6-23-2023

Assessing the Language of 2-year-olds: From Theory to Practice

Emily Jackson
University of Connecticut - Storrs

Dani Levine
University of Chicago

Jill de Villiers
Smith College, jdevilli@smith.edu

Aquiles Iglesias
University of Delaware

Kathy Hirsh-Pasek
Temple University

See next page for additional authors

Assessing the language of 2 year-olds: From theory to practice

Emily Jackson[1] | **Dani Levine**[2] | **Jill de Villiers**[3] | **Aquiles Iglesias**[4] | **Kathy Hirsh-Pasek**[5] | **Roberta Michnick Golinkoff**[4]

[1]University of Connecticut, Storrs, Connecticut, USA

[2]University of Chicago, Chicago, Illinois, USA

[3]Smith College, Northampton, Massachusetts, USA

[4]University of Delaware, Newark, Delaware, USA

[5]Temple University, Philadelphia, Pennsylvania, USA

Correspondence
Emily Jackson.
Email: emily.k.jackson@uconn.edu

Funding information
Institute of Education Sciences, Grant/Award Numbers: R305A110284, R324A160241

Abstract

Early screening for language problems is a priority given the importance of language for success in school and interpersonal relationships. The paucity of reliable behavioral instruments for this age group prompted the development of a new touchscreen language screener for 2-year-olds that relies on language comprehension. Developmental literature guided selection of age-appropriate markers of language disorder risk that are culturally and dialectally neutral and could be reliably assessed. Items extend beyond products of linguistic knowledge (vocabulary and syntax) and tap the *process* by which children learn language, also known as fast mapping. After piloting an extensive set of items (139), two phases of testing with over 500 children aged 2; 0–2; 11 were conducted to choose the final 40-item set. Rasch analysis was used to select the best fitting and least redundant items. Norms were created based on 270 children. Sufficient test-retest reliability, Cronbach's alpha, and convergent validity with the MB-CDI and PPVT are reported. This quick behavioral measure of language capabilities could support research studies and facilitate the early detection of language problems.

1 | INTRODUCTION

Early language skills form a critical basis for cognitive growth (Bower et al., 2020; Slusser et al., 2019) as well as social development (Astington & Jenkins, 1999; Levinson et al., 2020). However, language

growth is enormously variable across children (Fernald & Marchman, 2011; Huttenlocher et al., 2010). A brief efficient assessment of early language skills would be valuable not only for researchers, but also for practitioners such as early childcare providers, speech-language pathologists, and pediatricians. These practitioners have the opportunity to use early language screenings to identify and provide support for young children whose language status suggests a possible disorder (Larson, 2016). There is mounting evidence that early language intervention is likely to lead to better outcomes for children (Roberts & Kaiser, 2015; Wake et al., 2011). Some researchers (Larson, 2016) have suggested that universal language screening for young children would allow for timely identification of language disorders, hearing loss, autism spectrum disorders (ASD), and genetic disorders (Kaiser et al., 2022). One such successful screening initiative, Universal Newborn Hearing Screening (UNHS) was implemented in several states and led to positive effects on the age of diagnosis, referral, and enrollment in Early Intervention, and hearing aid fitting or cochlear implantation (Halpin et al., 2010). Identifying these types of disorders in infancy and early childhood allows for specialized services during the key period before school entry. In the United States, children under the age of three with developmental delays or disabilities are entitled to Early Intervention services through Part C of the Individuals with Disabilities Education Act (IDEA) (Lipkin et al., 2015).

However, the value of early screening for language has been called into question. The US Preventive Service Task Force (Siu, 2015, p. 1) wrote that "current evidence is insufficient to assess the balance of benefits and harms of screening for speech delay and disorders in children aged 5 years or younger." We challenge the notion that language screeners should be used only for children already suspected to be at risk. It is important for parents and other professionals to determine that children are developing on target for their age. Wider screening may identify children with mild or moder- ate speech and language disorders that would have languished without support otherwise (Kaiser et al., 2022).

For 2-year-olds, there are many fewer options for language screening than for older children. Multiple omnibus screening tools (e.g., ASQ, CDI, PEDS) or screeners for language (e.g., MB-CDI, PLS-5 screener) have been developed (Bricker et al., 1999; Fenson, 2007; Glascoe, 1997; Ireton & Ireton, 1992; Zimmerman et al., 2011). Despite their popularity, meta-analyses suggest that most of these tools have relatively poor *predictive* value for later outcomes in language or schooling (Sim et al., 2019). In particular, early screeners may over-select a group called "late talkers", a sizable proportion of whom resolve their difficulties by age four or five (Dollaghan, 2013). That is, even if a group of children can be identified who are below their peers in language production, it does not mean that they will have a language disorder at an older age. Yet, reported comprehension difficulties appear to be more persistent. For this reason, Leonard (2014) suggested that the field could benefit from developing measures of comprehension in the toddler years. Often, a referral is made when caregivers begin to notice delays in their toddler's expressive language development. Yet, we may be missing signs that present before a young child demonstrates a delay in expressive language.

A pair of new language screeners for children between 3 and 7 years of age that assess language comprehension via a touchscreen provided the impetus for the screener presented in this paper. The Quick Interactive Language Screener or QUILS (Golinkoff et al., 2017) is for English-reared children and the QUILS:ES for Spanish-English bilingual children (Iglesias et al., 2021). While the overall structure of these screeners are the same as the Quick Interactive Language Screener: Toddlers (QUILS:TOD), new items and new ways to keep the interest of such young children were required.

TABLE 1 Comparison of the QUILS:TOD test with current language assessments.

Test	Assesses language comprehension	Dynamic stimuli	Assesses vocabulary and syntax	Assesses Product and process	Takes <30 min	Can be given by teacher or aide	Lowest age of administration	Requires touch screen response
MB-CDI[a]	Yes	No	Yes	No	Yes	No	0;8	No
PPVT-5	Yes	No	No	No	Yes	No	2;6	No
CELF-P3	No	No	Yes	No	No	No	3;0	No
WJ-3	No	No	No	No	No	No	2;0	No
MULLEN	No	No	No	No	No	No	0	No
PLS-5	No	No	Yes	No	No	No	0	No
WPPSI-3	No	No	No	No	No	No	2;6	No
QUILS:TOD	Yes	Yes	Yes	Yes	Yes	Yes	2;0	Yes

Note: Note that global screening tools such as the *ASQ* (Ages and Stages Questionnaire) and *PEDS* (Parents' Evaluation of Developmental Status are not included).

Abbreviations: CELF-P3, Clinical Evaluation of Language Fundamentals® Preschool-Third Edition; MB-CDI, MacArthur-Bates Communicative Development Inventory; Mullen, Mullen Scales of Early Learning; PLS-5, Preschool Language Scales Fifth Edition; PPVT-5, Peabody Picture Vocabulary Test-Fifth Edition; QUILS:TOD, Quick Interactive Language Screener: Toddlers; WJ-3, Woodcock Johnson Test of Achievement-3; WPPSI-3, Wechsler Preschool and Primary Scale of Intelligence Third Edition.

[a]The MB-CDI is a parent report instrument.

1.1 | The design of the QUILS:TOD Language Screener

Table 1 provides a comparison of our design decisions with the properties of existing instruments. For one, the QUILS:TOD taps into language comprehension and not parent-report measures or picture-pointing tasks that children lose interest in. The MB-CDI was a boon to researchers, but it may be more difficult for parents to judge comprehension (Conti-Ramsden & Durkin, 2012; Xue et al., 2022). As language is embedded in interaction, children can use tone of voice, routines, and context to infer what is being said, rather than relying on the language per se (Chapman, 1978; Shatz, 1978). Nonethe- less, the comprehension indices on measures such as the MB-CDI (Fenson, 2007) have proved more predictive of long-term language problems than the expressive indices (Leonard, 2014; Paul, 1996).

Assessing language comprehension rather than production has many advantages, based on existing research and practical considerations. Production tasks create particular social *demands*, for shy children or for children with little experience with conventional testing situations. In comparison, the QUILS:TOD presents items as a game on a touchscreen that increases compliance by reducing social demands. Furthermore, comprehension avoids the problem of the examiner understanding the child's often idiosyncratic speech; wide variation exists in the acquisition of consonant sounds at age 2 (Crowe & McLeod, 2020). Furthermore, *judging* the linguistic significance or adequacy of a verbal response takes advanced training.

Of course, comprehension tests come with their own special demands. The child must be interested enough to analyze the pictured material in light of the question being asked. The foils must represent suitable contrasts that the child can recognize as potential answers, and also include likely errors that children might make based on existing research. It was therefore essential to use information from existing developmental science to design the subtests with precision.

No test has used dynamic stimuli. An advantage of the use of touchscreen tablets is the possibility of portraying verbs and events, requiring children to make fewer inferences from static pictures. However, it was necessary to work around the problem that 2-year-olds might not be able to attend to

two simultaneous moving stimuli to make their choice. Therefore, at the decision point for the child's response, the images became static.

As on some other tests, for example, MB-CDI, both vocabulary and syntax are assessed. Even at 2 years of age, children's knowledge of language incorporates more than words, but also important areas of syntax. The use of dynamic stimuli enabled the testing of several syntactic contrasts such as *word order*, as in reversible transitive sentences (the car bumped the truck vs. the truck bumped the car) (de Villiers & de Villiers, 1973; Hirsh-Pasek & Golinkoff, 1999) and subject versus object wh-questions (as in "Who hit the boy?" vs. "Who did the boy hit?") (Seidl et al., 2003)., Similarly, negation requires attention to word order and to the modifier (e.g., "He has *no* shoes") rather than simply attending to the final word to understand the meaning. A touchscreen presentation could test for other forms of syntactic knowledge such as reversible prepositions that required children's understanding of word order through descriptions with spatial prepositions as modifiers. A presentation of two plausible depictions ("The book is on the cup" vs. "The cup is on the book") requires that children understand the preposition "on" to choose the correct relation between the nouns.

Most importantly, given dialect variation in the obligatoriness of early morphemes—(possessive/s/, plural/s/, s third person/s/, and past tense -ed), we did not test for contrasts that would be biased against many speakers of American dialects such as African American English (Seymour et al., 1998). Thus, our assessment focused on products of learning that develop during this period and could be relia- bly assessed via the comprehension of nouns, verbs, adjectives, negation, reversible prepositions, Wh-questions, and reversible transitive sentences.

For vocabulary, prepositions were also used as two-year-old children demonstrate an understanding of early prepositions such as "in" and "on" while other prepositions such as "between" or "in front" are not well mastered until age 3 (Golinkoff et al., 2017). The testing of verb vocabulary was enhanced by the use of dynamic stimuli; nouns and adjectives were tested with static stimuli.

Measuring process. As with the two QUILS assessments for preschoolers, we argued that the products of learning are highly influenced by children's linguistic, cultural, and socioeconomic environments (de Villiers & Johnson, 2007; Kochanoff et al., 2003). Assessments that focus solely on products would fail to distinguish between children who experience language learning difficulties and those who simply have had quantitatively and qualitatively different exposure to language. Here we capitalized on the fact that typically developing children can rapidly learn new word meanings ("fast mapping": Carey & Bartlett, 1978) and extend those words in previously unattested contexts, such as a duck in a pond to a rubber duck in the bathtub. Young children can identify the referent for a new noun when the other objects to choose already have names in their vocabulary (Golinkoff et al., 1992; Rice et al., 2000) and even extend the novel noun in the presence of another unknown object and two other known choices. They can also learn and extend novel verbs (Golinkoff et al., 1996; Akhtar & Tomasello, 1997; Fisher, 2002), and novel adjectives (Syrett et al., 2014; Waxman & Hatch, 1992) after a few exposures.

Thus, the ability to learn new words from linguistic context was considered fruitful to include in our screener based on prior experimental research (Carey, 2010; Fisher, 2002; Naigles, 1990). In each case, we assessed whether the child could learn a new item from context, and then whether they could generalize or transfer the new form to a new context. Generalization is the hallmark of language learning right from the beginning: children do not learn formulaic expressions but can hear a new item once and use it rapidly in new ways.

Any instrument that tested children directly (unlike the MB-CDI) needed to be short and engaging to young children. Specifically, the goal was to identify items that can flag children at risk for a language disorder by creating a short screener in which the children are active participants. In addition, to engage in the universal screening some have called for (e.g., Kaiser et al., 2022), the screener

should be able to be delivered by a person without special professional training given the cost of professional administration. The properties that made this possible with the QUILS:TOD were that it had (a) automatic and standard narration; (b) automatic registration of responses; (c) automatic scoring; and (d) automatic reporting of results. Designed specially to target the age range of 2; 0–2; 11 as an important but neglected time representing the beginning of grammar as well as words, the QUILSTOD is given at an age at which early potential delays can be detected and intervention can be instrumental in getting children on track (Lipkin, et al., 2020).

Finally, despite the rise in touchscreen technology usage by young children, little formal research has been conducted on 2-year-olds. The breakthrough work of Friend and colleagues (Friend et al., 2018; Friend & Keplinger, 2003) showed that touchscreens could work for such young children. However, the comprehension tools created so far by that team are restricted to vocabulary though toddlers are embarked on more than word learning; the comprehension of early grammar needs to be included. Our first goal then was to determine whether children could successfully perform our intended tasks on a tablet and, if not, whether we could assist them to be successful. We refined the methodology now available for touchscreens to test whether 2-year-olds could respond reliably, because the advantages of automated delivery and scoring are enormous.

1.2 | Stages of development for the QUILS:TOD

Our assessment was developed in four stages across 4 years. We began with initial piloting of the original design to determine the feasibility of using a touchscreen with young children to item development, item tryouts, and the final creation of the QUILS:TOD. Each stage was informed by direct work with children in the target age group and consultation with experts in the field.

2 | METHODS

The steps below contain information about the participants at each stage of the project. Participants included children between the ages of 2; 0 and 2; 11 and their caregivers. All procedures involving human subjects in this study were approved by the Institutional Review Boards at each of the universities where the project was conducted. The present study was conducted according to guidelines laid down in the Declaration of Helsinki, with written informed consent obtained from a parent or guardian for each child before any assessment or data collection.

2.1 | Feasibility of using a touchscreen and modifications for 2-year-olds

The goal of the QUILS:TOD is to have a self-contained, automatic screening measure, with pre-recorded prompts. For the QUILS:TOD, all voiceovers were recorded by a female speaker with a lively, child-friendly voice and regionally neutral American accent. The development of QUILS and QUILS:ES had revealed that children between 3 and 5 years could readily use a touchscreen to respond to multiple choice items on the screen from an automated narration, but we did not initially know if 2-year-olds could do this.

One problem was providing a clear indication to the child that it was time to make a choice, without adult prompting. Our response to this issue was to use a red circle on each of the choice options signaling the end of the prompt and serving as a visual cue for the child to respond. After a child's

FIGURE 1 Sample QUILS: TOD administration setup.

response, a yellow border surrounded each option and a reinforcement sound played (e.g., a whistle or a drumbeat) regardless of whether the child answered correctly.

During this first step, we also determined allowable response time limits and feedback frequency. Between each subtest (between 2 and four items on the final screener), children viewed short, animated scenes (such as a giraffe flying a plane) and were presented with verbal encouragement (e.g., "Great job! Let's do some more!").

Additionally, we compared children's willingness to touch the screen to indicate their answer versus their ability to drag an object on the screen to place it in a requested position in relation to another object. No significant differences were found between children's accuracy on touch response and drag response items, suggesting that dragging and dropping an object on the screen could act as an effective tool for testing linguistic contrasts that are more difficult to convey with static images. Note that in the case of drag items, the choice of action is less constrained even with a small number of objects depicted. Both the touch and the drag responses get coded and scored automatically by the program.

Figure 1 shows a two-year-old responding on a tablet. The child was seated at a low table on which the tablet rested, usually propped at an angle for ease of view. In all cases, the examiner was seated adjacent to the child, helping to maintain the child's focus if it strayed. If the child embarked on a conversation with the examiner, or needed something, the test could be paused and restarted at the same point. In no case after the initial training did the examiner assist the child in choosing an answer. On occasion, a parent or teacher was seated in the room, or another tester, but they also were instructed not to intervene in the testing.

2.2 | Item selection and elicitation process

The second step focused on deciding which specific items to use and how to elicit a response from the child. Given what we know about children's language skills at age 2 and our focus on comprehen- sion, we determined that our tool needed to assess: Product Vocabulary (Nouns, Verbs, Adjectives), Product Syntax (Negation, Wh-Questions, Reversible Transitive Sentences, Reversible Prepositions), and Process Vocabulary (Noun Learning, Adjective Learning, and Verb Learning) (see Table 2).

TABLE 2 Distribution of QUILS: TOD subtests by area.

Vocabulary	Syntax	Process
Nouns	Wh-questions	Noun learning
Verbs	Negation	Verb learning
Adjectives	Reversible transitives	Adjective learning
	Reversible prepositions	

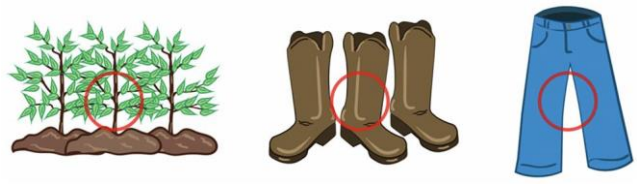

FIGURE 2 Noun Item 1. "Find the pants".

2.2.1 | Items to test for the products of vocabulary learning

Nouns. Young children's vocabularies generally contain a large number of nouns, particularly concrete nouns whose referents are tangible (e.g., ball) (Fenson et al., 1994; Gentner et al., 2001). The noun vocabularies of typically developing children expand rapidly during the first 2 years and beyond and provide a foundation for learning other word classes (Gleitman et al., 2005). In contrast, children who have small noun vocabularies prior to age 3 tend to continue to lag behind their peers in vocab- ulary knowledge throughout childhood, adolescence, and young adulthood (Rescorla, 2011; Rice & Hoffman, 2015). Friend et al. (2018) found that testing receptive vocabulary in the second year of life predicted vocabulary comprehension and kindergarten readiness at age four, and was a stronger predictor than parent reports.

Nouns for piloting were selected from lexical databases and other published resources and reduced throughout the testing phases. For each noun item, a target noun is presented with two foils—three picture choices in all—on the screen (see Figure 2). The narration instructs children to "Find the [target noun]," and children touch the screen to select their answer. Two foils were selected for each target noun based on research on lexical development (e.g., Golinkoff et al., 1995; Markman, 1989). Foils were chosen based on thematic, conceptual, or phonological characteristics. Since there were only two foils per target word, they did not come from all three categories. An example is seen in Figure 2.

Verbs. Verbs refer to actions and states—such as *move* and *think*—and form the main part of the predicate of a sentence (Golinkoff & Hirsh-Pasek, 2006). They label aspects of the world that are dynamic and could be transitory. Verbs tend to enter young children's lexicons later than nouns in many languages, but typically developing children nonetheless comprehend many verbs by the second year (Bornstein et al., 2004).

Much like nouns, the imageability, or ability to readily generate a visual image of a verb's meaning, is an important factor determining the age of acquisition of verbs (Ma et al., 2009). The verb subtests required animation as young children have difficulty interpreting statically represented action events (Cocking & McHale, 1981; Friedman & Stevenson, 1975). The items chosen were verbs (in present progressive verb form) that name actions (e.g., *washing*) pictured with common nouns. Verbs were represented by short, animated, video clips to capture their dynamic nature. Each verb was paired with

FIGURE 3 Adjective Item 26. "Which one is cold? Point to cold".

TABLE 3 Vocabulary types and items overview.

Type	Sample item
Nouns	Find the pants.
Verbs	Find: She is marching.
Adjectives	Which one is cold? Point to cold.

one foil of the same transitivity status. Verbs were selected if they were familiar and easily depicted. That is, the verb "jump" (an intransitive verb) was presented against "run," another intransitive verb. Agents remained consistent across the animations for the target verbs and the foils. Thus, children could not make selections based on a favorite character. The video clips were shown sequentially to prevent 2-year-olds from having to look at two videos containing dynamic events at once. Although prior research used dynamic, video events at test, they contained multiple, longer exposures to these events (Golinkoff et al., 2013; Maguire et al., 2006). The QUILS:TOD was designed to test many constructs quickly given 2-year-olds attention span. Once the children saw both animations, side by side still frames appeared on the screen and the narration instructed children to "Find the [target verb]." Children then touched the screen to select their answer.

Adjectives. Adjectives primarily modify nouns to provide additional details and distinguish between items in the same category. Evidence suggests that children begin to understand adjectives such as colors as early as 14 months (Booth & Waxman, 2009). However, adjectives tend to be acquired later than nouns and verbs, in part due to the conceptual complexity of this word class—children have to selectively attend to a feature of an object that is independent of the object itself (Kowalski & Zimiles, 2006). Children were presented with an illustration depicting three objects (see Figure 3). The narration asked children "Which one is [adjective]? Point to [adjective]," or "Who is [adjective]? Point to [adjective]." Two contrasting items were selected as foils for the target adjective. The foils selected aimed to represent clear distinctions from the target adjective.

Sample items chosen for Vocabulary Product are included in Table 3.

2.2.2 | Items to test for the products of syntax learning

Negation. Unlike affirmative statements, sentences involving negation are challenging for children under 2 years to understand, yet children show significant improvement during their third year in the ability to interpret such sentences (Grigoroglou et al., 2019; Nordmeyer & Frank, 2014). Negation items were developed using common nouns and verbs that could be depicted in a static image. Two static images were presented on the screen simultaneously, side by side. The child was asked to indicate which of the two people did not have a named object or was not doing an action. Similar to the presentation scheme of the verb items, the narration first instructed the children to assess the scene by

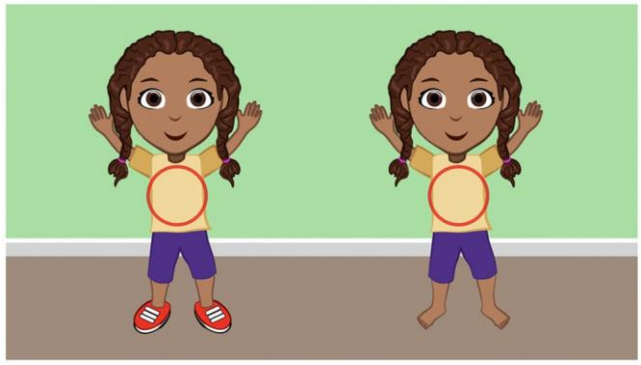

FIGURE 4 Negation Item 5. "Look at the girls. Which one has no shoes?".

saying, "Look at the [girls or boys]." The introductory phrase gave children the opportunity to examine both options before being asked to make a selection. Following this prompt, children were asked a question targeting negation such as, "Which one has no shoes?" (see Figure 4). Given that negation is difficult for children at this age, as well as the binary logic of negation, a third alternative foil was not used.

Wh-Questions. In a typical developmental sequence for *wh*-question acquisition, *what*, *where*, and *who* are generally learned prior to *why*, *when*, and *how* (Bloom et al., 1982; Rowland et al., 2003). The syntactic function of the earlier acquired wh-words (w*hat*, *where*, and *who)* may also be simpler because these questions ask for the key sentence components that they replace. As a result, these earlier wh-words serve relatively simple within-clause grammatical functions. The later acquired wh-words (*why*, *when*, and *how*) serve more complex functions in that they are asking for informa- tion concerning relations among an event's components (Bloom et al., 1982). As a result, only earlier acquired wh-words were included in the assessment.

"Where" questions are early acquired, but on a test, asking a question such as "Where is the hat?" becomes a test of vocabulary rather than a test of "where" questions. Other types of "where" questions target general knowledge and are more conceptually demanding, such as "Where does a bird live?". To avoid these complexities and focus on testing the syntax, we included only subject and object questions from simple sentences. Subject and object wh-questions emerge in children's comprehension (and sometimes in production) by the end of the second year of life (Goodwin et al., 2012; Seidl et al., 2003; Stromswold, 1995). Object questions (e.g., *What is the cat picking up?*) are more difficult than subject questions (e.g., *Who is hitting the drum?*) for children (de Villiers et al., 2008; Seidl et al., 2003).

Children were presented with an illustration depicting an event (e.g., a baby hitting a drum) with accompanying neutral narration which refers to different components of the scene (e.g., Look! A mommy, a baby, and a drum!). They were then asked a wh-question about a particular aspect of the event (e.g., "Who is hitting the drum?") and, in this case, shown the three items in the scene as foils. Wh-questions included in the final version of the QUILS:TOD include *who* and *what* questions.

Reversible Transitives. Reversible transitive sentences contain verbs that accept one or more objects and indicate the direction in which an activity is transferred from one actor (the agent) to another (the patient). Understanding these relationships requires an understanding of not only the label of the agent and the patient but also their syntactic relationship. This highlights the importance of word order information. For example, order information is necessary to distinguish *the boy chased the dog* from *the dog chased the boy*. This ability is present in some children even during the second

FIGURE 5 Reversible Transitive Item 38. "Make the car bump the truck!".

year of life before children can produce such sentences (de Villiers & de Villiers, 1972; Dittmar et al., 2011; Gertner et al., 2006; Hirsh-Pasek & Golinkoff, 1996).

Inclusion of the names of objects and actions that children likely understand by the age of two allows for this subtest to narrow in on the syntactic relationship rather than a test of vocabulary. An illustration of two objects was presented, and the children were prompted to "Make the [object] [*verb*] the [other object]." (see Figure 5) This subtest also explored the use of dragging as a response. Children had to identify which of the objects was the agent of the verb and drag it to the other object, which prompted an animation of the verb. The animation of the verb occurred regardless of whether children dragged the correct agent of the verb to the correct direct object. However, children were only awarded credit if they selected the correct agent and direct object. Children must make sense of the word order in the prompt to recreate the event.

Reversible Prepositions. Prepositions capture relations between other words in a clause, as in, "The book is *on* the cup." Prepositions were tested in two ways: in the vocabulary area and in the syntactic area. This is because prepositions can be understood as words containing a semantic meaning (as in "The tree is *behind* the man") as well as requiring that children understand how syntax signals a relationship between two entities. That is, children must not only understand the semantics of the prepositions (this is assessed in the vocabulary area) but must also recognize the importance of order information in prepositions that signal a relationship between two entities. For example, word order information is necessary to distinguish *the ball is behind the dog* from *the dog is behind the ball*.

Children's knowledge of prepositions lags somewhat behind their knowledge of verbs, but by age 3 typically developing children can both comprehend and produce many spatial prepositions (Bower et al., 2020; Schlosser et al., 2012). Some prepositions (e.g., *under*) are acquired earlier than others (e.g., *between*; Clark, 2009). The prepositions piloted included a combination of earlier acquired prepositions (*in*, *on*, *under*, and *in front of*) and later acquired prepositions (*above*, *below*, *between*, and *behind*) (Clark, 2009). We selected objects whose name children know by age three based on CDI norms (Dale & Fenson, 1996). Inclusion of drag and drop items within the preposition subtest introduces the flexibility to manipulate the images on the screen and allows for a more accurate representation of children's understanding of the concept. This feature does not restrict children to one of two predetermined choices, a target or a foil, as children can elect to drag the target object to any position on the screen. Therefore, dragging reduces chance responding, requiring children to understand the syntactical relation specified in the prompt. (e.g., "Put the dog *under* the chair"). Samples of the final items for testing Syntax Product are included in Table 4.

2.2.3 | Items to test for the process of vocabulary learning

Noun Learning. The processes involved in learning new nouns, particularly concrete nouns that label objects in situations of referential ambiguity, have been examined extensively (Golinkoff et al., 1994; Swingley, 2010). The initial steps are *fast mapping* (Carey & Bartlett, 1978) which involves inferring

TABLE 4 Syntax types and sample items overview.

Type	Item
Negation	Look at the girls. Which one has no shoes?
Wh-Questions	Who is hitting the drum?
Reversible transitives	Make the car bump the truck.
Reversible prepositions	Find: The book is on the cup.

a link between a novel sound unit (e.g., *cow*) and an object (e.g., a specific cow, present when the label is produced). Following this initial mapping, the next step is *extension* (Golinkoff et al., 1995; Markman & Hutchinson, 1984), or the process by which the label is applied to other members of the same basic category (e.g., other cows).

Fast mapping involves making a snap decision about word meaning based on whatever information is available at the time the novel noun is introduced; this includes using the principle of mutual exclusivity (Markman, 1989, 1992; Merriman et al., 1989) to eliminate alternative objects in the general vicinity that already have known labels. Extension requires children to use the *categorical scope* or the *taxonomic principle* (Golinkoff et al., 1994; Markman & Hutchinson, 1984) to generalize a label to other items in the same category. Of course, "words" formed from fast mapping and extension represent only partial knowledge. Their meanings will expand to include additional, more specific information acquired over time (Bion et al., 2013; Horst & Samuelson, 2008; Yurovsky et al., 2014). Yet the initial inferential processes of fast mapping and extension are critical components of lexical acquisition (Zosh et al., 2013). Children with language delays or disorders have difficulty fast mapping and require significantly more exposure to novel words than typically developing children (Gray, 2004; Rice et al., 1990, 1992).

Some rigorous fast mapping tasks use one novel object as the target and three familiar objects as foils (e.g., Golinkoff et al., 1992; Golinkoff et al., 2017). During early piloting, we presented all noun learning tasks with three possible objects, one novel object, and two foils. However, these tasks proved to be too difficult for some 2-year-olds and a subset of the items were reduced to a single novel item and one foil. In fact, even for older children tested with the QUILS, it proved difficult for them to remember items they did not implicitly choose (Aravind et al., 2018). The target object in the extension trial was an object that would receive the same name. However, it was designed to be different in color, pattern, or orientation from the target object in the fast-mapping trial. Items had two trials (see Figure 6). In the first trial, children were asked to find a novel object (e.g., "Can you find a *noof*?"). They viewed two objects on the screen: one that was a known object (e.g., a fork) for which children likely knew a name and the target, that is, one that was a novel object of a similar shape, and size. Immediately following this first trial, a second trial asked children to find another instance of the same item (e.g., "Can you show me another *noof*?"). They were presented with a new array of two objects: one novel object (i.e., a dissimilar novel object) and a novel exemplar of the object labeled in the first trial (i.e., a *noof* with different coloring or orientation).

Adjective Learning. Learning new adjectives requires children to recognize that a novel descriptor is being used to highlight a feature of one item among other, similar items (e.g., Waxman & Klibanoff, 2000). Children generally learn to link novel adjectives to object features only after having a name for the object itself (Gelman & Markman, 1985). Additionally, there are developmental gains across the preschool years in children's abilities to extend novel adjectives to features of objects from diverse basic-level categories, such as extending a novel label for a feature of a basket to the same feature of a spoon (Waxman & Klibanoff, 2000).

The adjective learning subtest was adapted from a task by Gelman and Markman (1985) using novel, salient visual patterns. These patterns were applied to familiar objects for which young children typically have labels by age 2 (Dale & Fenson, 1996), so that children would not simply map the

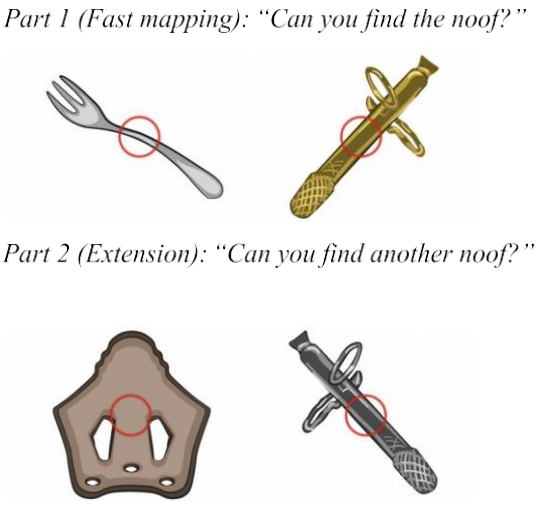

FIGURE 6 Noun Learning Item 18.

novel adjective onto a novel object by mistake. Adjective mapping trials examined one of two types of relations between the labeled item and the target item (i) within a basic-level category (e.g., a "tezzy" house and a different tezzy house): and (ii) across basic-level categories (e.g., "tezzy" house and a "tezzy" car). The foils were selected to test whether the children would map the new adjective to the novel property rather than to the object itself.

Items had two trials, one that required fast mapping and a second that required extension of the new adjective. In the first trial, children were shown a target property on a familiar object (e.g., polka dot pattern on a house) through the use of an ostensive label (e.g., "Wow! This house is very *tezzy*! Look! This house is very *tezzy*!"). The image moved up on the screen and became smaller as three additional images appeared below it: one object from the same basic-level category with a different property (e.g., a house with a solid color pattern), one object from the same basic-level category with the target property (e.g., polka dot pattern on a house) and one object from the same basic-level category with a different non-target property (e.g., a swirly pattern on a house). Children were asked to choose the object that shared the labeled target property (e.g., "What else is *tezzy*?"). A second trial asked children to find another instance of the property (e.g., "Show me what else is *tezzy!*") as they viewed a similar array of two new objects with the standard object remaining on the screen.

Verb Learning. To learn a new verb, children must not only map the novel verb onto its mean- ing but must also consider the argument slots (or nouns) around the verb like algebraic unknowns that can be filled in by many subjects and objects. Otherwise, their learning is narrow and formu- laic, as in "kissing-can–only-be-done-by–mommy." Early in development, children's understanding of verbs may be limited to particular agents and objects (Huttenlocher et al., 1983; Tomasello, 2000). The ability to extend verbs to new agents and objects is present as early as the third year (Forbes & Poulin-Dubois, 1997; Golinkoff, et al., 1996) and continues to develop through and beyond the preschool years (Forbes & Farrar, 1993; Kersten & Smith, 2002; Seston et al., 2009). Children with language difficulties, however, have poorer verb learning skills than their typically developing peers (Eyer et al., 2002; Johnson & de Villiers, 2009; Windfuhr et al., 2002).

Items had two trials, one that required fast mapping and a second that required extension of the new verb. In the first trial, children viewed a dynamic event with one or two characters engaged in a novel action (see Figure 7). The narration described the novel action using a novel verb, for example,

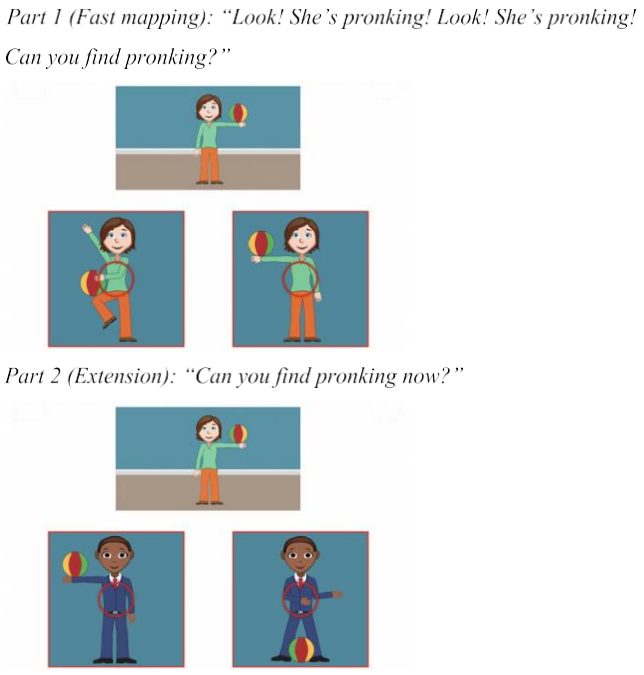

FIGURE 7 Verb Learning Item 33.

TABLE 5 Process types and sample items overview.

Type	Sample item
Adjective learning	What else is tezzy? Show me what else is tezzy?
Noun learning	Can you find the noof? Can you find another noof?
Verb learning	Can you find pronking? Can you find pronking now?

"Look! She's *pronking*! Look! She's *pronking*!" The dynamic event then froze and became smaller on the screen. Two options appeared: the target action being performed by the character from the dynamic scene and a novel action being performed by the same character. Children were asked to choose the labeled target action from these two options ("Can you find *pronking*?"). After this first trial, a second trial asked children to find another instance of the same verb (e.g., "Can you find *pronking* now?"). They were presented with two new options: the target action being performed by a novel character and a novel action being performed by the same novel character. To correctly answer this item, children must first fast map the novel verb (e.g., *pronking*) onto the action (e.g., rolling a ball from arm to arm over one's shoulders) and then extend this verb to another instance of the action with different actors.

The final items for testing Vocabulary Process are included in Table 5.

2.3 | Selecting the final items

2.3.1 | Testing phase I—Pilot testing

Following conventional evidence-based practice in psychometrics, the development team tried out more than three times the number of items to appear in the final screener (Schmeiser & Welch, 2006). Anticipating that the final screener would contain 40 items, Phase I, the pilot testing, included 139 items, with 38 items that tapped vocabulary, 35 items that tapped syntax, and 65 that tapped process. The participants for Initial Pilot Testing included 174 children (M_{age} = 30; 6, range = 22; 7–39; 1, 81 Male). Most of the children were not given all subtests, as the item sets were developed over time and tested in batches. Due to overall test length, children who did complete multiple subtests often did so across several sessions. Piloting was crucial for screener design. For example, given pilot children's performance compared to chance, it became clear that limiting foil choices (from 3 to 2 choices, or from 4 to 3) was essential for getting reliable performance.

To identify the best items from piloting for Phase II, Rasch analysis was used to analyze each subtest in Phase I (Piloting) and select a subset of items that ranged in difficulty across the two age bands. Based on these results, the initial 139 items were reduced to 73 items.

2.3.2 | Testing phase II—First item tryout testing

For the next round of testing the remaining 73 items were administered in 4 short blocks ranging from 5 to 10 min each. A total of 252 monolingual English-speaking preschoolers (mean age = 29.8 months, 138 male) from diverse socioeconomic backgrounds in Massachusetts, Delaware, and Pennsylvania participated in Phase II. The sample consisted of 133 children ages 2; 0–2; 6 and 119 children ages 2; 6-2; 11 (see Table 6).

Based on the data from Testing Phase II, two age bands of 2;0–2;6 and 2;6-3;0 were compared to determine differences between younger and older 2-year-olds. A Rasch model was used to calculate infit and outfit values. Rasch analysis assesses whether items are appropriate or redundant and has become a well-respected procedure for test development (Bond & Fox, 2013; Linacre, 2006). The

TABLE 6 Composition of First Item Tryout and Second Item Tryout sample populations.

	First item tryout	Second item tryout
Total *N*	252	448
Age		
2; 0–2; 6: *n* (%)	133 (52.78)	200 (44.64)
2; 6–2; 11: *n* (%)	119 (47.22)	248 (55.36)
Mean age (years): *M* (*SD*)	2; 6 (0; 3)	2; 6 (0; 3)
Gender		
Male: *n* (%)	138 (54.76)	232 (51.79)
Female: *n* (%)	114 (45.24)	216 (48.21)
SES		
Low: *n* (%)	51 (20.24)	169 (37.72)
Mid: *n* (%)	135 (53.57)	279 (62.28)
Not reported: *n* (%)	66 (26.19)	-

Rasch models test whether children pass items that reflect their ability, rather than pass a random selection of easy and hard items. Our procedures followed closely those conducted for the earlier instruments QUILS (Golinkoff et al., 2017) and QUILS:ES (Iglesias et al., 2021), as well as advice from classic and recent papers on Rasch applied to language measurement (Bond & Fox, 2013; Linacre, 2006).

Using these analyses, items that were redundant were removed. Wright maps plotting person ability against item ability identified items that were redundant with other items. In addition, Infit and Outfit values were used to delete items that were outside the model, as recommended by the recent meta-analysis of Rasch methodology applied in the area of language measurement (Aryadoust et al., 2021). The test was reduced from 73 to 51 items, with a small number of items simplified for further testing with a wider sample during Testing in Phase III.

2.3.3 | Testing phase III—Second item tryout testing

The reduced set of 51 items was separated into two blocks for Testing Phase III. Items were presented in two blocks due to the number of items to lessen the load on the children's attention and increase compliance. The order of block presentation was randomized across participants. The development team and collaborators administered the 51-item version of the screener to the final data collection sample of 448 children from childcare centers, preschools, Early Head Start programs, and laboratory settings in Massachusetts, Pennsylvania, Delaware, Oklahoma, California, New York, Texas, and Tennessee (see Table 6). We had a compliance rate of approximately 90%, indicating that the majority of children who began testing completed all of the items.

After completion of Testing Phase III, the development team removed problematic items following Rasch analyses as in Phase II as well as DIF analyses. DIF was conducted with respect to gender to ensure the items were not biased against girls or boys. The DIF analyses of Testing Phase III for the 51-item test revealed 2 items with significant DIF, one favoring boys and one favoring girls. The decision was made to keep both items since neither group is disadvantaged by the inclusion of both of these items, as per usual practice (Nandakumar, 1993).

Children's performance on the final items can be found in Table 7 which includes the descriptive statistics for each subtest and the screener as a whole.

2.3.4 | Testing phase IV - Final test creation

The final QUILS:TOD consists of the best 40 items culled from two rounds of testing. Separate Rasch analyses were conducted on each area of the final QUILS:TOD as well as on the overall screener. Fit

TABLE 7 Descriptive statistics for QUILS: TOD.

	Mean	*SD*	*Range*
QUILS:TOD vocabulary	6.73	2.39	0–11
QUILS:TOD syntax	7.92	2.58	1–13
QUILS:TOD process	6.59	3.34	0–16
QUILS:TOD overall	21.23	6.74	4–39

Note: The maximum possible score for each area of the QUILS:TOD was 11 for vocabulary, 13 for syntax, 16 for process, and 40 for the overall score. The sample for whom descriptives are shown is the subset of children from Phase III who completed both blocks of the screener, $N = 403$.

TABLE 8 Distribution of the areas, types, and number of items per subtest on the QUILS:TOD.

Subtest	Area	Response type	Event type	Items per subtest	Number of options (target and foils)
Nouns	Vocab	Touch	Static	4	3
Negation	Syntax	Touch	Static	4	2
Fast mapping adjectives	Process	Touch	Static	4 (2 and 2 extensions)	2 items with 2 2 items with 3
Verbs	Vocab	Touch	Dynamic	3	2
Fast mapping nouns	Process	Touch	Static	8 (4 and 4 extensions)	4 items with 2 4 items with 3
Adjectives	Vocab	Touch	Static	4	3
Wh- Questions	Syntax	Touch	Dynamic	3	2 items with 3 1 items with 2
Novel verbs	Process	Touch	Dynamic	4	2
Reversible prepositions	Syntax	Touch Drag	Static Dynamic	2 1	2
Reversible transitives	Syntax	Drag	Dynamic	3	2
Total				40	

statistics for each of the areas and overall were close to the expected value of 1. Fit statistics were also investigated at the item level for all areas and overall. More emphasis was placed on the Infit Mean-Square (MNSQ) because it is a weighted measure and is sensitive to the study subjects near the item level on the underlying ability continuum. (A mean-square value of 0.5–1.5 is regarded as productive for measurement. See https://www.rasch.org/rmt/rmt162f.htm.) Infit MNSQ values for all items in each of the three areas and overall were within the expected range of 0.8–1.2.

Further evidence of reliability was established by the person and item separation values (1.1–7.3), suggesting that each of the areas and the screener overall can successfully differentiate the different proficiencies of children and that items are well spread along the measures of difficulty.

The number of items per subtest, type of response required, and number of options are detailed in Table 8. These items have appropriate discriminative power across the range of abilities in the sample, which is especially important in the lowest performing group of children. The item set is also unbiased with respect to gender. For no subtest is the overall performance near ceiling even for the older age group.

3 | RESULTS

3.1 | Development of norms

The QUILS:TOD standard scores are generated based on age norms and the QUILS:TOD raw scores. The standard score reflects each child's performance as compared to the norms generated from the final norming sample of children, separately for each age group (2; 0–2; 6 and 2; 6–2; 11). The norming sample was a subsample of 270 children (135 female, 135 male) drawn from the Phase III sample (N = 448), stratified by SES status and gender to match the U.S. census, as shown in Table 9. This included 104 children ages 2; 0–2; 6 (M = 2; 3; SD = 0; 2) and 166 children ages 2; 6–2; 11 (M = 2; 9; SD = 0; 2). The norms will be available in the test manual once it is published.

TABLE 9 Composition of the norming sample for the QUILS: TOD.

	Final norming sample
Total *N*	270
Age	
2; 0–2; 6: *n* (%)	104 (38.52)
2; 6–2; 11: *n* (%)	166 (61.48)
Mean age (years): *M* (*SD*)	2; 7 (0; 3)
Gender	
Male: *n* (%)	135 (50.00)
Female: *n* (%)	135 (50.00)
SES	
Low: *n* (%)	149 (55.19)
Mid: *n* (%)	121 (44.81)

Information on socioeconomic status (SES) was provided either in the form of mother's self-reported educational attainment or by enrollment in a low-income childcare center. The majority of the children tested were from low SES families (55.2%), and 44.8% were from mid-SES families. The percentage of mid-SES families is close to the percentage reported in the 2019 U.S. census data for females age 18–39 years having an education level of an associate's degree and above (45.3%). Parents reported home language at time of screening, and we selected only monolingual English speakers for the norming sample.

Demographic data for race were available for 95.9% of the norming sample. Of those who reported this information, 61.4% were White, 29.3% were Black/African American, 5.8% were multiracial, 2.3% were Asian, and 1.5% were other races. Additionally, 94.8% of parents reported whether their child was of Hispanic origin; of those who reported on it, 12.1% of children were of Hispanic origin.

The standard scores for the baby QUILS were normalized to the bell-shaped distribution. In the area (Vocabulary, Syntax, Process) scores, and these area scores were produced for each of the two age groups in the standardization sample. Next, each area score variable was transformed so that its shape matched the bell-shaped curve with a mean of 100 and a standard deviation of 15. A scaled score was then created by summing the three standardized area scores, and the norming process was repeated on this scaled score to derive a standardized overall score. As with the area scores, the scaled score was transformed so that its shape matched the bell-shaped curve with a mean of 100 and a standard deviation of 15. Finally, norms tables were developed by comparing each standard score to its corresponding raw score. This process of transforming raw scores to normalized standard scores represents the most common application of a "nonlinear area conversion" (Thorndike, 1982, p. 115).

Cut scores were determined by considering the role of the QUILS:TOD in screening children at risk for language impairment. A language screener should try to identify all children who might be at risk, as the cost associated with missing vulnerable children is greater than the cost of unnecessarily screening children who will pass a more comprehensive test. For that reason, the development team judged scoring below the 25th percentile to be a conservative estimate of risk, given that the population of children with language impairments is estimated to lie between 7% and 12% (Leonard, 2014; Tomblin et al., 1997). The cut scores are based on this 25th percentile. The norms will be included in the manual when the screener is published.

A traditional norming approach was problematic because of the jumps between norms across age intervals. That is, the very same raw test score would be interpreted differently for two individuals

who are in two successive age bands, yet only a few days apart in age. New continuous norming approaches (e.g., GAMLSS) are recently becoming popular (Voncken, et al., 2021), but they may need larger samples for accuracy. We chose to use the same norming procedures here as for the QUILS and QUILS:ES screeners, but used the two age bands rather than year-long bands given the rapid rate of language acquisition during the third year of life. In future as larger samples accrue, it may be possible to use GAMLSS to refine the norms on QUILS:TOD and avoid the discontinuities of traditional methods. Further, beginning with the groundbreaking work of Roger Brown (1973), it has been shown that at this early point in language development, age is a poor predictor of where children stand relative to their peers. Hence, we adopted two large age bins—the first and second halves of the third year.

3.2 | Validity

The development team examined construct validity and convergent validity. Construct validity aimed to determine whether the screener measured significant and meaningful aspects of language development for this age range. The design section above establishes the basis for construct validity in reviewing past research on these types of linguistic forms in the developmental literature, though usually on small samples. That is, research on transcripts and in controlled experiments with typically developing children suggests that their receptive language incorporates these linguistic constructs at 2–3 years of age. It is yet another matter to demonstrate that screeners of large numbers of children can show predictable and meaningful variability across this short age range. Figure 8 shows that there is growth over the year between 24 and 36 months in the areas considered. Children who are 24–29 months do markedly less well than children aged 30–36 months in each of the area scores. Chance values vary across vocabulary (37.6%), syntax (47.4%), and process (31.7%) calculated separately by area as per number of items and foils reported in Table 8. We compared each age group's performance to both chance and ceiling values. For both age groups, the level of performance is significantly higher than chance (all $p < 0.01$) and below ceiling (all $p < 0.001$). Each item was chosen because it discriminated successfully in the Rasch model between low and high scoring children.

To assess convergent validity, parents and caregivers of 205 children filled out the MacArthur-Bates Communicative Development Inventory Short Form (MB-CDI; Fenson, 2007), and 196 children were tested on Form A of the Peabody Picture Vocabulary Test–Fourth Edition (PPVT-4; Dunn & Dunn, 2007). Across these samples, 129 children were tested on both the MB-CDI and PPVT. The PPVT was administered by the same examiners who tested the children on the QUILS:TOD, though at the point of testing they were blind to their resulting scores. Both the MB-CDI and PPVT-4 assess aspects of language development and have demonstrated validity and reliability. Tests measure and emphasize different aspects of language; nonetheless, we would expect reasonably high correlations among the different tests. Since there are no existing measures to test Process at age 2, we expect tests of vocabulary to correlate with this area as well as it is an important component for learning vocabu- lary (see Table 10).

The correlations indicate that the overall QUILS:TOD score correlates significantly with MB-CDI and PPVT-4 raw, total scores. Although the MB-CDI and PPVT-4 are both normed for children aged 2, the MB-CDI is only normed through 30 months, and the PPVT is only normed from 30 months. As a result, for the purpose of this validity study, both measures were used outside of standardization, and we used raw scores for our calculations. The correlations with the MB-CDI are shown in Figure 9 and correlations with the PPVT are shown in Figure 10.

The area scores (i.e., Vocabulary, Syntax, and Process) also correlate significantly with these assessments. Given that the PPVT-4 is a behavioral measure of receptive vocabulary and the

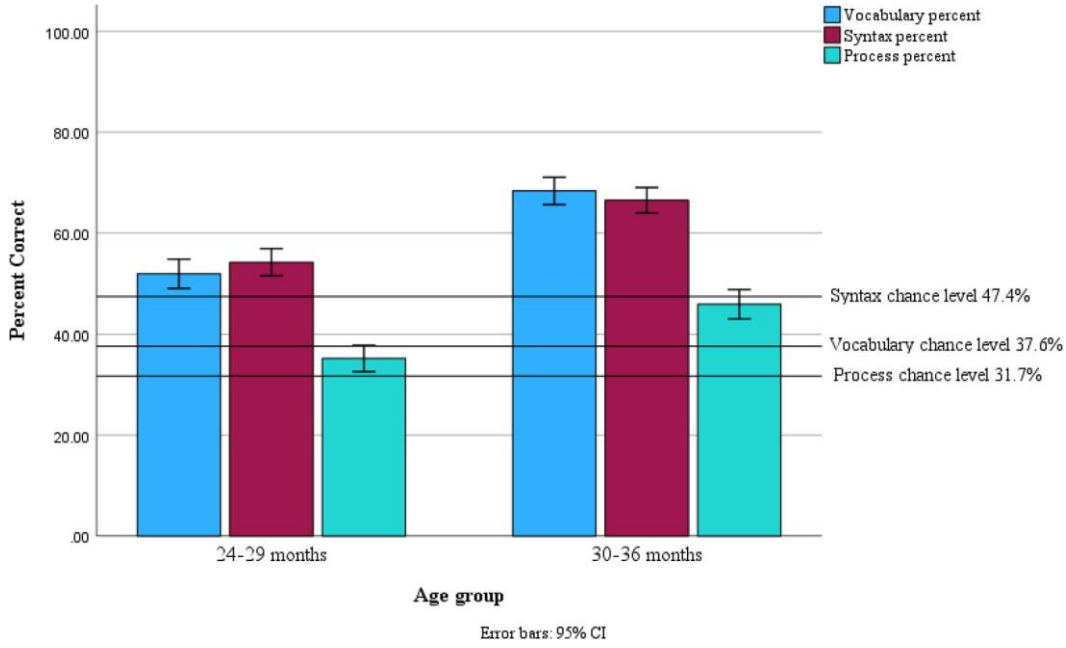

FIGURE 8 Distribution of average QUILS: TOD percent correct scores by age group.

TABLE 10 Convergent validity coefficients by age band.

Age band 24–29 months	**QUILS:TOD Subtests**				
Age partialed correlations		**Vocabulary**	**Syntax**	**Process**	**Overall**
MB-CDI	Pearson correlation	0.23[a]	0.29[b]	0.22[a]	0.35[b]
	n	88	88	88	88
PPVT	Pearson correlation	0.39[b]	0.37[b]	0.44[b]	0.51[b]
	n	78	78	78	78
Age band 30–36 months					
MB-CDI	Pearson correlation	0.28[b]	0.39[b]	0.29[b]	0.39[b]
	n	104	104	104	104
PPVT	Pearson correlation	0.61[b]	0.59[b]	0.47[b]	0.67[b]
	n	112	112	112	112

Note: MB-CDI, MacArthur-Bates Communicative Development Inventory Short Form (MB-CDI; Fenson, 2007); PPVT-4, Peabody Picture Vocabulary Test-Fourth Edition (PPVT-4; Dunn & Dunn, 2007).

[a]Fisher z-transformed correlation is significant at the 0.05 level (2-tailed).

[b]Fisher z-transformed correlation is significant at the 0.01 level (2-tailed).

MB-CDI is a parent-report measure of productive vocabulary, the development team predicted that the QUILS:TOD would correlate more strongly with the PPVT-4 than the MB-CDI, and analyses confirmed this prediction. This does not appear to be an artifact of a ceiling effect for the MB-CDI at 36 months. The scatterplots do reveal a ceiling effect for the MB-CDI, but for children in both age groups. The correlations for the final 40 item QUILS:TOD and MB-CDI is 0.319 ($p < 0.01$) for children 24.0–29.9 months and is 0.385 ($p < 0.01$) for children 30.0–35.9 months. This demonstrates that the correlations are in fact weaker with MB-CDI for the younger than the older group. It appears

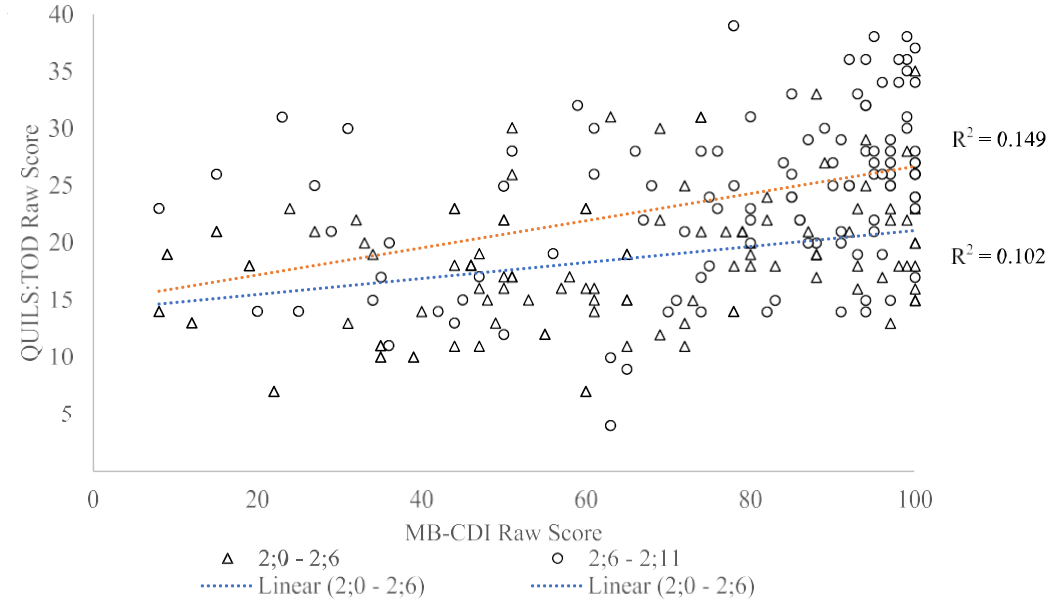

FIGURE 9 Correlation of QUILS: TOD raw scores with MB-CDI raw scores.

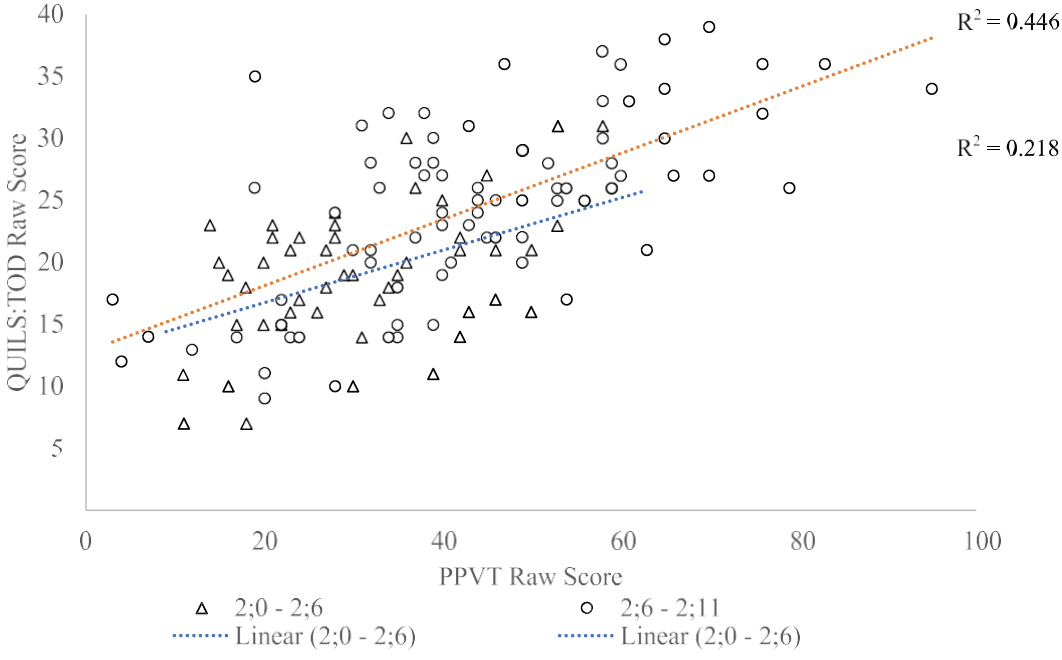

FIGURE 10 Correlation of QUILS: TOD raw scores with PPVT raw scores. Each point represents an individual participant. Triangles indicate the lower age-band from 2; 0–2; 6. Circles indicate the higher age-band from 2; 6-2; 11.

that the QUILS:TOD detects a wider range of ability than the MB-CDI. Thus, these results, together with the results of the construct validity tests, provide confirmation that the QUILS:TOD is measuring important aspects of language development for young children from the ages of 2; 0 through 2; 11.

3.2.1 | Test-retest reliability

The reliability of a test asks whether the scores are stable for one individual at different times. Eighty-six of the children participating in the Second Item Tryout were randomly assigned to take the QUILS:TOD a second time. The time interval between the first and second testing ranged from 3 to 5 weeks. The test-retest coefficient for the whole test was 0.75, indicating that scores from the QUILS:TOD possess reasonable stability across short time periods.

3.2.2 | Internal consistency reliability

A statistical test of internal consistency, Cronbach's coefficient alpha, was calculated to determine whether the items formed a coherent set reflecting a single construct, in this case, language ability, even though the items may vary in difficulty. Coefficient alpha provides a lower bound value of test reliability and is considered to be a conservative estimate of a test's reliability (Carmines & Zeller, 1979; Reynolds et al., 2009). The coefficient alpha is 0.68 for the Vocabulary area, 0.67 for the Syntax area, 0.76 for the Process area, and 0.86 for the overall QUILS:TOD. These good to high coefficient values demonstrate that items are coherent in measuring the unidimensional construct underlying each of the areas of the screener, as well as the overall QUILS:TOD as a language comprehension screener for young children. Age-partialled correlations among the areas are provided in Table 11 and all are statistically robust.

3.2.3 | Predictive validity

An important value of a screener is not just that it can provide an accurate estimate of current functioning, but also whether it can predict future language skills. A subset of the children who had completed the QUILS:TOD was selected to complete the QUILS as a measure of predictive validity. It is important to note that there are no items on the QUILS that also appear on the QUILS:TOD. Furthermore, there are some subtests that only appear on one of the screeners. For example, QUILS:TOD contains subtests like negation with no counterpart on QUILS, and QUILS contains subtests of conjunctions and embedded sentences that are not on QUILS:TOD. The tests of constructs that do share a common theme have methodological differences in the two age ranges. For the validation study, participants were tested when they were within the age range for QUILS. The number of participants was balanced across test sites. A total of 31 children had complete data for the QUILS:TOD, MB-CDI, and QUILS. Participant characteristics can be found in Table 12.

Predictive validity was calculated by comparing children's performance at both time points. The QUILS:TOD and MB-CDI scores were collected concurrently and the QUILS score was collected after the child turned 3. The correlations between measures were controlled for the child's age at

TABLE 11 Age-Partialed correlations for each QUILS: TOD Area score.

	QUILS:TOD vocabulary	QUILS:TOD syntax
QUILS:TOD vocabulary		
QUILS:TOD syntax	0.529***	
QUILS:TOD process	0.413***	0.387***

Note: *** = $p < 0.001$.

TABLE 12 Composition of the predictive sample for the QUILS: TOD with the QUILS.

Predictive validity sample	24.0–29.9 months	30.0–35.9 months
N	14	17
Mean age (months): *M* (*SD*)	26.6 (1.6)	32.6 (1.6)
Gender		
Male: *n*	6	12
Female: *n*	8	5

TABLE 13 Predictive validity correlations between measures controlling for age.

	MB-CDI	QUILS:TOD
QUILS	0.28	0.67***

Note: These correlations were conducted using raw scores, controlling for age at QUILS: TOD administration and months elapsed between QUILS:TOD and QUILS administration. Children in this sample were first tested on the QUILS: TOD between 24.0 and 35.9 months and were tested on the QUILS after their third birthday.

***, $p < 0.001$.

the time of QUILS:TOD administration and the time elapsed between the two assessments. These correlations can be found in Table 13. The correlation between the QUILS:TOD and the QUILS was statistically significant (r (31) = 0.67, $p < 0.001$). Interestingly, the earlier MB-CDI did not predict significantly to later QUILS.

4 | DISCUSSION

The purpose of this paper was to establish whether we could develop a behavioral measure of children's language capabilities at age two that is grounded in theory and research and that fulfills practical requirements. We culled the existing literature for potentially meaningful markers of language disorder risk that could be reliably assessed in 2-year-olds and were not going to be culturally and dialectically biased. Focusing on *product* and *process* allowed us to go beyond what the child had learned (product) to assess the child's ability to learn new language items (process). Thus, it is a more balanced test that should not compromise a child who is from an under-resourced environment. Focusing on receptive language allowed us to tap into knowledge not readily evident in the speech of young children and reduced the possibility of misidentifying children who were either late-talkers or reticent to speak. We also were guided by a desire to make this tool engaging for toddlers and easily accessible to a broad range of potential users (e.g., childcare providers or pediatricians) with little or no training in administering and scoring psychological assessments.

Our results answered affirmatively that a screener that meets these criteria can be developed. The QUILS:TOD, a self-contained touchscreen tool that assesses vocabulary product, syntax prod- uct, and language learning process, is valid and reliable based on data from a large, diverse group of 2-year-olds. The QUILS:TOD will need further scrutiny to ensure that the tool has adequate sensitivity and specificity, taking into consideration that the harm of false positives (misidentification as having a language problem) is relatively minor, compared to the false negative, namely mistaking that no problem exists (Glascoe, 2001). Ideally, these screeners should meet the same standards of sensitivity and specificity as standardized diagnostic tests, though usually specificity is sacrificed in favor of sensitivity so that the screener "catches" more children potentially at risk. Pace et al. (2021)

have also demonstrated that the QUILS has good sensitivity and specificity despite its brevity. Further work on sensitivity and specificity on the two-year-old test will add considerably to the value of this assessment. However, it should be noted that gold standard indices of language disorder against which to judge the QUILS:TOD are not simple at age two (Nelson et al., 2006).

Studies of predictive validity to later indices of language and general development are also important. We have only completed one small study so far, predicting only as far as QUILS in the fourth year of life, because the instrument is new and was completed at the beginning of the COVID pandemic. As with Friend et al. (2018), we expect the screener to outperform parental reports and potentially predict to school readiness later. Based on our preliminary data, the QUILS:TOD correlates highly with children's later performance on the QUILS. The MB-CDI collected at the same time point at the QUILS:TOD did not demonstrate correlations with performance at age three on the QUILS. Hence there is continuity between the QUILS:TOD and the QUILS, and there may be value in measuring receptive language over this preschool timespan. With more research, it may be possible to link each area (vocabulary, syntax, and process) to a different dimension of linguistic success, but that has not yet been shown.

The remaining question is, have we improved the testing of 2-year-olds' language? The advantages of a touch screen delivery mean reduced demands for the tester, making screening more feasible in at lower cost (e.g., without a speech language pathologist), and with automatic scoring and reports (Golinkoff et al., 2017). But the other design choices—the use of dynamic stimuli, the possibility of drag responses, intermittent cartoons for encouragement– were designed to make testing more feasible at the two-year-old level and to reduce the interfering effects of limited attention span and unfamiliarity with the static, cartoon, representation of events of paper and pencil tests. It remains for more research to determine whether these makes the QUILS:TOD a *better* estimate of young children's language ability than the current alternatives.

ACKNOWLEDGMENTS

The research was supported by the Institute of Education Sciences, grants R305A110284 and R324A160241. The authors declare no conflicts of interest with regard to the funding source for this study.

ORCID

Emily Jackson https://orcid.org/0000-0003-1096-0331

REFERENCES

Akhtar, N., & Tomasello, M. (1997). Young children's productivity with word order and verb morphology. *Developmental Psychology*, *33*(6), 952–965. https://doi.org/10.1037/0012-1649.33.6.952

Aravind, A., de Villiers, J., Pace, A., Valentine, H., Golinkoff, R., Hirsh-Pasek, K., Iglesias, A., & Wilson, M. S. (2018). Fast mapping word meanings across trials: Young children forget all but their first guess. *Cognition*, *177*, 177–188. https://doi.org/10.1016/j.cognition.2018.04.008

Aryadoust, V., Ng, L. Y., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing*, *38*(1), 6–40. https://doi.org/10.1177/0265532220927487

Astington, J. W., & Jenkins, J. M. (1999). Relation between language and theory-of-mind development. *Developmental Psychology*, *35*(5), 1311–1320. https://doi.org/10.1037/0012-1649.35.5.1311

Bion, R. A., Borovsky, A., & Fernald, A. (2013). Fast mapping, slow learning: Disambiguation of novel word–object mappings in relation to vocabulary learning at 18, 24, and 30 months. *Cognition*, *126*(1), 39–53. https://doi.org/10.1016/j.cognition.2012.08.008

Bloom, L., Merkin, S., & Wootten, J. (1982). *Wh"-Questions: Linguistic factors that contribute to the sequence of acquisition* (pp. 1084–1092). Child Development.

Bond, T. G., & Fox, C. M. (2013). *Applying the Rasch model: Fundamental measurement in the human sciences*. Psychology Press.

Booth, A. E., & Waxman, S. R. (2009). A horse of a different color: Specifying with precision infants' mappings of novel nouns and adjectives. *Child Development*, *80*(1), 15–22. https://doi.org/10.1111/j.1467-8624.2008.01242.x

Bornstein, M. H., Hahn, C. S., & Haynes, O. M. (2004). Specific and general language performance across early childhood: Stability and gender considerations. *First Language*, *24*(3), 267–304. https://doi.org/10.1177/0142723704045681

Bower, C. A., Foster, L., Zimmermann, L., Verdine, B. N., Marzouk, M., Islam, S., Golinkoff, R., & Hirsh-Pasek, K. (2020). Three-year-olds' spatial language comprehension and links with mathematics and spatial performance. *Developmental Psychology*, *56*(10), 1894–1905. https://doi.org/10.1037/dev0001098

Bricker, D., Squires, J., Mounts, L., Potter, L., Nickel, R., Twombly, E., & Farrell, J. (1999). *Ages and stages questionnaire*. Paul H. Brookes.

Brown, R. (1973). *A first language: The early stages*. Harvard University Press.

Carey, S. (2010). Beyond fast mapping. *Language Learning and Development*, *6*(3), 184–205. https://doi.org/10.1080/15475441.2010.484379

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *Papers and Reports on Child Language Development*, *15*, 17–29.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Sage Publications.

Chapman, R. S. (1978). Comprehension strategies in children. In J. F. Kavanaugh & W. Strange (Eds.), *Speech and language in the laboratory, school, and clinic* (pp. 308–327). MIT Press.

Clark, E. V. (2009). *First language acquisition*. Cambridge University Press.

Cocking, R. R., & McHale, S. (1981). A comparative study of the use of pictures and objects in assessing children's receptive and productive language. *Journal of Child Language*, *8*, 1–13. https://doi.org/10.1017/s030500090000297x

Conti-Ramsden, G., & Durkin, K. (2012). Language development and assessment in the preschool period. *Neuropsychology Review*, *22*(4), 384–401. https://doi.org/10.1007/s11065-012-9208-z

Crowe, K., & McLeod, S. (2020). Children's English consonant acquisition in the United States: A review. *American Journal of Speech-Language Pathology*, *29*(4), 2155–2169. https://doi.org/10.1044/2020_ajslp-19-00168

Dale, P. S., & Fenson, L. (1996). Lexical development norms for young children. *Behavior Research Methods, Instruments, & Computers*, *28*(1), 125–127. https://doi.org/10.3758/bf03203646

de Villiers, J., Roeper, T., Bland-Stewart, L., & Pearson, B. (2008). Answering hard questions: Wh-movement across dialects and disorder. *Applied PsychoLinguistics*, *29*(1), 67–103. https://doi.org/10.1017/s0142716408080041

de Villiers, J. G., & de Villiers, P. A. (1973). Development of the use of word order in comprehension. *Journal of Psycholinguistic Research*, *2*(4), 331–341. https://doi.org/10.1007/bf01067055

de Villiers, J. G., & Johnson, V. (2007). In R. Wagner, A. Muse, & K. Tannebaum (Eds.), *Implications of new vocabulary assessments for minority children, Vocabulary acquisition: Implications for reading comprehension* (pp. 157–181). Guilford press.

de Villiers, P. A., & de Villiers, J. G. (1972). Early judgments of semantic and syntactic acceptability by children. *Journal of Psycholinguistic Research*, *1*(4), 299–310. https://doi.org/10.1007/bf01067785

Dittmar, M., Abbot-Smith, K., Lieven, E., & Tomasello, M. (2011). Children aged 2; 1 use transitive syntax to make a semantic-role interpretation in a pointing task. *Journal of Child Language*, *38*(5), 1109–1123. https://doi.org/10.1017/s0305000910000747

Dollaghan, C. (2013). Late talker as a clinical category: A critical evaluation. In L. A. Rescorla & P. S. Dale (Eds.), *Late talkers: Language development, interventions, and outcomes*. Brookes.

Dunn, L. M., & Dunn, D. M. (2007). *PPVT-4: Peabody picture vocabulary test*. Pearson Assessments.

Eyer, J. A., Leonard, L. B., Bedore, L. M., McGregor, K. K., Anderson, B., & Viescas, R. (2002). Fast mapping of verbs by children with specific language impairment. *Clinical Linguistics and Phonetics*, *16*(1), 59–77. https://doi.org/10.1080/02699200110102269

Fenson, L. (2007). *MacArthur-Bates Communicative Development Inventories (MB-CDIs): User's guide and technical manual*. Paul H. Brookes.

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

Fernald, A., & Marchman, V. A. (2011). *Causes and consequences of variability in early language learning. Experience, variation, and generalization: Learning a first language (trends in language acquisition research)* (pp. 181–202). John Benjamins.

Fisher, C. (2002). Structural limits on verb mapping: The role of abstract structure in 2.5-year-olds' interpretations of novel verbs. *Developmental Science*, *5*(1), 55–64. https://doi.org/10.1111/1467-7687.00209

Forbes, J. N., & Farrar, M. J. (1993). Children's initial assumptions about the meaning of novel motion verbs: Biased and conservative? *Cognitive Development*, *8*(3), 273–290. https://doi.org/10.1016/s0885-2014(93)80002-b

Forbes, J. N., & Poulin-Dubois, D. (1997). Representational change in young children's understanding of familiar verb meaning. *Journal of Child Language*, *24*(2), 389–406. https://doi.org/10.1017/s0305000997003127

Friedman, S. L., & Stevenson, M. B. (1975). *Developmental changes in the understanding of implied motion in two-dimensional pictures* (pp. 773–778). Child Development.

Friend, M., & Keplinger, M. (2003). An infant-based assessment of early lexicon acquisition. *Behavior Research Methods, Instruments, & Computers*, *35*(2), 302–309. https://doi.org/10.3758/bf03202556

Friend, M., Smolak, E., Liu, Y., Poulin-Dubois, D., & Zesiger, P. (2018). A cross-language study of decontextualized vocabulary comprehension in toddlerhood and kindergarten readiness. *Developmental Psychology*, *54*(7), 1317–1333. https://doi.org/10.1037/dev0000514

Gelman, S. A., & Markman, E. M. (1985). Implicit contrast in adjectives vs. nouns: Implications for word-learning in preschoolers. *Journal of Child Language*, *12*(1), 125–143. https://doi.org/10.1017/s0305000900006279

Gentner, D., Boroditsky, L., Bowerman, M., & Levinson, S. (2001). Individuation, relativity, and early word. *Language, Culture and Cognition*, *3*, 215–256.

Gertner, Y., Fisher, C., & Eisengart, J. (2006). Learning words and rules: Abstract knowledge of word order in early sentence comprehension. *Psychological Science*, *17*(8), 684–691. https://doi.org/10.1111/j.1467-9280.2006.01767.x

Glascoe, F. P. (2001). Are overreferrals on developmental screening tests really a problem? *Archives of Pediatrics and Adolescent Medicine*, *155*(1), 54–59. https://doi.org/10.1001/archpedi.155.1.54

Glascoe, F. P. (1997). *Parents' evaluations of developmental status*. Ellsworth and Vandermeer Press.

Gleitman, L. R., Cassidy, K., Nappa, R., Papafragou, A., & Trueswell, J. C. (2005). Hard words. *Language Learning and Development*, *1*, 23–64. https://doi.org/10.1207/s15473341lld0101_4

Golinkoff, R. M., de Villiers, J. G., Hirsh-Pasek, K., Iglesias, A., Wilson, M. S., Morini, G., & Brezack, N. (2017). *User's manual for the quick interactive Language Screener (QUILS): A measure of vocabulary, syntax, and language acquisition skills in young children*. Paul H. Brookes Publishing Company.

Golinkoff, R. M., & Hirsh-Pasek, K. (2006). Introduction: Progress on the verb learning front. In *Action meets word: How children learn verbs* (pp. 3–28).

Golinkoff, R. M., Hirsh-Pasek, K., Bailey, L. M., & Wenger, N. R. (1992). *Young children and adults use lexical principles to learn new nouns* (Vol. 28). Developmental Psychology.

Golinkoff, R. M., Jacquet, R. C., Hirsh-Pasek, K., & Nandakumar, R. (1996). Lexical principles may underlie the learning of verbs. *Child Development*, *67*(6), 3101–3119. https://doi.org/10.1111/j.1467-8624.1996.tb01905.x

Golinkoff, R. M., Ma, W., Song, L., & Hirsh-Pasek, K. (2013). Twenty-five years using the intermodal preferential looking paradigm to study language acquisition: What have we learned? *Perspectives on Psychological Science*, *8*(3), 316–339. https://doi.org/10.1177/1745691613484936

Golinkoff, R. M., Mervis, C. B., & Hirsh-Pasek, K. (1994). Early object labels: The case for a developmental lexical principles framework. *Journal of Child Language*, *21*(1), 125–155. https://doi.org/10.1017/s0305000900008692

Golinkoff, R. M., Shuff-Bailey, M., Olguin, R., & Ruan, W. (1995). Young children extend novel words at the basic level: Evidence for the principle of categorical scope. *Developmental Psychology*, *31*(3), 494–507. https://doi.org/10.1037/0012-1649.31.3.494

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

Gray, S. (2004). Word learning by preschoolers with specific language impairment. *Journal of Speech, Language, and Hearing Research*, *47*(5), 1117–1132. https://doi.org/10.1044/1092-4388(2004/083)

Grigoroglou, M., Chan, S., & Ganea, P. A. (2019). Toddlers' understanding and use of verbal negation in inferential reasoning search tasks. *Journal of Experimental Child Psychology*, *183*, 222–241. https://doi.org/10.1016/j.jecp.2019.02.004

Halpin, K. S., Smith, K. Y., Widen, J. E., & Chertoff, M. E. (2010). Effects of universal newborn hearing screening on an early intervention program for children with hearing loss, birth to 3 yr of age. *Journal of the American Academy of Audiology*, *21*(03), 169–175. https://doi.org/10.3766/jaaa.21.3.5

Hirsh-Pasek, K., & Golinkoff, R. M. (1996). The intermodal preferential looking paradigm: A window onto emerging language comprehension. In D. McDaniel, C. McKee, & H. S. Cairns (Eds.), *Methods for assessing children's syntax* (pp. 105–124). The MIT Press.

Hirsh-Pasek, K., & Golinkoff, R. M. (1999). *The origins of grammar: Evidence from early language comprehension.* MIT press.

Horst, J. S., & Samuelson, L. K. (2008). Fast mapping but poor retention by 24-month-old infants. *Infancy*, *13*(2), 128–157. https://doi.org/10.1080/15250000701795598

Huttenlocher, J., Smiley, P., & Charney, R. (1983). Emergence of action categories in the child: Evidence from verb meanings. *Psychological Review*, *90*(1), 72–93. https://doi.org/10.1037/0033-295x.90.1.72

Huttenlocher, J., Waterfall, H., Vasilyeva, M., Vevea, J., & Hedges, L. V. (2010). Sources of variability in children's language growth. *Cognitive Psychology*, *61*(4), 343–365. https://doi.org/10.1016/j.cogpsych.2010.08.002

Iglesias, A., de Villiers, J., Golinkoff, R. M., Hirsh-Pasek, K., & Wilson, M. S. (2021). *User's manual for the quick interactive Language Screener - ES ™ (QUILS-ES™): A measure of vocabulary, syntax, and language acquisition skills in young bilingual children*. Brookes Publishing Co.

Ireton, H., & Ireton, H. (1992). Child development inventory.

Johnson, V. E., & de Villiers, J. G. (2009). Syntactic frames in fast mapping verbs: Effect of age, dialect, and clinical status. *Journal of Speech, Language, and Hearing Research*, *52*(3), 610–622. https://doi.org/10.1044/1092-4388(2008/07-0135)

Kaiser, A. P., Chow, J. C., & Cunningham, J. E. (2022). A case for early language and behavior screening: Implications for policy and child development. *Policy Insights from the Behavioral and Brain Sciences*, *9*(1), 120–128. https://doi.org/10.1177/23727322211068886

Kersten, A. W., & Smith, L. B. (2002). Attention to novel objects during verb learning. *Child Development*, *73*(1), 93–109. https://doi.org/10.1111/1467-8624.00394

Kochanoff, A., Hirsh-Pasek, K., Newcombe, N., & Weinraub, M. (2003). Using science to inform preschool assessment. *Center for Improving Resources in Children's Lives*.

Kowalski, K., & Zimiles, H. (2006). The relation between children's conceptual functioning with color and color term acquisition. *Journal of Experimental Child Psychology*, *94*(4), 301–321. https://doi.org/10.1016/j.jecp.2005.12.001

Larson, A. L. (2016). Language screening for infants and toddlers: A literature review of four commercially available tools. *Communication Disorders Quarterly*, *38*(1), 3–12. https://doi.org/10.1177/1525740115627420

Leonard, L. B. (2014). *Children with specific language impairment*. MIT Press.

Levinson, S., Eisenhower, A., Bush, H. H., Carter, A. S., & Blacher, J. (2020). Brief report: Predicting social skills from semantic, syntactic, and pragmatic language among young children with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, *50*(11), 4165–4175. https://doi.org/10.1007/s10803-020-04445-z

Linacre, J. M. (2006). Data variance explained by Rasch measures. *Rasch Measurement Transactions*, *20*, 1045. Lipkin,

P. H., Macias, M. M., Norwood, K. W., Brei, T. J., Davidson, L. F., Davis, B. E., Ellerbeck, K. A., Houtrow, A. J., Hyman, S. L., Kuo, D. Z., Noritz, G. H., Yin, L., Murphy, N. A., Levy, S. E., Weitzman, C. C., Bauer, N. S., Childers Jr, D. O., Levine, J. M., Peralta-Carcelen, A. M., & Voigt, R. G. (2020). Promoting optimal development: Identifying infants and young children with developmental disorders through developmental surveillance and screening. *Pediatrics*, *145*(1). https://doi.org/10.1542/peds.2019-3449

Lipkin, P. H., Okamoto, J., Norwood, K. W., Adams, R. C., Brei, T. J., Burke, R. T., Davis, B. E., Friedman, S. L., Houtrow, A. J., Hyman, S. L., Kuo, D. Z., Noritz, G. H., Turchi, R. M., Murphy, N. A., Allison, M., Ancona, R., Attisha, E., De Pinto, C., Holmes, B., & Young, T. (2015). The Individuals with Disabilities Education Act (IDEA) for children with special educational needs. *Pediatrics*, *136*(6), e1650–e1662. https://doi.org/10.1542/peds.2015-3409

Ma, W., Golinkoff, R. M., Hirsh-Pasek, K., McDonough, C., & Tardif, T. (2009). Imageability predicts the age of acquisition of verbs in Chinese children. *Journal of Child Language*, *36*(2), 405–423. https://doi.org/10.1017/s0305000908009008

Maguire, M. J., Hirsh-Pasek, K., & Golinkoff, R. M. (2006). 14 A unified theory of word learning: Putting verb acquisition in context. *Action meets word: How children learn verbs*, 364.

Markman, E. M. (1989). *Categorization and naming in children: Problems of induction*. MIT Press.

Markman, E. M. (1992). Constraints on word learning: Speculations about their nature, origins, and domain specificity. In M. R. Gunnar & M. Maratsos (Eds.), *Modularity and constraints in language and cognition* (pp. 59–101). Lawrence Erlbaum Associates, Inc.

Markman, E. M., & Hutchinson, J. E. (1984). Children's sensitivity to constraints on word meaning: Taxonomic versus thematic relations. *Cognitive Psychology*, *16*, 1–27. https://doi.org/10.1016/0010-0285(84)90002-1

Merriman, W. E., Bowman, L. L., & MacWhinney, B. (1989). *The mutual exclusivity bias in children's word learning* (pp. i–129). Monographs of the Society for Research in Child Development.

Naigles, L. (1990). Children use syntax to learn verb meanings. *Journal of Child Language*, *17*(2), 357–374. https://doi.org/10.1017/s0305000900013817

Nandakumar, R. (1993). A fortran 77 program for detecting differential item functioning through the mantel-haenszel statistic. *Educational and Psychological Measurement*, *53*(3), 679–684. https://doi.org/10.1177/0013164493053003009

Nelson, H. D., Nygren, P., Walker, M., & Panoscha, R. (2006). Screening for speech and language delay in preschool children: Systematic evidence review for the US preventive services task Force. *Pediatrics*, *117*(2), e298–e319. https://doi.org/10.1542/peds.2005-1467

Nordmeyer, A. E., & Frank, M. C. (2014). The role of context in young children's comprehension of negation. *Journal of Memory and Language*, *77*, 25–39. https://doi.org/10.1016/j.jml.2014.08.002

Pace, A., Curran, M., Van Horne, A., de Villiers, J. G., Iglesias, A., Golinkoff, R. M., Hirsh-Pasek, K., & Wilson, M. (2021). *Sensitivity and Specificity of the QUILS screener for detection of children at risk for DLD*. Poster presented at the conference of the American Speech and Hearing Association.

Paul, R. (1996). Clinical implications of the natural history of slow expressive language development. *American Journal of Speech-Language Pathology*, *5*(2), 5–21. https://doi.org/10.1044/1058-0360.0502.05

Rescorla, L. (2011). Late talkers: Do good predictors of outcome exist? *Developmental Disabilities Research Reviews*, *17*(2), 141–150. https://doi.org/10.1002/ddrr.1108

Reynolds, C. R., Livingston, R. B., & Willson, V. (2009). Measurement and assessment in education.

Rice, M. L., Buhr, J., & Oetting, J. B. (1992). Specific-language-impaired children's quick incidental learning of words: The effect of a pause. *Journal of Speech, Language, and Hearing Research*, *35*(5), 1040–1048. https://doi.org/10.1044/jshr.3505.1040

Rice, M. L., Buhr, J. C., & Nemeth, M. (1990). Fast mapping word-learning abilities of language-delayed preschoolers. *Journal of Speech and Hearing Disorders*, *55*(1), 33–42. https://doi.org/10.1044/jshd.5501.33

Rice, M. L., Cleave, P. L., & Oetting, J. B. (2000). The use of syntactic cues in lexical acquisition by children with SLI. *Journal of Speech, Language, and Hearing Research*, *43*(3), 582–594. https://doi.org/10.1044/jslhr.4303.582

Rice, M. L., & Hoffman, L. (2015). Predicting vocabulary growth in children with and without specific language impairment: A longitudinal study from 2; 6 to 21 years of age. *Journal of Speech, Language, and Hearing Research*, *58*(2), 345–359. https://doi.org/10.1044/2015_jslhr-l-14-0150

Roberts, M. Y., & Kaiser, A. P. (2015). Early intervention for toddlers with language delays: A randomized controlled trial. *Pediatrics*, *135*(4), 686–693. https://doi.org/10.1542/peds.2014-2134

Rowland, C. F., Pine, J. M., Lieven, E. V., & Theakston, A. L. (2003). Determinants of acquisition order in wh-questions: Re-Evaluating the role of caregiver speech. *Journal of Child Language*, *30*(3), 609–635. https://doi.org/10.1017/s0305000903005695

Schlosser, R. W., Shane, H., Sorce, J., Koul, R., Bloomfield, E., Debrowski, L., Neff, A., Miller, S., & Schneider, D. (2012). Animation of graphic symbols representing verbs and prepositions: Effects on transparency, name agreement, and identification. *Journal of Speech, Language, and Hearing Research*, *55*(2), 342–358. https://doi.org/10.1044/1092-4388(2011/10-0164)

Schmeiser, C. B., & Welch, C. J. (2006). Test development. *Educational Measurement*, *4*, 307–353.

Seidl, A., Hollich, G., & Jusczyk, P. W. (2003). Early understanding of subject and object wh-questions. *Infancy*, *4*(3), 423–436. https://doi.org/10.1207/s15327078in0403_06

Seston, R., Golinkoff, R. M., Ma, W., & Hirsh-Pasek, K. (2009). *Vacuuming with my mouth?: Children's ability to comprehend novel extensions of familiar verbs* (Vol. 24, pp. 113–124). Cognitive Development.

Seymour, H. N., Bland-Stewart, L., & Green, L. J. (1998). Difference versus deficit in child African American English. *Language, Speech, and Hearing Services in Schools*, *29*(2), 96–108. https://doi.org/10.1044/0161-1461.2902.96

Shatz, M. (1978). On the development of communicative understandings: An early strategy for interpreting and responding to messages. *Cognitive Psychology*, *10*(3), 271–301. https://doi.org/10.1016/0010-0285(78)90001-4

Sim, F., Thompson, L., Marryat, L., Ramparsad, N., & Wilson, P. (2019). Predictive validity of preschool screening tools for language and behavioural difficulties: A PRISMA systematic review. *PLoS One*, *14*(2), e0211409. https://doi.org/10.1371/journal.pone.0211409

Siu, A. L. (2015). Screening for speech and language delay and disorders in: US preventive services task Force recommendation statement. *Annals of Internal Medicine*, *163*(10), 778–786. https://doi.org/10.7326/m15-2223

Slusser, E., Ribner, A., & Shusterman, A. (2019). Language counts: Early language mediates the relationship between parent education and children's math ability. *Developmental Science*, *22*(3), e12773. https://doi.org/10.1111/desc.12773

Stromswold, K. (1995). The acquisition of subject and object wh-questions. *Language Acquisition*, *4*(1), 5–48. https://doi.org/10.1207/s15327817la0401&2_1

Swingley, D. (2010). Fast mapping and slow mapping in children's word learning. *Language Learning and Development*, *6*(3), 179–183. https://doi.org/10.1080/15475441.2010.484412

Syrett, K., Arunachalam, S., & Waxman, S. R. (2014). Slowly but surely: Adverbs support verb learning in 2-year-olds. *Language Learning and Development*, *10*(3), 263–278. https://doi.org/10.1080/15475441.2013.840493

Thorndike, R. L. (1982). Educational measurement: Theory and practice. The improvement of measurement in education and psychology.

Tomasello, M. (2000). Do young children have adult syntactic competence? *Cognition*, *74*(3), 209–253. https://doi.org/10.1016/s0010-0277(99)00069-4

Tomblin, J. B., Records, N. L., Buckwalter, P., Zhang, X., Smith, E., & O'Brien, M. (1997). Prevalence of specific language impairment in kindergarten children. *Journal of Speech Language Hearing Research*, *40*(6), 1245–1260. https://doi.org/10.1044/jslhr.4006.1245

Voncken, L., Albers, C. J., & Timmerman, M. E. (2021). Bias-variance trade-off in continuous test norming. *Assessment*, *28*(8), 1932–1948. https://doi.org/10.1177/1073191120939155

Wake, M., Tobin, S., Girolametto, L., Ukoumunne, O. C., Gold, L., Levickis, P., Sheehan, J., Goldfeld, S., & Reilly, S. (2011). Outcomes of population based language promotion for slow to talk toddlers at ages 2 and 3 years: Let's Learn Language cluster randomized controlled trial. *British Medical Journal*, *343*(aug18 2), d4741. https://doi.org/10.1136/bmj.d4741

Waxman, S. R., & Hatch, T. (1992). Beyond the basics: Preschool children label objects flexibly at multiple hierarchical levels. *Journal of Child Language*, *19*(1), 153–166. https://doi.org/10.1017/s0305000900013672

Waxman, S. R., & Klibanoff, R. S. (2000). The role of comparison in the extension of novel adjectives. *Developmental Psychology*, *36*(5), 571–581. https://doi.org/10.1037/0012-1649.36.5.571

Windfuhr, K. L., Faragher, B., & Conti-Ramsden, G. (2002). Lexical learning skills in young children with specific language impairment (SLI). *International Journal of Language & Communication Disorders*, *37*(4), 415–432. https://doi.org/10.1080/1368282021000007758

Xue, Y., Bandel, E., Vogel, C., & Boller, K. (2022). Psychometric properties of parent-and staff-reported measures and observational measures of infant and toddler development in Early Head Start. *Early Childhood Research Quarterly*, *61*, 132–144. https://doi.org/10.1016/j.ecresq.2022.06.003

Yurovsky, D., Fricker, D. C., Yu, C., & Smith, L. B. (2014). The role of partial knowledge in statistical word learning. *Psychonomic Bulletin & Review*, *21*, 1–22. https://doi.org/10.3758/s13423-013-0443-y

Zimmerman, I. L., Steiner, V. G., & Pond, R. E. (2011). *Preschool Language Scales* (5th ed.). Pearson. (PLS-5).

Zosh, J. M., Brinster, M., & Halberda, J. (2013). Optimal contrast: Competition between two referents improves word learning. *Applied Developmental Science*, *17*(1), 20–28. https://doi.org/10.1080/10888691.2013.748420

How to cite this article: Jackson, E., Levine, D., de Villiers, J., Iglesias, A., Hirsh-Pasek, K., & Michnick Golinkoff, R. (2023). Assessing the language of 2 year-olds: From theory to practice. *Infancy*, 1–28. https://doi.org/10.1111/infa.12554