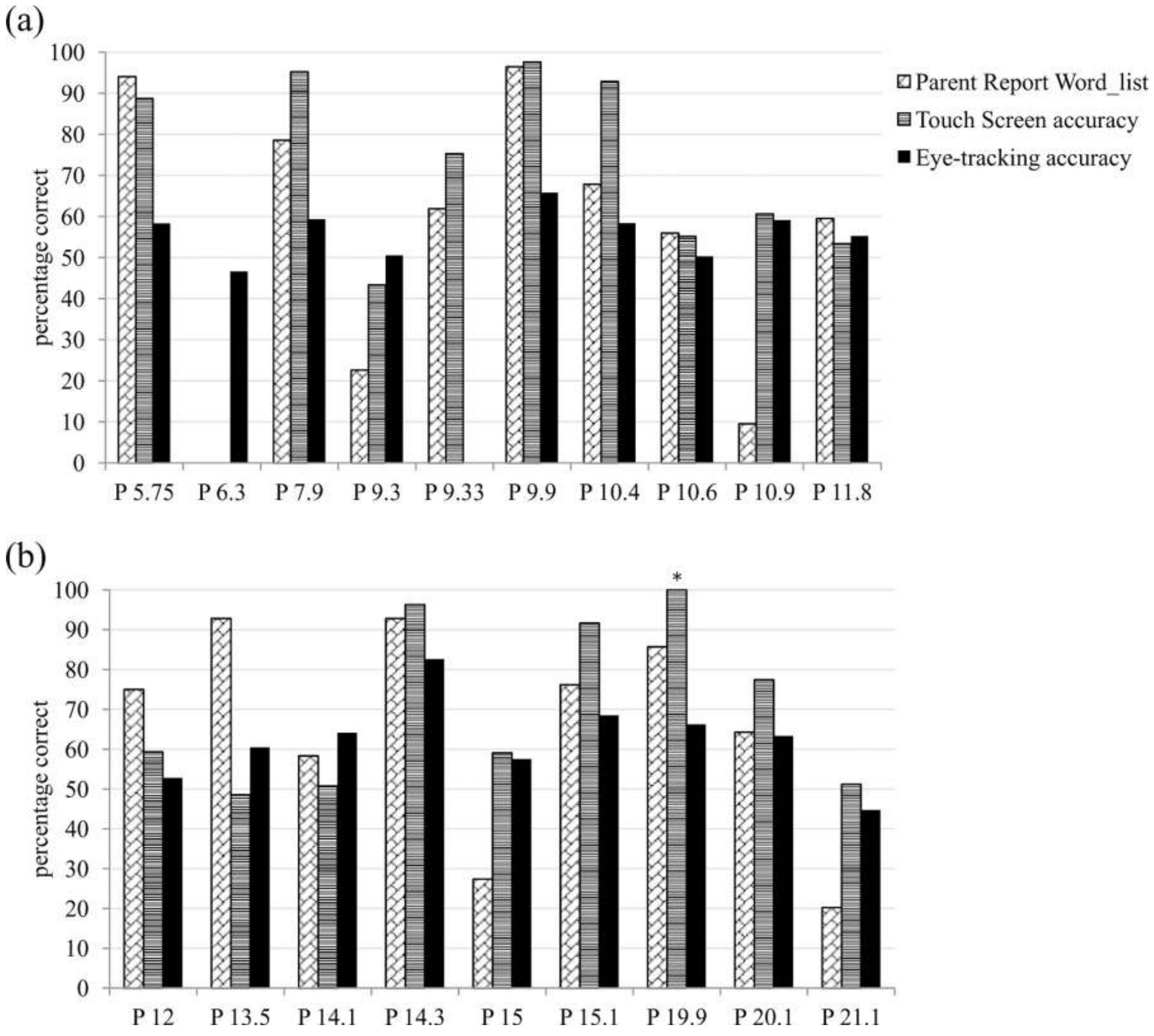

Figure 4.

Individual performance on three measures of receptive vocabulary for words overlapping across the three assessments (percentage accuracy): (a) younger group and (b) older group. *For this participant based on words in Block 1 only (28 words).

Table 1.

Descriptive characteristics of the participants.

	N	Mean	Standard deviation	Range
Chronological age (months)	19	149.9	53.2	69–254
Adapted ADOS total	16	20.75	5.01	13–28
ADOS Module 1 total	2	21.5	6.36	17–26
Vineland Adaptive Behavior Composite (Standard Score)[a]	14	46.6	10.7	27–65
Vineland Communication Domain (Standard Score)	17	46.1	10.8	26–65
Raven Colored Progressive Matrices ratio nonverbal IQ scores	18	62.7	29	28.4–130.4
PPVT-4 ratio scores	18	27.6	15	9.5–64
Ratio males/females	15/4			

ADOS: Autism Diagnostic Observation Schedule; PPVT: Peabody Picture Vocabulary Test.

[a] Four caregivers did not complete other sections of the Vineland besides the Communication domain; therefore, the Adaptive Behavior Composite score could be computed only for 14 participants.

Table 2.

Individual performance across different assessments.

Participant	Raven floor	PPVT floor	Word list—% always understood	Word list—% maybe understood	Vineland communication standard score	Eye tracking—% valid	Eye tracking—% accuracy	Touch screen—% valid	Touch screen—% accuracy
P 5.75	No	No	92.50	7.43	65	71.43	58.32	96.43	89.74
P 6.3	No	No	55.00	41.00	62	80.95	46.59		
P 7.9	No	No	75.65	6.64	53	76.19	59.33	100.00	95.24
P 9.3	Yes	Yes	19.33	28.62	45	45.24	50.54	65.48	43.37
P 9.33	Yes	Yes	59.11	32.71	43			55.95	75.29
P 9.9	No	No	91.82	8.18	37	96.43	65.80	100.00	97.62
P 10.4	No	No	57.62	23.05	61	60.71	58.35	100.00	92.86
P 10.6	Yes	Yes	44.60	36.80		59.52	50.31	73.81	55.16
P 10.9	Yes	Yes	6.30	35.93		42.86	59.13	96.43	60.64
P 11.8	Yes	Yes	46.72	22.99	48	76.19	55.26	66.67	53.43
P 12	Yes	Yes	60.37	26.79	50	35.71	52.76	96.43	59.31
P 13.5	Yes	Yes	94.85	2.21	42	76.79	60.42	85.71	48.59
P 14.1	Yes	Yes	52.74	45.42	47	50.00	64.15	94.05	50.76
P 14.3	Yes	No	81.41	10.41	40	55.95	82.62	98.81	96.34
P 15			23.87	29.92		85.71	57.51	78.57	59.09
P 15.1	Yes	Yes	65.41	28.57	49	71.43	68.50	100.00	91.67
P 19.9	Yes	Yes	71.79	7.33	40	94.05	66.23	96.43[a]	100.00[a]
P 20.1	No	No	56.72	24.63	30	53.57	63.33	95.24	77.45
P 21.1	Yes	Yes	26.37	31.50	26	48.81	44.71	100.00	51.19
Group data:									
N above floor	6	7	19	19	16	18	18	18	18

PPVT: Peabody Picture Vocabulary Test.

[a] For this participant based on words in Block 1 only (28 words).

Table 3.

Correlations among measures.

	Raven matrices raw score	PPVT-4 raw score	Caregiver checklist words understood	Vineland-II receptive communication raw score	Mean % looking time at target	Touch-screen accuracy	Adapted ADOS total
Raven matrices raw score		0.559 *	0.519 *	0.044	0.385	0.649 *	−0.534 *
PPVT-4 raw score			0.683 **	−0.334	0.705 **	0.802 **	−0.636 *
Caregiver checklist words understood				0.021	0.498 *	0.600 **	−0.430
Vineland-II receptive communication raw score					0.065	0.066	0.127
Mean % looking time at target						0.642 **	−0.413
Touch-screen accuracy							−0.517 *

ADOS: Autism Diagnostic Observation Schedule; PPVT: Peabody Picture Vocabulary Test.

** $p < 0.001$;

* $p < 0.05$.