CrossMark

Applied Psycholinguistics **37** (2016), 529–549
doi:10.1017/S0142716415000107

Understanding and assessing word comprehension

BEVERLY A. GOLDFIELD, CHRISTINA GENCARELLA, and KEVIN FORNARI
Rhode Island College

Received: February 16, 2014 Accepted for publication: December 31, 2014

ADDRESS FOR CORRESPONDENCE
Beverly A. Goldfield, Department of Psychology, Rhode Island College, 600 Mt. Pleasant Avenue, Providence, RI 02908. E-mail: bgoldfield@ric.edu

ABSTRACT
The intermodal preferential looking (IPL) task was developed to assess comprehension in infants and toddlers. We extend this methodology to examine word comprehension in preschool children using two measures: proportion of looking time to target (LTT) and longest look (LL) to target. Children (3–6 years) were tested with the IPL for comprehension of nouns, verbs, and adjectives. Both LTT and LL scores showed that, across all ages, eye gaze to the target word increased from baseline to test; there were higher scores for nouns compared to verbs and adjectives. We also compare IPL performance to scores on a standardized test of receptive vocabulary (the Peabody Picture Vocabulary Test—Fourth Edition). Correlations with Peabody Picture Vocabulary Test scores were stronger for LTT than LL measures. The IPL may provide an alternative method for assessing word comprehension in preschool children with behavioral limitations.

The data that inform theories of language acquisition are typically measurements of language production. Language comprehension, particularly in the early years, is more difficult to assess. Among the earliest studies of comprehension are vocabulary data from parent report, as in research by Huttenlocher (1974), Benedict (1979), and Harris, Yeeles, Chasin, and Oakley (1995). These studies address basic questions regarding word comprehension, including age of onset, rate of development, types of words comprehended, and the relationship between comprehension and the onset of word production, but diary reports are labor intensive, yielding small samples with potential problems in generalizability. Parent-report data from larger samples are available from vocabulary checklists such as the MacArthur–Bates Communicative Development Inventories (Fenson et al., 1994) that have been adapted for more than 60 languages, making cross-linguistic comparisons a significant feature of this methodology. In the English version of the MacArthur–Bates Communicative Development Inventories, Words and Gestures, words are arranged in 19 semantic categories such as people, food, toys, actions, household objects, and so on. Parents of children age 8 to 16 months are asked to

© Cambridge University Press 2015 0142-7164/15

https://doi.org/10.1017/S0142716415000107 Published online by Cambridge University Press

check words their child understands and/or says. Parent reports have been a rich source of information about early word comprehension, but a disadvantage is the variability inherent in parent judgment of comprehension.

Another approach to assessing word comprehension utilizes behavioral measures carried out in the lab or in the child's home environment. Children are asked to choose an object or to carry out an action named by the experimenter (e.g., Goldin-Meadow, Seligman, & Gelman, 1976; Harris et al., 1995). Such measures depend crucially on the ability of young children to cooperate and to have the requisite motor skills to execute the behavior. An alternative experimental paradigm used to study comprehension processes in both children and adults makes use of a far less-demanding skill: visual attention as measured by the direction and duration of eye gaze. The use of visual attention to index how cognitive systems process spoken language is the basic premise of varied studies in the "visual world paradigm" (VWP; Allopenna, Magnuson, & Tanenhaus, 1998; Cooper, 1974; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). In VWP research, participants attend to a visual display while hearing an utterance; eye movements to items in the display are recorded to infer varied aspects of linguistic processing, including phonological, semantic, and pragmatic associations. A strength of the VWP is that the same basic methodology can be used to explore language processing in adults, children, and atypical populations (Huettig, Rommers, & Meyer, 2011). For example, both children (Johnson & Huettig, 2011) and adults (Huettig & Altmann, 2005) attend to a color-matched distractor (e.g., eye gaze to a picture of a *red* plane rather than a *yellow* plane) when hearing the word *strawberry*, suggesting that a spoken word activates conceptual information (color) related to that word, even if the hearer does not yet know the specific color term. Similarly, eye movements of both typically developing adolescents and adolescents with autism who viewed a set of four objects (e.g., pills, hammer, medal, and hamster) while hearing a test sentence (e.g., *Joe stroked the hamster quietly*) were affected by semantic associations between the sentence verb (*stroke*) and target object (hamster) and by the extent to which the target word (*hamster*) phonologically overlapped one of the competitor objects (hammer). In this study, the use of an eye-tracking measure avoided metalinguistic skills that might explain the poor performance of participants with autism in previous studies that have examined comprehension of words in context (Brock, Norbury, Einav, & Nation, 2008).

A closely related methodology for examining comprehension in very young children is the intermodal preferential looking (IPL) task. The IPL was adapted (Golinkoff, Hirsh-Pasek, Cauley, & Gordon, 1987) from work on perception (Spelke, 1979) and research on signal detection (Thomas, Campos, Shucard, Ramsay, & Shucard, 1981). It was developed to test the ability of infants and toddlers to match speech with an image of its referent using the child's visual attention as the dependent variable. The IPL measures word comprehension by comparing a child's visual gaze to two images (target and distracter) displayed on a video or computer monitor before (baseline trial) and after (test trial) the target image is labeled. Images depicted on the monitor can be static (e.g., photographs of objects, persons, and places to depict nouns and adjectives) or dynamic (e.g., videos of actors performing actions to represent verbs or sentences). Comprehension is defined as an increase in visual attention to the target image during the

https://doi.org/10.1017/S0142716415000107 Published online by Cambridge University Press

test trial compared to baseline presentation. Visual attention can be scored during presentation by a hidden observer, or the sequence and duration of visual fixations can be videotaped and scored offline. Recent advances in eye-tracker technology have made automatic scoring possible and enhance the reliability and efficiency of this methodology. The IPL has been used to assess varied aspects of word comprehension during the first 2 years, including first words comprehended (Bergelson & Swingley, 2012; Tincoff & Jusczyk, 1999, 2012), comprehension of nouns and verbs (Golinkoff, et al., 1987), noun extension (Meints, Plunkett, & Harris, 1999), thematic roles linked to verbs (Meints, Plunkett, & Harris, 2008), word learning (Schafer, 2005; Schafer & Plunkett, 1998), the reliability and validity of parent report (Houston-Price, Mather, & Sakkalou, 2007; Styles & Plunkett, 2009), and the relationship between word comprehension and production (Reznick & Goldfield, 1992). The looking while listening task is a variant of the IPL in which the focus is on the time course of fixations to target versus distracter in order to document developmental change in the speed and efficiency of linguistic processing (e.g., Fernald, Perfors, & Marchman, 2006: Fernald, Swingley, & Pinto, 2001; Fernald, Zangl, Portillo, & Marchman, 2008).

The measurement of word comprehension is also of interest in applied contexts, such as testing for diagnostic and educational evaluations. Assessment of receptive vocabulary is crucial, for example, in determining the developmental status of young children with speech delays related to specific language impairment or more pervasive developmental disorders. However, standardized cognitive and language assessments typically depend on the child's ability and willingness to engage with the examiner, point to pictures, choose among objects, or act out commands. For example, items for children age 15–22 months on the receptive language subtest of the Mullen Scales of Early Learning (Mullen, 1995) request that the child wave bye, clap hands, and give the parent a toy named by the examiner. At 23–32 months, children are shown a series of black-and-white line drawings and are asked to point to body parts and objects such as *cat*, *car*, and *doll*. Preschoolers are asked to follow spatial directions (e.g., "Put the teddy bear *under* the table") and to point to items and pictures illustrating color terms (e.g., "Point to the *red* crayon") and size (e.g., "Touch the *smaller* box"). On the Preschool Language Scale (Zimmerman, Steiner, & Pond, 2002), comprehension is similarly tested by asking the child to manipulate toys (e.g., "Put *some* blocks here") and to point to pictures of attributes and actions (e.g., "Which one is *wet?* Point to *washing*") named by the examiner.

One of the most widely used standardized tests of word comprehension is the Peabody Picture Vocabulary Test—Fourth Edition (PPVT-4, Dunn & Dunn, 2007). The PPVT-4 is designed to assess level of word comprehension from age 2.5 years through adulthood using a test booklet that displays four color line drawings per page. The child is asked to point to the picture named by the examiner or to indicate the number of her choice. The PPVT-4 measures word comprehension with a total of 228 target words that consist of 75% nouns, 18% verbs, and 7% attributes. There are two alternate forms (A and B) of the test. The PPVT-4 provides standardized scores as well as age and grade equivalents. The PPVT-4 was normed on a representative, English-speaking sample of 3,540. It is used in schools and clinics as a screening instrument and in the assessment of language impairment

https://doi.org/10.1017/S0142716415000107 Published online by Cambridge University Press

in children, adults with aphasia, and geriatric cases with deteriorating language abilities. In addition to its usefulness in school and clinic, the PPVT-4 is also widely used in clinical and developmental research to establish the comparability of groups or as an outcome measure.

However, because the PPVT-4 and similar standardized tests used to assess word comprehension require sustained attention, engagement and cooperation with an examiner, and gestural or verbal responses, they may be unreliable or challenging to use with children under the age of 3 and older children with limitations in attention, social engagement, motor skills, or verbal ability. As an alternative, practitioners might consider the IPL. Although the IPL is a research methodology that was designed for testing infants and toddlers with a limited behavioral repertoire, it may be an appropriate alternative for children who, although chronologically older, pose similar challenges for language assessment. Children with physical impairments such as cerebral palsy may find it difficult to point or speak. Children diagnosed with developmental delays associated with autism spectrum disorders (ASD) have limited attention span and may be unwilling or unable to participate in the social engagement required for most standardized testing situations. Reliable developmental assessments become increasingly critical during the preschool years, when evaluations are needed for educational planning and placement.

The IPL has been used to study various aspects of word comprehension in normally developing infants and toddlers, but there has been little research extending the methodology to older, normally developing preschool or school-age children. It is not clear that a methodology with a format designed for the capabilities and attention span of children during their first 2 years will appeal to older children. Moreover, the number of words tested using the IPL with infants and toddlers has typically been limited to 12 or fewer words (e.g., Golinkoff et al., 1987; Houston-Price et al., 2007; Reznick, 1990; Styles & Plunkett, 2009); a meaningful assessment of comprehension with older children would require considerably expanding the number of word trials. Finally, IPL studies with infants and toddlers have indexed change in visual attention using measures of latency, proportion of looking time, and longest look. However, longer bouts of sustained attention and increasing sensitivity to the comprehensibility of the stimulus presentation emerge beyond the infancy period (Richards & Anderson, 2004), so it is not clear which of these several measures are appropriate indices for preschool children.

As an initial step in extending this laboratory methodology to older children and for potential use in applied contexts, the present study uses the IPL to assess word comprehension in a sample of normally developing preschool children. We test comprehension of nouns, verbs, and adjectives in 3-, 4-, 5-, and 6-year-olds, comparing two measures of visual attention. We also administer the PPVT-4 (Dunn & Dunn, 2007) and test the relationship between IPL and PPVT-4 scores. If a test of word comprehension using the IPL methodology correlates positively with a standardized test such as the PPVT-4, it suggests an important first step toward extending the IPL to assess various aspects of word comprehension in older, preschool children with behavioral limitations that might constrain their performance on standardized tests.

https://doi.org/10.1017/S0142716415000107 Published online by Cambridge University Press

METHOD

Participants

A total of 75 children participated in the study. Eight children were eliminated from the analyses because of instrument problems, leaving a total sample of 67 children; 19 (10 females, 9 males) at age 3 ($M = 3$ years, 6 months [3;6], range = 3;0–3;11), 16 (8 females, 8 males) at age 4 ($M = 4;5$, range = 4;1–4;11), 15 (6 females, 9 males) at age 5 ($M = 5;4$, range = 5;0–5;11), and 17 (6 females, 11 males) at age 6 ($M = 6;5$, range = 6;0–6;11). Participants were all normally developing children with no uncorrected hearing or visual impairments. English was the primary language used at home by parents and children. Participants were recruited from a college laboratory preschool and from a commercial mailing list using names and addresses of families with children in the target age range. Parents received a letter briefly describing the study and inviting them to contact the lab for further information. Interested parents were provided with details about the research and were scheduled for an appointment. Parents received a $15.00 gift card to a toy/book store for their participation.

Materials

For the IPL, we created a slideshow to assess comprehension of 21 nouns, 8 verbs, and 4 adjectives. Words were selected from various word frequency sources, including the *American Heritage Word Frequency Book* (Carroll, Davies, & Richman, 1971), *Basic Reading Vocabularies* (Harris & Jacobson, 1982), and the *Educators' Word Frequency Guide* (Zeno, Ivens, Millard, & Duvvuri, 1995). We selected vocabulary items that these sources indicated were in the spoken and/or reading vocabularies of children age 3 through 8 years. In IPL studies of word comprehension with infants and toddlers, each target word is typically matched to one distracter item. Each word tested on the PPVT-4 includes one target word with three distracters. Pilot testing indicated that sets of three words, one target and two distracters, were appropriate for the automated sequence of trials presented to the 3- to 6-year-olds in the present study.

For each noun and adjective trial, we created a set of three words that word frequency sources indicated were at a similar comprehension level (e.g., all three words were vocabulary appropriate for 5-year-olds). Nouns include words for people, animals, food, vehicles, household items, places, clothing, and body parts. Noun sets were selected from the same semantic category (e.g., a set of three animals, or three foods, or three vehicles). Three of the target nouns (*alphabet*, *tool*, and *furniture*) represent category names with the target image an appropriate exemplar (e.g., images of hammer, corn, and helmet for the target word *tool*). Adjectives include two sets of physical attributes (e.g., images that represent *sharp / deep / wide*) and two sets that depicted emotions (e.g., images of facial expressions representing *afraid / angry / sad*).

For verb trials, the three verbs in each set were matched for syntactic (transitive vs. intransitive) category and also represented similar levels of comprehension based on the word frequency sources. Four of the verb sets consisted of intransitive

https://doi.org/10.1017/S0142716415000107 Published online by Cambridge University Press

Table 1. *Target and distracter words*

Nouns	Adjectives	Verbs
crab: beaver, parrot	**sharp:** deep, wide	**reach:** bend, flap
coconut: mushroom, onion	**hollow:** steep, shallow	**limp:** creep, shrug
helicopter: bulldozer, crane	**ashamed:** surprised, happy	**fold:** climb, toss
canyon: desert, harbor	**afraid:** angry, sad	**scoop:** crush, shred
straw: basket, knot		**spread:** lick, stack
toad: goose, cricket		**twist:** attach, knit
alphabet (letters): numbers, colors		**leap:** gallop, stretch
collar: shawl, apron		**weep:** stumble, shiver
pond: forest, field		
tool (hammer): corn, helmet		
beak: paw, fin		
furniture (couch): plate, closet		
porch: factory, school		
moose: lizard, snail		
barrel: lantern, pebble		
heel: forehead, wrist		
sailor: pirate, soldier		
beetle: buffalo, canary		
pepper: lemon, nut		
flute: violin, tambourine		
cork: baton, kettle		

Note: The target words are in boldface type.

verbs (e.g., *limp / creep / shrug*) with a single actor portraying the body movement, and four verb sets were transitive, with an actor performing an action on an object (e.g., *scoop* rice / *crush* paper cup / *shred* paper). None of the nouns, adjectives, or verbs appeared on the PPVT-4. One word from each noun, adjective, and verb set was randomly assigned to be the target word, and two were the distracters. All target and distracter items are presented in Table 1.

On the slideshow nouns and adjectives are represented by color photographs downloaded from various online photo archives; within each set, photos were matched for visual salience as much as possible. For each verb, we videotaped the same female actor performing the action against a white background, with a similar pace for each of the three actions depicted. On each trial, the three images in a set are displayed with one image centered at the top of the array and two items at the bottom to the right and left of the center (see Figure 1). The position of the target word in the array (top, bottom left, or bottom right) was counterbalanced across the 33 trials. Word class (nouns, verbs, and adjectives) and the relative comprehension level of each word set were randomized across trials. Although some IPL studies of word comprehension test both words in a paired display, with each item serving as a target and a distracter on different trials within or across participants (e.g., Bergelson & Swingley, 2012; Houston-Price et al., 2007), the goal of the present study was to test all children in the sample on the same set of

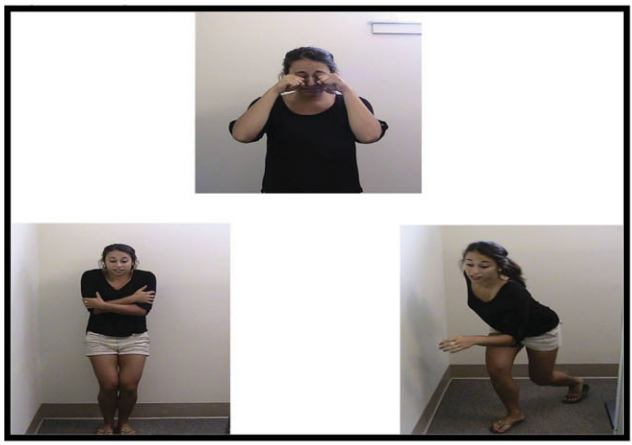
https://doi.org/10.1017/S0142716415000107 Published online by Cambridge University Press



Noun target: *canyon*



Adjective target: *hollow*



Verb target: *weep*

Figure 1. (Color online) Examples of noun, adjective, and verb triads.

https://doi.org/10.1017/S0142716415000107 Published online by Cambridge University Press

33 target words, 33 target items, and 66 distracters, in much the same way that the PPVT-4 presents a unique display of four items for each target word.

The audio accompanying each trial was digitally recorded using an MXL condenser microphone and Dell laptop computer. The same female voice was recorded for all audio material, using the characteristic varied intonation of child-directed speech. Baseline trials were accompanied by "Wow look up here" or "Hey what do you see?" randomly assigned across trials. Nouns were introduced during the labeling trial using the sentence frame "Look at the *x*" as in "Look at the *crab*." Adjectives occurred with "Where is she/it *x*?" as in "Where is she *afraid*?" or "Where is it *sharp*?" Verbs were introduced with "Look at her *x*" as in "Look at her *limp*." Each noun, adjective, and verb was also recorded as a single word utterance for presentation during the test trial. Tobii Studio software was used to create an audio–video slideshow with the timing and sequence of presentation described below.

For administration of the PPVT-4 (Dunn & Dunn, 2007), the easel test booklet for Form A and the corresponding recording forms were used with all participants.

Procedure

During a 10-min warm-up period, parents completed a background information form while children played with a puzzle or crayons and paper and interacted with the experimenter. Children were randomly assigned to begin testing with the IPL or the PPVT-4. For the IPL testing, children were seated 26 inches in front of a Tobii T60XL eye-tracker computer monitor with a 24-in. screen. The testing session began with a brief calibration. We tested 33 words in five blocks of six trials and a final block of three trials with a 3-s animation (e.g., a marching Mickey Mouse) between blocks. Each block consisted of a random mix of noun, adjective, and verb trials.

Each word was tested with a baseline, label, and test phase. Each baseline phase began with a twinkling red star centered on a blank blue background (1 s) designed to elicit attention to the middle of the screen, followed by a set of one target and two distracter images. Noun and adjective sets appeared for 3 s; verb sets appeared for 5 s to provide additional time to view the three dynamic images. The baseline phase was accompanied by an auditory prompt ("Wow look up here" or "Hey what do you see?"). During the 2-s label phase, the twinkling red star on a blank blue background reappeared, accompanied by audio presentation of the target word in its sentence frame (e.g., "Look at the crab"). During the final test phase, the set of one target and two distracter images in the same orientation reappeared for 3 s and the target word was repeated (e.g., *crab*). The total time for the slideshow with animations between blocks was 5 min 28 s. The sequence and duration of visual fixations to the target and distracter images during baseline and test trials were automatically recorded by the Tobii eye tracker.

For the PPVT-4 testing, child and examiner sat at a small table, with the child facing the PPVT-4 easel test booklet. Children were tested using the standard procedures for administration, beginning with training items and following instructions for establishing the basal and ceiling sets. Administration of the PPVT-4 typically required 10–15 min to complete.

https://doi.org/10.1017/S0142716415000107 Published online by Cambridge University Press

Measures

IPL trials with no visual attention recorded during baseline and/or test trials were eliminated from the analyses. There were 35 children with a full set of 33 trials; 12 with 32 trials; 6 with 31 trials; 3 with 30 trials; 4 with 29 trials; 3 with 26 trials; and 1 child each completing 25, 24, 23, and 22 trials. Children at 3, 4, 5, and 6 years missed 5.1%, 3.7%, 5.8%, and 4.8% of the 33 trials, respectively. Missing trials accounted for 5.1% of all noun trials, 2.9% of adjective trials, and 5.4% of verb trials.

There are several measures available from the eye-tracker recordings of children's visual attention throughout the IPL slideshow. Latency measures record the time between the onset or offset of the target word and the first fixation to the target versus the distracter item. Latency measures, along with the frequency of fixations, have typically been used in looking while listening tasks that examine the time course of processing linguistic input (e.g., Fernald et al., 2008). However, most of the IPL studies investigating word comprehension across a set of vocabulary items (e.g., Bergelson & Swingley, 2012; Houston-Price et al., 2007; Meints et al., 1999, 2008; Reznick, 1990; Schafer & Plunkett, 1998; Styles & Plunkett, 2009; Tincoff & Jusczyk, 1999) have utilized one or both of the following measures: (a) the relative proportion of visual attention to the target versus the distracter during the baseline and test trials, and (b) the longest look to the target versus the distracter during the baseline and test trials. These measures make optimal use of the time available to the child for inspection of the displays. Moreover, similar patterns of word comprehension have been found for these two measures by Meints et al. (1999, 2008), by Schafer and Plunkett (1998), and by Styles and Plunkett (2009) in IPL studies with children age 12 to 36 months of age. In the present study, we extend these two measures to examine comprehension of the 33 vocabulary items by preschool children.

Looking time to target (LTT) is the percentage of time a child looked at the target image (versus the two distracter images) calculated as total visual fixation time to the target image divided by the total visual fixation time to the target and the two distracter images. Mean LTT scores during baseline and test trials were calculated for each child for the 21 nouns, for the 8 verbs, and for the 4 adjectives. Our second dependent variable is based on the longest look (LL) to the target versus the two distracters. We created an LL difference score, calculated as LL to target minus the sum of LLs to the two distracter items. Mean LL scores during baseline and test trials were calculated for the 21 nouns and the 4 adjectives, which had equivalent 3-s baseline and test trials, but not for the 8 verbs, which had longer baseline (5 s) than test trials (3 s).

For the PPVT-4 measure of word comprehension, we calculated a total raw score for each child following the test guidelines in which the total number of errors is subtracted from the final item of the ceiling set.

RESULTS

We first consider whether the IPL methodology, with a format designed for the capabilities and attention span of children during their first 2 years, can be

https://doi.org/10.1017/S0142716415000107 Published online by Cambridge University Press

Table 2. *Mean (standard deviation) percentage looking time to target scores for IPL and PPVT Raw Scores*

	IPL						
	Nouns		Adjectives		Verbs		PPVT
Age	Baseline	Test	Baseline	Test	Baseline	Test	Raw Score
3	0.31 (0.04)	0.52 (0.09)	0.28 (0.09)	0.34 (0.15)	0.36 (0.08)	0.46 (0.11)	89.2 (24.1)
4	0.32 (0.02)	0.56 (0.07)	0.31 (0.07)	0.47 (0.11)	0.33 (0.07)	0.50 (0.10)	103.2 (17.1)
5	0.31 (0.02)	0.62 (0.09)	0.29 (0.07)	0.47 (0.16)	0.31 (0.08)	0.54 (0.09)	117.2 (20.6)
6	0.32 (0.05)	0.71 (0.10)	0.27 (0.06)	0.63 (0.17)	0.32 (0.04)	0.59 (0.10)	124.0 (11.7)

Note: IPL, Intermodal preferential looking task; PPVT, Peabody Picture Vocabulary Test.

appropriately extended to children age 3 to 6 years. As a measure of engagement with the task, we calculated the percentage of time children looked at the target and distracter items during the time available for their inspection during the slideshow: 3-s baseline for nouns and adjectives; 5-s baseline for verbs; 3-s test for all words. We found that, overall, children attended nearly equally to the stimuli during the baseline (76%) and test (77%) trials. Moreover, this pattern was evident at all ages, with engagement for 3-, 4-, 5-, and 6-year-olds at 77%, 76%, 72%, and 78%, respectively, during baseline trials and 80%, 77%, 74%, and 78%, respectively, during test trials.

To examine word comprehension, we compare visual attention to the target versus the distracters during baseline and test trials using two measures previously described, LTT and LL. Our *t* tests revealed no sex differences for either measure on any word class, so scores for males and females are combined in the following analyses. Means and standard deviations for LTT scores during baseline and test trials for nouns, adjectives, and verbs at each age and for PPVT-4 raw scores at each age are presented in Table 2. To examine change in visual attention to the target image as a function of trial, word class, and age, we conducted a $2 \times 3 \times 4$ mixed design analysis of variance (ANOVA) on arc sine transformations of the percentage scores. There were two levels of trial (baseline vs. test), three levels of word class (nouns, verbs, and adjectives), and four levels of age (3, 4, 5, and 6 years). Trial and word class are within factors; age is a between factor.

To infer comprehension on the IPL, we expect an increase in visual attention to the target image during the test trial compared to baseline presentation. Across all ages and the three word classes, mean LTT scores indicate that children looked at the target image during the baseline trial about 32% of the time, which is what we might expect if children, on average, distribute their attention across the three images before the target is labeled. During the test trial, attention to the target image nearly doubles to 58%. The ANOVA analysis indicates a main effect of trial, $F(1, 63) = 323.7$, $p < .001$, with an effect size of $\eta^2 = 0.83$. Thus, across the four age groups, there is evidence for overall comprehension of the 33 words tested on the IPL.

https://doi.org/10.1017/S0142716415000107 Published online by Cambridge University Press

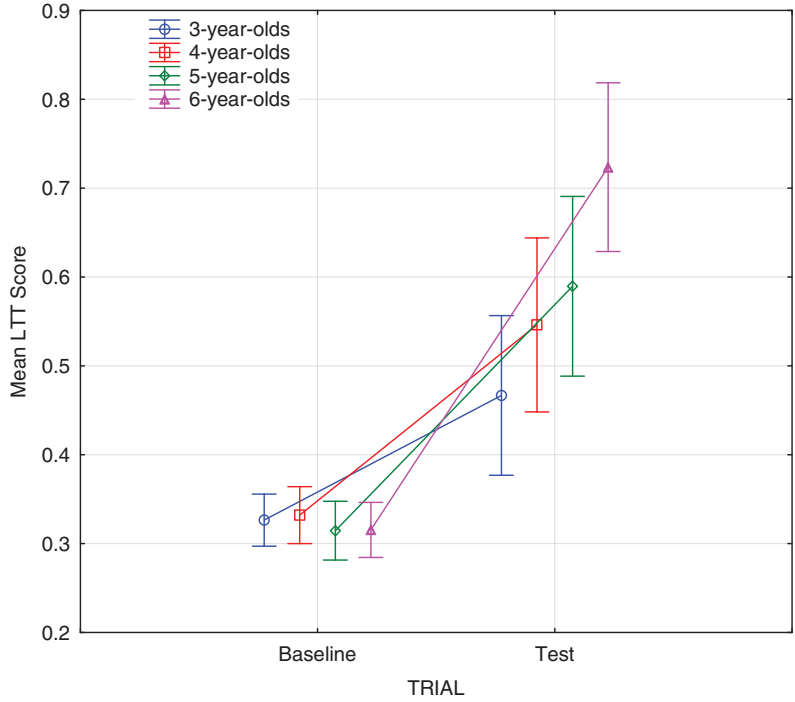


Figure 2. (Color online) Mean percentage of scores at baseline and test for each age.

There is also a main effect of age, F (3, 63) = 12.7, $p < .001$, with an effect size of $\eta^2 = 0.37$. Across baseline and test trials, LTT scores are lowest for the 3-year-olds ($M = 0.39$) and increase with age for children at 4 ($M = 0.43$), 5 ($M = 0.45$), and 6 ($M = 0.51$) years. However, as can be seen in Table 2 and in Figure 2, baseline scores are similar across the four age groups, whereas test trial scores increase differentially with age. This is confirmed by a Trial × Age interaction, F (3, 63) = 16.4, $p < .001$, with an effect size of $\eta^2 = 0.43$, and by post hoc analyses. Scheffe tests indicate no age differences during the baseline trials ($p > .05$). During the test trials, visual attention is higher for the 6-year-olds when compared to 3- ($p < .001$), 4- ($p < .001$), and 5- ($p < .01$) year-olds. LTT test scores are also higher at 5 years compared to 3 years ($p < .05$); there are no differences for the 3- versus 4-year-old and 4- versus 5-year-old comparisons ($p > .05$). Moreover, at all ages, LTT scores are higher at test compared to baseline (at 3 years, $p < .01$; at 4, 5, and 6 years, $p < .001$), indicating overall word comprehension for each age group.

There is a main effect of word class, F (2, 126) = 25.1, $p < .001$, with an effect size of $\eta^2 = 0.28$. Across all ages, children's attention to the target image during the combined baseline and test trials is highest for nouns ($M = 0.49$), followed by verbs ($M = 0.45$) and adjectives ($M = 0.40$). However, as can be seen in Table 2

https://doi.org/10.1017/S0142716415000107 Published online by Cambridge University Press

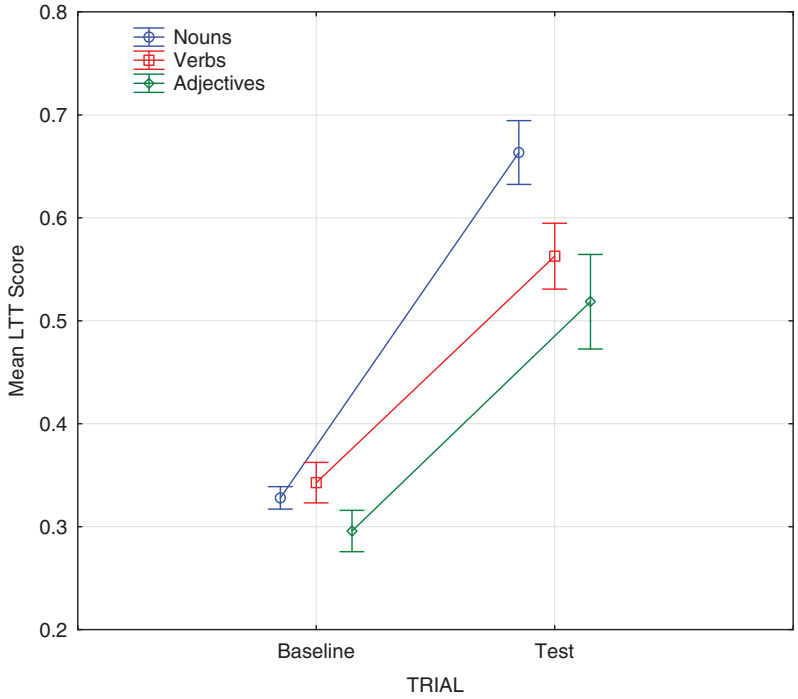


Figure 3. (Color online) Mean percentage of looking time to target scores at baseline and test for nouns, adjectives, and verbs.

and Figure 3, LTT scores for each word class are similar during baseline trials but increase differentially during the test trial. This is confirmed by a Trial × Word Class interaction, F (2, 126) = 14.3, $p < .001$, with an effect size of $\eta^2 = 0.18$, and by post hoc analyses. Scheffe tests indicate no word class differences during baseline trials ($p > .05$); during test trials, LTT scores for nouns are higher than for verbs or adjectives ($p < .001$), but verb versus adjective scores do not differ ($p > .05$). LTT scores are higher at test than at baseline for all three word classes ($p < .001$), suggesting that, across the four age groups, children comprehend nouns, verbs, and adjectives. Finally, word class does not interact with age, F (6, 126) = 2.04, $p > .05$, and there is no three-way interaction among trial, word class, and age, F (6, 126) = 1.4, $p > .10$.

A similar pattern emerged for word comprehension using a measure of LL to the target during baseline versus test trials. Means and standard deviations for LL difference scores during baseline and test trials for nouns and adjectives at each age are presented in Table 3. Note that baseline scores represent the longest look to the target image minus the sum of the longest looks to the two distracter images, and would be expected to be negative if children initially distribute their gaze across the three images. However, we would expect LL difference scores

https://doi.org/10.1017/S0142716415000107 Published online by Cambridge University Press

Table 3. *Mean (standard deviation) longest look to target difference scores (s)*

	Nouns		Adjectives	
Age	Baseline	Test	Baseline	Test
3	−0.61 (0.30)	0.06 (0.25)	−0.80 (0.45)	−0.60 (0.75)
4	−0.46 (0.20)	0.07 (0.25)	−0.50 (0.39)	−0.25 (0.36)
5	−0.43 (0.19)	0.19 (0.28)	−0.58 (0.39)	0.006 (0.40)
6	−0.64 (0.31)	0.67 (0.57)	−0.85 (0.42)	0.40 (0.71)

to increase in a positive direction during the test trial if children comprehend the target word. To examine change in visual attention to the target image as a function of trial, word class, and age, we conducted a 2 × 2 × 4 mixed design ANOVA. There were two levels of trial (baseline vs. test), two levels of word class (nouns and adjectives), and four levels of age (3, 4, 5, and 6 years). Trial and word class are within factors; age is a between factor.

The ANOVA analysis indicates a main effect of trial, F (1, 63) = 128.4, p < .001, with an effect size of η^2 = 0.67. Thus, the LL data also suggest overall comprehension of the 21 nouns and 4 adjectives across the four age groups.

There is a main effect of age, F (3, 63) = 7.6, p < .001, with an effect size of η^2 = 0.26. Across baseline and test trials, LL difference scores are lowest for the 3-year-olds (M = –0.48) and increase with age for children at 4 (M = –0.28), 5 (M = –0.20), and 6 (M = –0.10) years. However, as was the case for the LTT data, there is a Trial × Age interaction, F (3, 63) = 13.5, p < .001, with an effect size of η^2 = 0.39. Scheffe tests indicate no age differences during the baseline trials (p > .05). During the test trials, LL to the target is higher for the 6-year-olds when compared to 3- and 4 year-olds (p < .05); there are no other age differences. When we compare baseline to test for each age group, LL scores are higher at test compared to baseline for 5- (p < .05) and 6-year-olds (p < .01), but not for 4-year-olds (p > .10); 3-year-olds show a marginal difference (p < .10). Thus, the LL data reveal somewhat weaker effects than do the LTT scores, which demonstrated significantly higher scores at test compared to baseline for each age group.

There is a main effect of word class, F (1, 63) = 32.6, p < .001, with an effect size of η^2 = 0.34. Across all ages, children's LL difference scores to the target image during the combined baseline and test trials are higher for nouns (M = –0.14) compared to adjectives (M = –0.39). However, as can be seen in Table 3, LL scores are similar for nouns and adjectives during baseline trials but increase differentially during the test trial. This is confirmed by a Trial × Word Class interaction, F (1, 63) = 7.2, p < .01, with an effect size of η^2 = 0.10, and by post hoc analyses. Scheffe tests indicate no word class differences during baseline trials (p > .05); during test trials, LL scores are higher for nouns compared to adjectives (p < .001). LL scores are higher at test than at baseline for both nouns and adjectives (p < .001), suggesting that, across the four age groups, children

https://doi.org/10.1017/S0142716415000107 Published online by Cambridge University Press

Table 4. *Mean (standard deviation) percentage change scores for individual words in each word class*

Nouns		Adjectives		Verbs	
Word	*M* (*SD*)	Word	*M* (*SD*)	Word	*M* (*SD*)
Tool	0.43 (0.33)	Sharp	0.34 (0.33)	Spread	0.41 (0.32)
Coconut	0.42 (0.28)	Afraid	0.18 (0.30)	Scoop	0.35 (0.42)
Beetle	0.40 (0.34)	Hollow	0.14 (0.32)	Fold	0.33 (0.32)
Helicopter	0.38 (0.29)	Ashamed	0.11 (0.32)	Reach	0.23 (0.25)
Barrel	0.38 (0.29)			Twist	0.16 (0.36)
Flute	0.37 (0.32)			Limp	0.10 (0.26)
Alphabet	0.36 (0.32)			Leap	0.09 (0.32)
Porch	0.34 (0.30)			Weep	−0.17 (0.31)
Crab	0.31 (0.22)				
Pepper	0.31 (0.35)				
Furniture	0.30 (0.28)				
Straw	0.30 (0.33)				
Sailor	0.29 (0.30)				
Pond	0.29 (0.27)				
Heel	0.24 (0.40)				
Beak	0.22 (0.44)				
Moose	0.21 (0.33)				
Canyon	0.17 (0.31)				
Cork	0.11 (0.28)				
Collar	0.09 (0.33)				
Toad	−0.01 (0.44)				

comprehend both types of words. Finally, as was also the case for the LTT data, word class does not interact with age, $F\ (3, 63) = 1.8$, $p > .05$, and there is no three-way interaction among trial, word class, and age, $F\ (3, 63) = 1.7$, $p > .05$.

Although ANOVA analyses indicate overall comprehension of words tested on the IPL, there was considerable variation in mean scores for individual vocabulary items. Table 4 presents mean LTT change scores for individual nouns, verbs, and adjectives. LTT change scores were created by subtracting the mean LTT score during baseline from the mean LTT score during test for each of the 21 nouns, 8 verbs, and 4 adjectives. Within each word class, words have been arranged in descending order. For nouns, exemplars of semantic categories are distributed throughout this range. For example, among animal words, *beetle* is relatively easy, *crab* and *moose* more difficult, and *toad* represents the lowest score. Children scored higher on *pond* than on *canyon.* The highest scoring noun is a category name, *tool*, and children do moderately well on the two other category items, *alphabet* and *furniture*. On adjectives, children attained their highest score with *sharp* but have more difficulty with *hollow.* The facial expression depicted by *afraid* elicits higher comprehension scores than does *ashamed*. Verb score also vary considerably. In general, children do better on the four verbs that depict

https://doi.org/10.1017/S0142716415000107 Published online by Cambridge University Press

actions on objects (*spread* butter on bread, *scoop* rice, *fold* clothing, and *twist* cap on jar; $M = 0.31$) than on the four actions depicted without objects (*reach*, *limp*, *leap*, *and weep*; $M = 0.06$). It is possible that the objects provide contextual cues that enriched children's ability to interpret those actions.

Finally, we examined the relationship of children's performance on the IPL to their scores on the PPVT-4, a standardized test of word comprehension. Mean PPVT-4 raw scores for each age are displayed in Table 2. Corresponding standard scores indicate average receptive vocabulary for the 3-year-olds in this sample ($M = 99$); 4-, 5-, and 6-year-olds, however, achieved above-average standard scores of 120, 118, and 122, respectively.

To examine the relationship between the PPVT-4 and the LTT measure of comprehension, we calculated four Pearson product-moment correlation coefficients between PPVT-4 raw scores and LTT change scores (LTT during test minus LTT during baseline) for the 21 nouns, 8 verbs, 4 adjectives, and total test (all 33 words). In all four cases, there were positive, significant correlations. The strongest relationship was found for children's total PPVT-4 raw score and the mean LTT change score for the total 33 words tested on the IPL ($r = .51$, $p < .001$), followed by nouns ($r = .47$, $p < .001$), adjectives ($r = .41$, $p < .001$), and verbs ($r = .33$, $p < .01$). When controlling for age differences with partial correlations, the relationship between the PPVT-4 and the LTT change scores is attenuated, with significant correlations remaining for total words tested ($r = .26$, $p < .05$) and nouns ($r = .27$, $p < .05$) but not for verbs ($r = .10$, $p > .05$) or adjectives ($r = .15$, $p > .05$). The use of standardized PPVT-4 scores did not improve the correlations.

We also examined the relationship between the PPVT-4 and the LL measure of comprehension using LL change scores (test minus baseline). We calculated three Pearson product-moment correlation coefficients between PPVT-4 raw scores and LL change scores for the 21 nouns, 4 adjectives, and total test (25 words). Again, these correlations are positive and significant, but weaker than the corresponding LTT correlations, with relatively stronger relationships for total test ($r = .36$, $p < .01$), and adjectives ($r = .38$, $p < .01$) compared to nouns ($r = .24$, $p < .05$). However, when controlling for age with partial correlations, there are no significant relationships between the PPVT-4 and the LL measure of comprehension (nouns $r = –.01$, adjectives $r = .01$, and total test $r = .002$). Again, the use of standardized PPVT-4 scores did not improve the correlations.

DISCUSSION

The IPL task measures visual attention to a target image before and after the image is labeled. Comprehension is defined as an increase in attention to the target image during the test trial compared to baseline presentation. The IPL was developed to test various aspects of language comprehension in infants and toddlers who cannot be relied upon to follow directions, point to pictures, or act out commands. A goal of the study was to determine whether the IPL can be extended to measure word comprehension in preschool-age children. We tested 3-, 4-, 5-, and 6-year olds on 21 nouns, 8 verbs, and 4 adjectives using the IPL methodology. Nouns and adjectives were represented by triads (a target image and two distracters) of

https://doi.org/10.1017/S0142716415000107 Published online by Cambridge University Press

color photographs; verbs were depicted by triads of an actor performing actions. Children viewed each triad before (baseline) and after (test) the target was labeled. The data indicate that children at all ages were engaged by the task, with eye gaze recorded to the three stimuli in the array during approximately 75% of the time these were available for viewing during baseline and test trials.

We compared two measures of visual attention that have been frequently used in IPL studies with children under the age of 3 years. LTT scores represent the proportion of time that children look at the target image (vs. the two distracters) during the test trial versus baseline presentation. During the baseline trial, children at all ages focused on the target image about one-third of the time; after the target image was labeled, attention to the target nearly doubled, with age-related increases recorded during the test trial. Moreover, children's comprehension scores were higher for nouns compared to verbs and adjectives, a pattern that reflects developmental trends in word learning.

A similar pattern was found using a measure of the difference between the longest look directed to the target minus the longest look to the two distracters during test trials versus baseline presentation. Across the four age groups, longest looks to the target image relative to the two distracters increased significantly during the test trial compared to baseline and were higher for nouns compared to adjectives. However, specific comparisons revealed that longest looks increased from baseline to test for 5- and 6-year-olds but not for the 3- and 4-year-olds. It may be that the increased array of three stimuli used in the current IPL somewhat weakened the ability of the younger children to sustain longer looks to the stimuli during the relatively brief trial durations. It should be possible to explore this hypothesis by increasing trial duration in future research.

The LTT measure of comprehension is also favored over LL scores when we compared the relationship between children's performance on the IPL task to their scores on the PPVT-4 (Dunn & Dunn, 2007), a widely used standardized test of word comprehension. Correlations were consistently stronger for LTT versus LL measures of comprehension (including total test, nouns, and adjectives). The strongest correlation ($r = .51$) was obtained for the relationship between children's total PPVT-4 score and mean LTT score for all 33 test items on the IPL. However, when age is partialed out of the correlation, this relationship remains significant but is substantially weakened ($r = .26$). With age controlled for, none of the LL measures (total test, nouns, or adjectives) are significantly related to children's PPVT-4 score.

Overall, the data indicate that the IPL methodology can be successfully extended to measure word comprehension in older, preschool children. Most of these normally developing preschoolers maintained attention to the task throughout the 33 test trials, which included noun and adjective trials depicting static images and verb trials with dynamic stimuli. On average, they distributed their attention to each of the three images during the baseline trial and significantly increased attention to an image matching the target word during the test trial. However, for children age 3 to 6 years, the relative proportion of visual attention to the target versus the distracter across trials may be a more robust index of comprehension than the measure of longest look, at least in the case of the stimulus triads used in the current study.

https://doi.org/10.1017/S0142716415000107 Published online by Cambridge University Press

There are several important differences to note in vocabulary testing with the PPVT-4 and the IPL in the present study, which might contribute to the relatively modest correlations between the two measures. We selected the 33 words tested on the IPL from word frequency tables to represent a range of semantic categories and vocabulary levels for children aged 3 to 6 years. The test items of the PPVT-4, in contrast, have been carefully norm-referenced over the more than 50 years and several revisions that the test has been in use. A limitation of this preliminary study is that the total number of words tested is relatively small. Further research should extend the number of word trials, especially for verbs and adjectives. The IPL displays one target word and two distracters per trial, whereas the PPVT-4 depicts one target and three distracters per page. There was no overlap in the specific words tested on the two assessments. The IPL presents an automated, strictly timed sequence of baseline, label, and test trials for each word tested, whereas the PPVT-4 is untimed and administered by an examiner. The PPVT-4 items are presented in a sequence of increasing difficulty, with standardized criteria for establishing basal and ceiling items. On the IPL, children viewed a sequence of items that was randomly distributed with respect to age-appropriate vocabulary level. The child responds to the PPVT-4 with a point or by naming the number of her choice, whereas the IPL measures only visual attention. Despite these differences, and with age differences controlled, we found modest, significant correlations between LTT and PPVT-4 scores.

One advantage of the IPL is that it can be used to test verbs with ecologically valid dynamic stimuli: actors performing intransitive actions (e.g., *leap*) and actors operating on objects (e.g., *scoop*). In the current study, we tested nouns with static photos of objects and verbs with dynamic, video-recorded stimuli, and we found higher LTT scores for nouns versus verbs on the IPL test trials. Although noun comprehension precedes verb comprehension early in development, it is not clear whether the differences reported here reflect this same developmental sequence in word knowledge or more systematic procedural differences. For example, only 8 verbs were tested compared to 21 nouns. Although baseline trials for verbs were extended to 5 s, compared to 3 s for the static stimuli in noun trials, it is possible that dynamic stimuli require even longer intervals. In contrast, noun scores also exceeded scores for adjective trials, which were similarly represented by static photos. Further work will be needed to differentiate the contributions of attentional differences to the assessment of word comprehension using static versus dynamic stimuli.

When using dynamic stimuli to depict verbs, more complex scenarios are possible with multiple actors and objects (e.g., *give*, *take*, and *receive*). The static images of line drawings used to represent actions and events in standardized tests such as the PPVT-4 may be difficult for young children to interpret. Two-dimensional images of a momentary body posture must suggest an entire series of movements that unfold over time. In some cases, pictorial conventions such as motion lines used to enhance the image may be unknown or unfamiliar. Moreover, many verbs, such as those in the triad *limp*, *creep*, and *shrug* used in the present IPL assessment, may be impossible to depict in static images. Dynamic stimuli can also be used to test verbs beyond their specific, conventional meanings. For example, we tested the verb *spread* using a relatively conventional scenario of an

https://doi.org/10.1017/S0142716415000107 Published online by Cambridge University Press

actor spreading butter on bread. A less conventional depiction might be to use the verb in a non-food-related context, such as the action of spreading a deck of cards across the surface of a table. Dynamic events in the IPL methodology might also be used to assess adverbs such as *hastily* or *delicately* and locative prepositions such as *toward* or *near.*

We can begin to consider how the IPL methodology might be further extended to provide an alternative method for assessing word comprehension in children with attention, behavioral, and social limitations, whatever their chronological age. The IPL eliminates the need for the child to point to pictures or to manipulate objects and circumvents the challenges of testing children with physical limitations. The IPL also does not require social engagement with an experimenter, which hinders assessment for many children with ASD. A number of studies have used eye-tracking procedures to examine the eye gaze of children with autism and language delays in social and communicative contexts (e.g., Tenenbaum, Amso, Abar, & Sheinkopf, 2014). Moreover, there is limited evidence that the requirement to associate auditory with visual stimuli is appropriate for individuals with autism. Walker-Andrews, Haviland, Huffman, and Toci, (1994) report that children with autism were able to match events (e.g., dropping marbles into a cup) and their associated sound effects during an experimental session with the IPL, and children with autism age 27–41 months were able to match an subject–verb–object utterance to an appropriate video depiction (Swenson, Kelley, Fein, & Naigles, 2007). Reliable assessments of vocabulary comprehension in children with ASD are particularly important because comprehension tends to be relatively more impaired than expressive ability (Hudry et al., 2010).

In contrast, the IPL methodology depends on the scoring of visual attention, which poses problems for both measurement and interpretation. Eye-tracker monitoring allows for reliable and automatic calculation of visual gaze, but this is an expensive research technology that may not yet be widely available for clinical and educational use. Moreover, visual attention provides a relative measure of change rather than the absolute scoring of an item as correct or incorrect, available from the child's specific behavioral response on a standardized test. In our application of the IPL methodology to measure word comprehension in the present study, we calculated LTT change scores, which represented the change in visual attention from baseline to test trials. Clearly, negative change scores for an individual word suggest the child's lack of comprehension. A test of comprehension using the IPL methodology would be maximally useful if a criterion score were available for LTT change scores with positive values, that is, those words for which visual attention increased from baseline to test trials. For example, Reznick (1990) used the IPL to test word comprehension in children during the first 2 years and compared change scores of 5%, 10%, 15%, 20%, and 25%. He found that the 15% criterion produced the most reliable results for this age when factors such as developmental differences, individual differences, and test–retest reliability were considered. Reznick and Goldfield (1992) also used a criterion score of 15% increase in LTT from baseline to test for children assessed during the second year, and report that a spurt in the number of words comprehended overlapped with a spurt in word production. Killing and Bishop (2008) used the IPL to test word comprehension in children age 20 to 24 months and found that the number of words showing a 15%

https://doi.org/10.1017/S0142716415000107 Published online by Cambridge University Press

increase in attention from baseline to test correlated positively with parent-report measures of comprehension. It should be possible to develop similar criteria for comprehension data for preschool children.

Language comprehension is an important index of developmental status, and reliable measures are crucial for research into the onset and trajectory of word learning as well as for the practical concerns of educational planning and placement for children experiencing language delay. Standardized tests such as the PPVT-4 (Dunn & Dunn, 2007) are, for the most part, reliable and widely used for educational and clinical assessments. The IPL was developed as a research methodology to assess word comprehension in infants and toddlers who do not reliably follow directions, point to pictures, or act out commands. The present study suggests that the IPL methodology might be extended beyond research and adapted to provide an alternative tool for assessing word comprehension in older children with similar limitations.

ACKNOWLEDGMENTS

This research was supported by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under Grant 2 P20 GM103430.

REFERENCES

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language, 38*, 419–439.

Benedict, H. (1979). Early lexical development: Comprehension and production. *Journal of Child Language, 6*, 183–200.

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

Brock, J., Norbury, C., Einav, S., & Nation, K. (2008). Do individuals with autism process words in context? Evidence from language-mediated eye-movements. *Cognition, 108*, 896–904.

Carroll, J. B., Davies, P., & Richman, B. (1971). *American heritage word frequency book*. Boston: Houghton–Mifflin.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology, 6*, 84–107.

Dunn, L. M., & Dunn, D. M. (2007). *Peabody Picture Vocabulary Test* (4th ed.). Bloomington, IL: Pearson.

Fenson, L., Dale, P., Reznick, J. S., Bates, E., Thal, D., & Pethick, S. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development, 59*, 174–185.

Fernald, A., Perfors, A., & Marchman, V. A. (2006). Picking up speed in understanding: Speech processing efficiency and vocabulary growth across the second year. *Developmental Psychology, 42*, 98–116.

Fernald, A., Swingley, D., & Pinto, J. P. (2001). When half a word is enough: Infants can recognize spoken words using partial phonetic information. *Child Development, 72*, 1003–1015.

Fernald, A., Zangl, R., Portillo, A. L., & Marchman, V. A. (2008). Looking while listening: Using eye movements to monitor spoken language comprehension by infants and young children. In

https://doi.org/10.1017/S0142716415000107 Published online by Cambridge University Press

I. A. Sekerina, E. M. Fernandez, & H. Clahsen (Eds.), *Developmental psycholinguistics: On line methods in children's language processing* (pp. 97–135). Philadelphia, PA: John Benjamins.

Goldin-Meadow, S., Seligman, M. E., & Gelman, R. (1976). Language in the two-year-old. *Cognition, 4*, 189–202.

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*, 23–45.

Harris, A. J., & Jacobson, M. D. (1982). *Basic reading vocabularies*. New York: MacMillan.

Harris, M., Yeeles, C., Chasin, J., & Oakley, Y. (1995). Symmetries and asymmetries in early lexical comprehension and production. *Journal of Child Language, 22*, 1–18.

Houston-Price, C., Mather, E., & Sakkalou, E. (2007). Discrepancy between parental reports of infants' receptive vocabulary and infants' behavior in a preferential looking task. *Journal of Child Language, 34*, 701–724.

Hudry, K., Leadbitter, K., Temple, K., Sionims, V., McConachie, H., Aldred, C., et al. (2010). Preschoolers with autism show greater impairment in receptive compared with expressive language abilities. *International Journal of Language and Communication Disorders, 45*, 681–690.

Huettig, F., & Altmann, G. T. M. (2005). Word meaning and the control of eye fixation: Semantic competitor effects and the visual world paradigm. *Cognition, 96*, B23–B32.

Huettig, F., Rommers, J., & Meyer, A. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica, 137*, 151–171.

Huttenlocher, J. (1974). The origins of language comprehension. In R. L Solso (Ed.), *Theories in cognitive psychology*. Potomac, MD: Erlbaum.

Johnson, E. K., & Huettig, F. (2011). Eye movements during language-mediated visual search reveal a strong link between overt visual attention and lexical processing in 36-month-olds. *Psychological Research, 75*, 35–42.

Killing, S. E. A., & Bishop, D. V. M. (2008). Move it! Visual feedback enhances validity of preferential looking as a measure of individual differences in vocabulary in toddlers. *Developmental Science, 11*, 525–530.

Meints, K., Plunkett, K., & Harris, P. L. (1999). When does an ostrich become a bird: The role of typicality in early word comprehension. *Developmental Psychology, 35*, 1072–1078.

Meints, K., Plunkett, K., & Harris, P. L. (2008). Eating apples and houseplants: Typicality constraints on thematic roles in early verb learning. *Language and Cognitive Processes, 23*, 434–463.

Mullen, E. M. (1995). *The Mullen Scales of Early Learning*. Bloomington, IL: Pearson.

Reznick, J. S. (1990). Visual preference as a test of infant word comprehension. *Applied Psycholinguistics, 11*, 145–166.

Reznick, J. S., & Goldfield, B. A. (1992). Rapid change in lexical development in comprehension and production. *Developmental Psychology, 28*, 406–413.

Richards, J. E., & Anderson, D. R. (2004). Attentional inertia in children's extended looking at television. *Advances in Child Development and Behavior, 32*, 163–212.

Schafer, G. (2005). Infants can learn decontextualized words before their first birthday. *Child Development, 76*, 87–96.

Schafer, G., & Plunkett, K. (1998). Rapid word learning by 15-month-olds under tightly controlled conditions. *Child Development, 69*, 309–320.

Spelke, E. (1979). Perceiving bimodally specified events in infancy. *Developmental Psychology, 15*, 626–636.

Styles, S., & Plunkett, K. (2009). What is "word understanding" for the parent of a one-year-old? Matching the difficulty of a lexical comprehension task to a parental CDI report. *Journal of Child Language, 36*, 895–908.

Swenson, 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*, 542–557.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268*, 1632–1634.

https://doi.org/10.1017/S0142716415000107 Published online by Cambridge University Press

Tenenbaum, E. J., Amso, D., Abar, B., & Sheinkopf, S. J. (2014). Attention and word learning in autistic, language delayed, and typically developing children. *Frontiers in Psychology, 5*, Article 490.

Thomas, D. G., Campos, J. J., Shucard, D. W., Ramsay, D. S., & Shucard, J. (1981). Semantic comprehension in infancy: A signal detection analysis. *Child Development, 52*, 798–803.

Tincoff, R., & Jusczyk, P. W. (1999). Some beginnings of word comprehension in 6-month-olds. *Psychological Sciences, 10*, 172–174.

Tincoff, R., & Jusczyk, P. W. (2012). Six-month-olds comprehend words that refer to parts of the body. *Infancy, 17*, 432–444.

Walker-Andrews, A. S., Haviland, J. M., Huffman, L., & Toci, L. (1994). Preferential looking in intermodal perception by children with autism. *Journal of Autism and Developmental Disorders, 24*, 99–107.

Zeno, S. M., Ivens, S. H., Millard, R. T., & Duvvuri, R. (1995). *Educators' word frequency guide*. New York: Touchstone Applied Science Associates.

Zimmerman, I. L., Steiner, V. G., & Pond, R. E. (2002). *Preschool Language Scale—4*. Bloomington, IN: Pearson.

https://doi.org/10.1017/S0142716415000107 Published online by Cambridge University Press