8-7-2018

Emerging Language Comprehension in Toddlers with Significant Developmental Delays

Evelyn Fisher

EMERGING LANGUAGE COMPREHENSION IN TODDLERS WITH SIGNIFICANT DEVELOPMENTAL DELAYS: AN IRT APPROACH

by

EVELYN L. FISHER

Under the Direction of MaryAnn Romski, PhD

ABSTRACT

Language comprehension is critical to a wide variety of child outcomes, including academic success and emotional and social well-being. Effective intervention relies on valid, reliable language comprehension data to determine the intensity and techniques that are appropriate for an individual child. The present study investigated language comprehension in a sample of 113 toddlers with significant developmental delays using IRT methods. We found that the aggregate data adequately fit the Rasch model, though each measure also contained items with poor fit. Analyses of the correspondence between item difficulties and participant abilities generally supported the appropriateness of the measures for our sample, and indicated acceptable measurement precision for the majority of participants. Examination of the relative difficulty of

items revealed patterns that were largely consistent with the literature on typically developing children, with a few exceptions. Investigation of individual items showing the highest proportions of change in our sample indicated that parent-report items of moderate difficulty were most likely to reflect language comprehension improvement. Our findings inform clinical practice by underscoring the strengths and limitations of currently available measures. They also inform future measure development by emphasizing the benefits of integrating IRT methods in order to maximize both measurement precision and testing efficiency. Finally, they add to knowledge about language comprehension development in atypical populations.

INDEX WORDS: Developmental delays, Language comprehension, Receptive language, Item response theory, Toddlers, Augmentative and alternative communication

EMERGING LANGUAGE COMPREHENSION IN TODDLERS WITH SIGNIFICANT DEVELOPMENTAL DELAYS: AN IRT APPROACH

by

EVELYN L. FISHER

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

in the College of Arts and Sciences

Georgia State University

2018

EMERGING LANGUAGE COMPREHENSION IN TODDLERS WITH SIGNIFICANT DEVELOPMENTAL DELAYS: AN IRT APPROACH

by

EVELYN L. FISHER

Committee Chair: MaryAnn Romski

Committee: Robin Morris

Rose Sevcik

Elizabeth Tighe

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

May 2019

DEDICATION

For my partner, Evan Wong, whose patience and sense of humor have helped me get through graduate school. Also for my family, especially my parents, Jim and Faye Fisher, who have supported me in all my endeavors.

ACKNOWLEDGEMENTS

I would like to thank my committee members, Drs. Rose Sevcik, Robin Morris, and Elizabeth Tighe, and MaryAnn Romski, for all of their feedback and assistance throughout this project. I would also like to thank my fellow lab members, Devon Killoran and Courtney Everett, for their contributions in reliability coding. The studies were supported by grant DC-03799 From the National Institute on Deafness and Other Communication Disorders and grant R324A070122 from the Institute of Education Sciences.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

1.1 Purpose of the Study

Language comprehension is the ability to understand spoken language. The term is often used interchangeably with receptive language. A child's language comprehension is the combined product of several components, including lexical, syntactic, and pragmatic knowledge. Additionally, factors outside of the domain of language, such as social cognition, perceptual reasoning, memory, and processing speed, play important roles in supporting comprehension and contributing to its development over time.

Two distinct types of language comprehension referred to in the literature are linguistic comprehension, sometimes called pure comprehension, and language comprehension in context, sometimes called pragmatic or discourse comprehension (Miller & Paul, 1995). The former refers to comprehension demonstrated in response to a language stimulus alone, whereas the latter refers to comprehension demonstrated in response to both the language stimulus and other contextual cues, often social cognition-related ones. Both types of language comprehension are highly relevant to the ability to function in everyday life, as situations vary in the amount of contextual support that is available to a person.

Because so much of the information we are exposed to is delivered through the medium of language, language comprehension is critical to a child's ability to learn from the environment and experience a wide variety of desirable outcomes. Baseline language comprehension has proven to be a strong predictor of speech and language outcomes in longitudinal studies of many populations, including late talkers, preterm children, children with developmental disabilities, and pre-lingually deaf individuals who received cochlear implants (Lyytinen, 2005; Romski & Sevcik, 1993; Rousset, Dowell, & Leigh, 2016; Sevcik & Romski, 2005; Suh et al., 2017; Thal

& Tobias, 1991). Additional studies suggest that the influence of language comprehension extends beyond achievement in the domain of language, into psychological well-being and adaptive skills. Yew and O'Kearney (2013) reviewed the literature on emotional and behavioral outcomes among children with specific language impairment (SLI) and found that children with combined receptive-expressive impairments experienced higher levels of internalizing and externalizing psychological symptoms compared to children with only expressive impairments. Howlin, Mawhood, and Rutter's (2000) follow-up study of 20 adults diagnosed with developmental receptive language disorder in childhood indicated that 75% showed moderate or severe social impairments on a measure of adaptive skills.

Despite the importance of language comprehension, there are many challenges in accurately measuring this construct, especially among children with developmental disabilities. First, comorbid impairments or atypical profiles across neurodevelopmental domains complicate the interpretation of performance because factors other than language comprehension affect assessment outcomes. Second, measures designed for typically developing children may fail to discriminate among children with developmental disabilities with varying levels of language comprehension, due to either floor effects or insufficient numbers of items included for the low end of the ability range. Similarly, measures designed for typically developing children may also fail to detect change over time at the comparatively slower rate at which it occurs in children with developmental disabilities. Fourth, children with developmental disabilities are more likely to have difficulty tolerating long assessments, and so non-compliance, short attention spans, and interfering behavior all threaten assessment validity.

The application of item response theory (IRT) analyses to measures of language comprehension has potential to assist in addressing these problems. IRT is defined as a group of

probability models that specify the relationships between individual test items and a latent trait (Hambleton, 1991). Several tools of IRT, including unidimensionality and differential item functioning (DIF) analyses, allow us to test whether or not individual items all measure the same construct across contexts. Thus, IRT can be applied to determine whether or not comorbid impairments exhibited by children with developmental disabilities interfere with valid measurement of a given construct. Additionally, detailed statistics about individual items, including their difficulty levels and performance in discriminating among children with varying levels of the latent trait, allow us to evaluate the appropriateness of the measure for a given population. This can be accomplished by examining whether the measure contains sufficient number of items within the range of latent ability exhibited by children with developmental disabilities. Item-level analyses also allow us to determine the sensitivity of individual items to change over time, an application that is highly relevant to intervention studies including children with developmental disabilities. Finally, taking advantage of item-level statistics allows for the elimination of items with poor psychometric properties and the maximization of instrument validity using the fewest possible numbers of items.

In addition to improving measurement of language comprehension, IRT also has potential to reveal information about the nature of language comprehension and its development in children with developmental disabilities. For example, the order of items in several existing measures of language comprehensions was determined by item difficulty estimates derived from the typically developing normative sample. However, it cannot be assumed that rank order of items by difficulty will remain consistent when the items are applied to children with developmental disabilities. This rank ordering of items by difficulty provides suggestions about

the development of language comprehension over time in children with developmental disabilities by indicating the relationship between items and the latent trait.

The purpose of this study is to explore the construct language comprehension in toddlers with significant developmental delays and limited speech. In order to do this, we examined several measures of language comprehension by applying item response theory (IRT). We seek to 1) determine the psychometric properties of the measures; 2) evaluate the appropriateness of the measures for the sample; and 3) examine the development of comprehension over time. In the following introduction, we review the literature on language comprehension and explain principles and applications of IRT.

1.2 Language Comprehension

1.2.1 Comprehension development in typically developing children

Early language comprehension in infants is often conceptualized as arising from prerequisite skills that can be observed even prior to 8 months of age. These prerequisites include grossly intact hearing, as evidenced by motor responses to noise, and a tendency to pay attention to voices and faces. From 8 to 12 months, infants begin to show comprehension of a few words in the context of familiar routines. For example, an infant may respond to the direction "splash," only in the bathtub (Miller & Paul, 1995). Data from the *MacArthur-Bates Communicative Development Inventories* (CDI; 2006) standardization indicated that average receptive vocabulary more than triples between 8 and 12 months, increasing from 21 to 74 words. Later, at ages 12 to 18 months, infants are increasingly able to demonstrate understanding outside of routines, but comprehension remains limited to words that refer to objects and events in the immediate environment. Receptive vocabulary continues to increase over this time frame, reaching an average of 260 words by 18 months (Fenson et al., 2006).

Importantly, the language comprehension in context evidenced by infants is strongly supported by aspects of social cognition, especially joint attention and imitation skills, as well as learning and memory for routines or typical object-action relationships. Thus, infants sometimes appear to have true linguistic comprehension when they are actually relying on non-linguistic comprehension strategies (ex. following their mother's eye-gaze toward an object).

From 18 to 24 months, comprehension expands to include objects that are out of view as well as some two-word combinations (action-object, agent-action, possessor-possession). From 24 to 36 months, children begin to demonstrate some understanding of three-word constructions (agent-action-object), but have difficulty using information from word order to correctly interpret unlikely sentences (ex. "baby feeds mommy"). They also begin to understand some questions (who, what, where, and whose), spatial concepts (in and under), and comparative concepts (ex. first and bigger). From 36 to 48 months, children's syntactic comprehension expands to include use of word order cues. They also begin understanding "how" questions, and their repertoires of spatial and comparative concepts expand. Receptive vocabulary size is difficult to estimate in young children compared to infants due to large and rapid increases making parent report impractical. However, Chapman (1978) indicated estimates of 500 words at 24 months, 1000 words at 36 months, and 3,000 words at 48 months.

1.2.2 Comprehension in children with developmental disabilities.

Researchers have investigated language comprehension in a variety of specific conditions associated with developmental disabilities. Many of these studies have included comparisons of expressive and receptive language, as well as discussion about how deficits in other domains, especially oral motor and gesture, likely interfere with either the language comprehension development or the ability to demonstrate language comprehension.

Preterm children are at heightened risk of having developmental disabilities due to brain damage secondary to early medical complications. Studies of preterm children suggest that language is generally delayed, though the gap between preterm and full term children in performance on language measures decreases over time from toddlerhood to the school years (Luu et al., 2009). Poorer language outcomes among preterm children are associated with the presence of periventricular leukomalacia (PVL) and/or intraventricular hemorrhage (IVH), especially grade 3 and 4 IVH (Luu et al., 2009). In terms of specific domains of language, a 2011 meta-analysis indicated similar delays across expressive and receptive language skills (Barre, Morgan, Doyle, & Anderson, 2011).

Literature on children with cerebral palsy suggests that their language skills are relatively stronger compared to visuo-spatial skills (Fennell & Dikel, 2001). However, many children with cerebral palsy nonetheless have language impairments secondary to either general intellectual or oral motor impairments (Pirila et al., 2007; Sabbadini, Bonanni, Carlesimo, & Caltagirone, 2001). Researchers have suggested that receptive language may be a relative strength compared to expressive language for a subset of children with cerebral palsy, due to the fact that oral motor impairments, especially dysarthria and anarthria, may limit the development of expressive language (Geytenbeek, Heim, Vermeulen, & Oostrom, 2010). However, expressive and receptive language performance were found to be approximately equal in an epidemiological study of 84 five and six year old children diagnosed with cerebral palsy (Mei et al., 2016).

Literature on the language of children Down syndrome generally indicates relatively stronger receptive language compared to expressive language, which may be attributable to limitations placed on expressive language by severe impairments in articulation or oral motor skills (Luyster, 2011). However, a few studies that examined receptive language in a more fine

grained manner suggest that receptive syntax may also be an area of specific weakness, while receptive vocabulary is stronger (Abbeduto et al., 2003; Næss, Lyster, Hulme, & Melby-Lervåg, 2011).

Research on children with ASD suggests even delays in receptive and expressive language, which are widely understood to be the result of broad deficits in early social cognition that characterize ASD (Yoder, Watson, & Lambert, 2015). However, concerns have also been raised that standardized tests may underestimate the receptive language of children with ASD due to difficulties in gesture prohibiting pointing. A recent study of language comprehension in children with ASD using eye-tracking indicated that this methodology may be able to detect lexical knowledge that would have been missed if another response format was required (Venker, Eernisse, Saffran, & Ellis Weismer, 2013).

Children with any of the conditions above, as well as other conditions, may be appropriate candidates for AAC interventions due to having complex communication needs. Romski and Sevcik (1993) emphasize that, for a speech generating device (SGD) users, comprehension could occur via either of two paths: 1) comprehension of speech 2) comprehension of the visual symbols that are part of the AAC system. Thus, for some children with developmental disabilities, the use of AAC may support the development of comprehension by creating an alternative strategy for the child to acquire symbol-referent relationships in the presence of speech comprehension difficulties.

1.2.3 Measurement of comprehension

Accurate measurement of a child's language comprehension is critical to clinical decision-making for many reasons. First, language comprehension data can indicate whether or

not intervention is warranted. Among otherwise typically developing toddlers with expressive language delay, or late talkers, language comprehension skills distinguish between children at high and low risk of poor language outcomes in preschool (Lyytinen, 2005; Thal & Tobias, 1991), and for this reason, intervention is recommended for late talkers with comorbid receptive delays, but often regarded as unnecessary for late talkers with exclusively expressive delays (Paul, 2000). Second, language comprehension data indicates the degree of impairment among school-age children with language disorders, which can be used to determine the intensity of intervention and supports that are needed. According to Bishop and Edmundson's (1987) hierarchical model of vulnerability in language components, comprehension is foundational, and problems in comprehension characterize the most severe language disorders.

Third, for children who are receiving intervention services, accurate measurement of a child's language comprehension can be helpful in identifying specific techniques and targets. For example, children with strengths in receptive language compared to expressive language may benefit from techniques that focus on elicitation of speech ("Spoken Language Disorders," 2015). Children with combined expressive-receptive difficulties may benefit from techniques that focus on strategies to support comprehension, for example explicit instruction in using contextual cues during exchanges (Miller & Paul, 1995). Among people with complex communication needs, who require augmentative and alternative communication (AAC), language comprehension data may inform the selection of an AAC system according to the level of complexity that is appropriate for the person (Romski, Sevcik, & Adamson, 1997).

Despite its importance, there are many challenges to measuring language comprehension accurately and applying those measurements to conceptualize a child's language comprehension profile in a valid way. These challenges stem from the fact that language comprehension is a

highly complex, multifaceted, and dynamic construct. Complications in interpreting language comprehension data have two sources: 1) characteristics of the measure and 2) characteristics of the child.

1.2.3.1 Characteristics of the measure

Characteristics of the measure that affect the nature of the data it produces include a variety issues related to both measure format and content. In terms of format, one of the most prominent measure characteristics is whether the measure is parent report, directing testing, or a combination of both. There are pro and con arguments regarding the validity and reliability of each of these types. Parent report is helpful in identifying currently emerging behaviors that are infrequent and therefore unlikely to be observed in direct testing. However, parent report may contain inaccuracies due to several problems, like parent difficulties in recalling child behavior, biases toward portraying the child as being either more or less competent, or variability in how the questions are interpreted. The validity of direct testing is supported by the fact that it involves standardized procedures implemented by well-trained observers. However, direct testing is vulnerable to interference from temporary fluctuations in child mood and behavior, which can substantially impact scores among toddlers.

The materials used in the administration of direct testing measures also vary, and may include actual objects, miniatures, and two-dimensional illustrations. Two-dimensional illustrations could further be classified in different ways, including color vs. black and white and photos vs. drawings. Some language comprehension tasks require no materials (ex. compliance with a command like “stand up”). Materials could influence data on a child’s language comprehension via their effect on child motivation. Many young children find actual objects and miniatures more appealing than illustrations, and may be more engaged in those tasks (Pecyna &

Sommers, 1985). Additionally, some children may have difficulty recognizing representations of familiar objects and scenes in the forms of either miniatures or illustrations. This could lead to failure of an item despite adequate linguistic knowledge.

Finally, the type of response required varies across language comprehension items. In infants, responses may include natural reactions, such as turning toward sound or smiling in response to a sing-song voice. In older children, response types may include behavioral compliance, manipulating objects, gazing or pointing at a stimulus, or answering questions. In situations where there is a mismatch between child linguistic knowledge and ability to engage in the required response format, the measure may underestimate the child's skills. For example a child with significant oral motor impairment may be unable to respond to questions with speech even when they know the correct answer.

Measures of language comprehension also vary tremendously in terms of their content. Important aspects of content include the language components that are being probed, such as phonology, vocabulary, grammar, discourse, or pragmatics. Each of those components could further be subdivided in a variety of ways. For example, vocabulary could be broken down into parts of speech (i.e. noun, verb, adjective, etc.).

Content can also be understood as more or less taxing on other domains of cognition. For example, increasing the number of steps in a multi-step instruction causes the item to be more taxing on working memory. The disadvantage of such an item is that it may be more difficult to interpret the meaning of failure, because multiple skills are required for passing. However, it also may also have greater ecological validity, meaning that the item more accurately reflects the types of language comprehension-related tasks that are functionally necessary for everyday life.

Some test authors view skills from domains other than language to be so strongly related to functional outcomes that combining them with language into a single scale is justifiable. For example, the description of the Receptive Language scale of Mullen Scales of Early Learning (MSEL; Mullen, 1995) states that it assesses "auditory comprehension and auditory memory".

The content of items also varies in terms of what types of social or contextual cues are available to the child. As described above, early language comprehension is often context-dependent, and infants and toddlers rely on aspects of social cognition, including joint attention and imitation, in addition to linguistic knowledge, in order to demonstrate understanding. For this reason, many test items for infants and toddlers include cues. These cues are most commonly adult gestures, but also include gaze, demonstration of the correct response (attempting to elicit imitation from the child), voice tone, or placement of materials. Items with cues can be understood to help identify children with weaknesses in the social cognition foundations on which later linguistic comprehension is built.

Finally, one measure characteristic that pertains to both test format and content is the flexibility of the test items, or how much freedom the examiner has to choose the particulars of item administration. Test items could allow for flexibility in terms of who they are administered by (parent or examiner), the number of repetitions given, the materials used, or the language they contain. The disadvantage of greater flexibility is that procedures are less standardized, which could be detrimental to the validity and reliability of the measure. For example, inter-rater reliability may be lowered because two examiners do not make exactly the same choices in modifying items. The advantage of flexibility is that it allows the examiner to adapt procedures in order to explore specific constructs in an individual child, even when the characteristics of that child complicate the assessment. For example, in order to validly assess language

comprehension in a toddler with a very shy temperament, the examiner can allow the parent to administer items.

1.2.3.2 Characteristics of the child

Difficulties in measuring comprehension also stem from characteristics of child. This is particularly apparent when assessing children with intellectual and developmental disabilities. In terms of cognitive development across domains, profiles of strengths and weaknesses often show more variability among children with developmental disabilities, increasing the likelihood that limitations in another domain may interfere with language comprehension measurement. This is especially true for tests developed with typically developing children in mind, whose skills develop in comparatively more predictable patterns. This increased profile variability is due to the wide variety in etiologies of the disabilities, which relate to different neurological substrates. Several examples of the problem of profile variability are contained in the section on the literature language comprehension in children with developmental disabilities, including limitations in gesture among children with autism and oral motor development among children with cerebral palsy and Down syndrome.

In addition to differences in the development of skills that enable children to respond to items, children with developmental disabilities have much higher rates of comorbid sensory impairments, which may interfere with perceiving the materials or stimuli. Fleeting attention spans and interfering behavior may also make standardized assessment administration difficult (Akshoomoff, 2006). Among children with ASD, deficits in social reciprocity may limit the child's understanding of the pragmatics of the testing situation and make it difficult for examiners to achieve adequate motivation. Finally, measures designed for typically developing children may simply be too difficult for children with developmental disabilities. This would

result in floor effects, and ultimately the measure would provide little specific information about the child's language comprehension, other than the fact that it is delayed relative to peers.

1.2.3.3 Available measures

Given the complex nature of language comprehension and the strengths and weaknesses of currently available measures, a multi-method multi-informant approach is favored for generating valid and reliable information that can be used to tailor interventions among children with developmental disabilities (Conti-Ramsden & Durkin, 2012). Several measures are currently available that were designed, at least partially, with the unique needs of children with developmental disabilities in mind.

Use of the Mullen Scales of Early Learning (MSEL) for estimating the abilities of children with developmental delays, especially ASD in particular, is very common in the literature (Mullen, 1995; Zwaigenbaum et al., 2009). The Vineland Adaptive Behavior Scales (VABS) and the Vineland Adaptive Behavior Scales- Second Edition (VABS II) were designed specifically to assess the adaptive behavior of people with developmental disabilities (Sparrow, Ciccheti, & Balla, 1984; Sparrow, Cicchetti, & Balla, 2005). The Clinical Assessment of Language Comprehension (CALC) was designed with the objective of informing interventions for children with disabilities (Miller & Paul, 1995). Several studies have demonstrated that the Sequenced Inventory of Communication Development-Revised (SICD-R) distinguishes between children with a variety of disabilities, including hearing loss, Down syndrome, and autism, and typically developing children (Hedrick, Prather, & Tobin, 1984). The MacArthur-Bates Communicative Development Inventories (CDI) is often used to measure treatment progress in children with developmental disabilities who are older than the children in the normative samples for each form (Fenson et al., 2006). The Peabody Picture Vocabulary Test-Third

Edition (PPVT-III) and Peabody Picture Vocabulary Test-Fourth Edition (PPVT-4) have also been studied in a wide variety of special populations, including children with language delays and children with intellectual disability (Dunn & Dunn, 2007, 1997).

Additional measures of language comprehension in preschool-age children include the Clinical Evaluation of Language Fundamentals-Preschool (CELF-P; Semel, Wiig, & Secord, 2003). The standardization studies of the CELF-P2 included a subset of children with autism spectrum disorder (ASD), and additional studies have used the CELF-P to investigate language development in children with Down syndrome and other developmental disabilities (Liogier d'Ardhuy et al., 2015). Additionally, the Test for Auditory Comprehension of language (TACL; Carrow-Woolfolk, 2013) has been studied in a variety of clinical conditions, including hearing impairment, intellectual disability, and language disorders (Anderson, Hess, & Richardson, 1980; Davis, 1977). However, both the CELF-P and the TACL are intended for children ages 36 months and older. Thus, they are not appropriate for toddlers with developmental disabilities. The most recent version of the Preschool Language Scale (PLS-5; Zimmerman, Steiner, & Pond, 2011) included some children with developmental disorders and/or histories of serious medical conditions in the normative sample. Additionally, other studies of children with autism and other developmental disabilities have also applied this instrument (Hansen, Wadsworth, Roberts, & Poole, 2014; Hobson, Hobson, Malik, Bargiota, & Calo, 2013).

1.3 Measurement

Measurement is the assignment of numbers to objects or events according to systematic rules (de Ayala, 2008). Psychometrics is the study of measurement of mental traits and capacities. These mental traits and capacities are referred to as latent variables because they are not directly observable. Rather, they must be inferred or estimated from observable behavior.

That observable behavior usually takes the form of responses to test items. For example, on the PPVT-4, the latent variable of receptive vocabulary is estimated from the sum of the number of times the child points to the correct picture on a 2 x 2 array when presented orally with a word. In this, and all other psychological testing situations, the examiner does not directly observe the latent variable, but rather observes specific behaviors that we believe to have a relationship to it. Classical Test Theory (CTT) and Item Response Theory (IRT) are two techniques for measuring latent variables. This section reviews psychometric concepts and principles of both CTT and IRT.

1.3.1 Basic concepts in measurement

Reliability is defined as the consistency of test results. Several forms of consistency are relevant to evaluation of the quality of a test, including consistency across time, across examiners, across forms, and among the test items (Whitley, 2003). Consistency over time is referred to as test-retest reliability, and is usually assessed by administering the test to a group of participants twice over a relatively short time. Consistency across examiners is referred to as inter-rater reliability, and is assessed by having two different examiners score the same examinee. Consistency across forms is referred to as equivalent form reliability, and must be assessed in situations where tests developers create alternate forms by administering both forms to a group of participants. Consistency among test items is estimated using either split-half reliability or inter-item correlations.

Validity is defined as the extent to which a test actually measures the latent variable that it purports to measure. Whitley (2003) classifies the types of validity as content-related, criterion-related, and construct-related. Content-related validity refers to the extent to which the content of the measure accurately covers the full breadth of the latent variable. It is often

evaluated by seeking the opinions of experts on that latent variable. Criterion-related validity refers to the degree to which scores on the measure predict scores on a related measure. It is often evaluated by administering participants an additional, more established measure of the same latent variable. Finally, construct-related validity refers to the extent to which the measure reflects current theory about the latent variable. It is often evaluated by combining the evidence about whether or not the measure shows relationships to other latent variables that are consistent with hypotheses derived from the theory. For example, one would expect a valid measure of language comprehension to show a stronger correlation with expressive language compared to fine motor skills.

Another issue related to validity and reliability is measurement invariance. Measurement invariance refers to the idea that a quality test should measure the same latent variable regardless of who is being tested (de Ayala, 2008). In other words, the test should be independent of the participant or sample to which it is applied. For example, items on a questionnaire about social skills could have systematically different relationships to the latent variable social skills in men and women. This would suggest that an issue other than the latent variable, such as differences in contexts and experiences across genders, is interfering with measurement.

1.3.2 Classical Test Theory (CTT)

CTT dates back to early 20th century efforts to measure human intelligence (Spearman, 1904). The focus of statistical analyses in CTT is typically at the level of the whole test. However, simple item-level statistics may also be integrated in CTT, such as calculating the proportion of the normative sample who responded correctly to an item to estimate difficulty. CTT theorizes that an individual's observed score (O) is the sum of the true score (T), or true

latent variable level, and measurement error (*E*), expressed in the equation $O = T + E$. CTT also assumes that *E* varies randomly and does not differ across the latent variable continuum.

CTT remains in use in many places because of its long history, relative simplicity, and ability to be used with small sample sizes. However, there are several important disadvantages of CTT. First, the assumption that measurement error is the same across the latent variable continuum is questionable. Tests are often designed to be most appropriate for people within certain ranges of ability, usually the average range, and without specific attention to item properties at high and low ends of the latent variable continuum, it is unlikely that measurement error is truly equal. Second, both true scores and errors are unknown in the equation included above, and so CTT models are unfalsifiable. For many researchers, the lack of testable assumptions in CTT fully undermines its legitimacy as a scientific theory (Hambleton, 1991). Third, in CTT, individual items are typically weighted equally. Thus, CTT fails to take advantage of item-level statistics that distinguish among items that contain varying amounts of information with regard to the latent variable, which could improve measurement precision. Fourth, the statistical techniques used calculate reliability in CTT, including Cronbach's alpha for internal consistency, result in tests with more items appearing more reliable. Efforts to improve reliability by adding items rather than improving item quality could result in longer testing times.

Finally, the most common critique of CTT is the interdependence of measure and examinee. Each can be influenced by changes in the other. For example, within a CTT framework, observed scores are influenced by the difficulty of the measure, so examinees appear to have higher abilities on easy measures and lower abilities on harder measures. Because of this, it is difficult to compare measures developed with different normative samples and

examinees who were administered different measures. Thus, measures developed using only CTT lack measurement invariance.

1.3.3 Item Response Theory (IRT)

IRT was first described by Lord in 1950. However, its popularity as a psychometric technique grew dramatically in the 1980s and 1990s, as software programs capable of conducting IRT analyses became more widely available (Lord, 1980). Today, IRT is widely used in high stakes entrance exams, including the Graduate Requisite Exam, or GRE (ETS, 2017). Additionally, many modern neurodevelopmental and neuropsychological tests were developed using a combination of both CTT and IRT techniques.

The term IRT can be defined as a group of probability models that express the relationship between the response to an individual item and a latent variable (Hambleton, 1991). In IRT, both items and examinees have locations on the same latent variable continuum. A high quality measure is one that includes a collection of items that are able to precisely differentiate between examinees in all locations on the latent variable continuum.

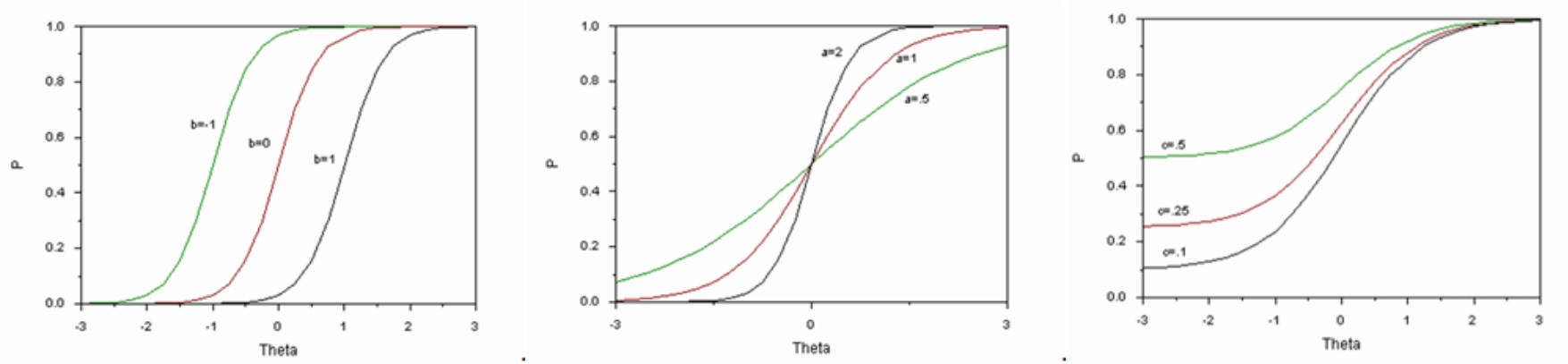
There are several assumptions that must be met for the majority of IRT models. First, measures should be undimensional, assessing only one latent variable. Next, items should also demonstrate local independence, meaning that they are related to one another only via the latent variable. Additionally, items should demonstrate montonicity, meaning that increased levels of the latent variable should be associated with greater probability of a correct response. Last, IRT models assume that the relationship between the latent variable (Θ) and probability of a correct response (p) can be expressed via a mathematical function.

1.3.3.1 Item Characteristic Curves (ICC) and parameters

When the mathematical function of Θ and p is graphed, it is called the Item Characteristic Curve, or ICC. For items with binary response options (i.e. correct or incorrect), the ICC takes the form of a logistic function. ICCs can be characterized by up to three relevant parameters. These parameters are illustrated in Figure 1.1. The first of parameter is item difficulty or *b*. Parameter *b* is the level of Θ, at which an examinee has a 50% chance of responding correctly. Parameter *b* determines the position of the ICC along the x-axis, and the ICC shifts right or left as *b* changes.

The second parameter is discrimination or *a*, which quantifies the item's ability to distinguish between examinees at different levels of Θ. Parameter *a* is visible on the ICC as the slope of the line. Steep slopes indicate that an item is able to discriminate among examinees with relatively small differences in Θ, whereas gradual slopes require larger differences in Θ for adequate discrimination. Higher levels of *a* are desirable because they indicate that a measure is able to make more precise estimates of Θ for the examinee.

The final parameter of interest in IRT is guessing or *c*. Parameter *c* is represented by the position of the line along the y-axis at very low levels of Θ. Parameter *c* illustrates the fact that examinees with low abilities on the latent variable continuum may nonetheless respond correctly by guessing. At increasing levels of *c*, the probability of this occurring is higher. Parameter *c* is particularly important on true/false and other multiple choice format measures.



a. *b* paramter, or difficulty b. *a* parmeter, or discrimination c. *c* parameter or guessing

Figure 1.1 IRT model parameters

1.3.3.2 Advantages of IRT

IRT involves attending to the technical properties of individual items in order to closely scrutinize the relationship between test behavior and latent variable (Embretson & Reise, 2000). This process has many advantages for those seeking to optimize measurement. First, information is known about the ability of the test to discriminate among people at all levels of the latent trait. At the test development stage, this could lead to improvements to address areas of the continuum with poor discrimination. If a test is already in use, users are at least aware of the level of caution that should be used in interpretation of scores for examinees likely to have latent variable levels at the extremes of the continuum.

Second, IRT models are falsifiable; they supply quantitative methods for identifying items that do not meet assumptions and therefore should not be included in the measure. For example, if an item fails to demonstrate monotonicity, this indicates that the item may not reflect the latent variable, and it should be removed. Third, knowledge of the technical properties of items allows for item-weighting to increase precision. Additionally, it simplifies the process of creating equivalent alternate forms of measures. Test developers begin with a bank of items, select pairs with similar parameters, and place one in each form. Fourth, test reliability in an IRT context does not directly depend on the number of items. Rather, it is calculated using item parameters at each level of the latent variable. Because of this, IRT tests are usually shorter.

Finally, the most important advantage of tests developed using IRT models is that they demonstrate measurement invariance. Item properties are examinee-free and latent variable estimates are item-free (Nasir, 2014). This enhances the validity of test results and facilitates comparison of results across examinees who took different tests and tests that were developed with different samples of examinees.

1.3.3.3 Applications of IRT

Test development is the most common context in which IRT is used. Within the area of test development, one of the most well-known IRT applications is computer adaptive testing (CAT). CAT design applies IRT principles to tailor the level of the test items administered to the examinee and derive precise estimates of latent variable in a time-efficient manner. IRT is also frequently used cross-cultural psychology in order answer questions regarding whether an item is valid across groups of participants who differ from one another. This is accomplished via analyses that test for differential item functioning, or DIF. DIF is present when examinees with equal levels of the latent variable from different groups do not have equal probability of responding correctly to an item. Finally, IRT can be used to deepen our understanding of cognition broadly by adding detailed information about the parameters of specific tasks in various populations. For example, Thiruselvam (2016) applied IRT methods to results from the California Verbal Learning Test-Second Edition (CVLT-II) in order to closely examine primacy and recency effects on recall in both healthy and memory clinic-referred adults.

1.4 Research Aims, Questions, and Hypotheses

1.4.1 Research Aim 1

Evaluate the psychometric properties of several language comprehension measures in a sample of children with significant developmental delays by examining the extent to which they meet the assumptions of IRT and appropriately fit an IRT model.

- Question 1: Does the aggregate data fit an IRT model with adequate separation and reliability?
- Question 2: Do the individual items fit the IRT model?

- Question 3: Is the item-level data unidimensional?
- Question 4: Do the items demonstrate local independence?

We hypothesized that the aggregate data would fit an IRT model with adequate separation and reliability. We hypothesized that the majority of individual items would fit the IRT model. We hypothesized that data would be unidimensional. Finally, we hypothesized that the data would show local independence.

1.4.2 Research Aim 2

Determine the appropriateness of the measure for the sample.

- Question 1: Do the items allow for distinguishing among participants with different levels of the latent variable?
- Question 2: For what range of ability levels do the measures show sufficient information?
- Question 3: Does the PPVT-III (4) adequately capture language comprehension in our sample?

We hypothesized that the items would allow for distinguishing among participants at different levels of the latent variable. We hypothesized that the measures would show sufficient information for the majority of our participants. We hypothesized that the PPVT-III (4) would not adequately capture language comprehension in our sample because many participants would be unable to achieve valid basal scores on it.

1.4.3 Research Aim 3

Examine the development of comprehension over time.

- Question 1: What are the characteristics of the items of lowest and highest difficulty?

- Question 2: Is the order of MSEL Receptive Language scale items by difficulty consistent with that of the normative sample?
- Question 3: How does the latent variable, language comprehension, change over time in the sample?
- Question 4: Which items are sensitive to the passage of time?

We hypothesized that the lowest difficulty items would be those that test for responsiveness to sound. The highest difficulty items would include comparative concepts and early numeracy skills. We hypothesized that the order of the MSEL items by difficulty would be consistent with that of the normative sample. We hypothesized that comprehension would improve over time in the sample. We hypothesized that parent report items would be more sensitive to the passage of time because parents are more likely to observe recently emerging changes.

2 METHODS

2.1 Participants

In the present study, we examined language comprehension in a total of 113 children who participated in either of two longitudinal studies of language development in toddlers with significant developmental delays. The overarching goals of these studies were to investigate the communication profiles of toddlers with significant developmental delays and to compare the effectiveness of several parent-implemented interventions designed to improve communication skills. Sixty-two children participated in a randomized comparison of one spoken and two augmented language interventions (Romski et al., 2010). Fifty-one children participated in a subsequent randomized comparison of two augmented language interventions (Romski et al., 2017).

Children were recruited through referrals from a variety of professionals in the Atlanta metropolitan area who frequently provide services to children with developmental delays, including pediatricians, neurologists, speech-language pathologists, and psychologists. Interested parents contacted the principle investigator to discuss participation. Selection criteria included child age between 24 and 36 months at the time of enrollment, at least primitive intentional communication abilities, upper-extremity gross motor skills that enabled the child to touch symbols on a speech-generating communication device, and a primary diagnosis other than delayed speech, hearing impairment, or autism. In addition, eligible participants exhibited significant developmental delays and risk for speech and language impairment, which was operationally defined as being able to produce fewer than 10 intelligible spoken words and having an age-equivalent score of less than 12 months on the Expressive Language scale of the Mullen Scales of Early Learning (MSEL; Mullen, 1995).

2.1.1 Toddlers

See Table 2.1 for a summary of toddler information. The toddler sample consisted of 79 boys and 34 girls. The mean child age at the beginning of the study was 30.6 months (*SD* = 5.3). Medical etiology of developmental delay included a wide variety of conditions, such as Down syndrome, preterm birth, cerebral palsy, and others.

Table 2.1 Toddler Information

	n	%
Gender		
Male	79	69.9
Female	34	30.1
Race		
White	64	56.6
Black or African American	36	31.9
Asian	10	8.8
Other	3	2.7
Medical Etiology		
Unknown or no condition	31	27.4
Down syndrome	28	24.8
Preterm birth	21	18.6
Cerebral palsy	19	16.8
Angelman syndrome	3	2.7
Epilepsy	3	2.7
Mitochondrial disorder	2	1.8
Neurofibromatosis	2	1.8
Other conditions	4	3.5

All toddlers underwent a developmental assessment before beginning the intervention. The average Early Learning Composite standard score on the MSEL observed in our sample was 58.53 (*SD* = 12.11). This score falls in the Very Low range, which was expected based on the inclusion criteria of significant developmental delays. Ninety-five percent of the sample scored more than one standard of deviation below the mean on the Early Learning Composite.

2.1.2 Parents

Each family chose one parent as the designated person who would complete intervention sessions with the child. One hundred and two mothers and eleven fathers participated in the

study. The mean parent age was 37.5 years (*SD* = 5.7). See Table 2.2 for a summary of parent information.

Table 2.2 Parent Information

	n	%
Gender		
Male	11	9.7
Female	102	90.3
Race		
White	68	60.2
Black or African American	36	31.9
Asian	8	7.1
Other	1	0.9
Education*		
High school	9	8.1
Some college	19	17.1
Bachelor degree	51	45.9
Graduate or professional degree	32	28.8

Note: *N = 111 for education because two parents did not report this.

2.2 Procedures

2.2.1 Assessments

See Table 2.3 for a summary of the measures administered at each time point. Before beginning the interventions, children and parents completed a battery of assessments designed to allow researchers to evaluate each child's development across several domains, including communication, visual-spatial skills, motor skills, and adaptive behavior. Parents also completed questionnaires regarding the children's medical and intervention histories. The assessment

battery was re-administered to the dyads immediately following the intervention, three months after intervention, six months after intervention, and twelve months after intervention.

Table 2.3 Measures at Each Time Point

	Pre Intervention	Post Intervention	3 Months Post	6 Months Post	12 Months Post
MSEL	X				
VABS (II)	X				X
CALC	X	X	X	X	X
SICD-R	X	X	X	X	X
CDI	X[a]	X[a]	X[b]	X[b]	X[b]
PPVT-III (IV)					X

Note: MSEL = Mullen Scales of Early Learning, VABS (II) = Vineland Adaptive Behavior Scales (Second Edition), CALC = Clinical Assessment of Language Comprehension, SICD-R = Sequenced Inventory of Communication Development-Revised, CDI = MacArthur-Bates Communicative Development Inventories, PPVT-III (IV) = Peabody Picture Vocabulary Test- Third Edition (Fourth Edition).

[a] CDI Words and Gestures form was used; [b] CDI Words and Gestures form or Words and Sentences Form was used, depending on the child's vocabulary size.

2.2.2 Interventions.

After completing pre-intervention assessments, parent-child dyads were randomly assigned to one of four language interventions across the two studies: Augmented Communication-Input (AC-I), Augmented Communication-Output (AC-O), Augmented Communication-Input and Output Hybrid (AC-IO), or Spoken Communication (SC). See Table 2.4 for a comparison of the interventions. In the AC-I language intervention, the interventionist or parent encouraged the child to use a speech-generating device (SGD) to communicate by modeling SGD use without requiring the child use it. In the AC-O language intervention, the interventionist or parent required the child to use the SGD to produce augmented words through verbal, visual, and

physical hand-over-hand prompts. In the AC-IO intervention, the interventionist or parent both modeled SGD use and required the child to use the SGD to produce augmented words through verbal, visual, and physical hand-over-hand prompts. In the SC language intervention, the parent or interventionist visually and verbally prompted the child to produce spoken words.

Table 2.4 Comparison of Intervention Types

Component	SC	AC-I	AC-O	AC-I/O
Target Vocabulary	I/P and child use speech to communicate	I/P uses the speech-generating device to provide comm. input to child	Child uses the speech-generating device to communicate	I/P uses the speech-generating device to provide comm. input; the child uses speech-generating device to communicate
Mode	Individualized vocabulary of spoken words	Individualized vocabulary of visual-graphic symbols + words	Individualized vocabulary of visual-graphic symbols + words	Individualized vocabulary of visual-graphic symbols + words
Strategies	I/P encourages and prompts the child to produce spoken words	I/P provides vocabulary models to child using the device; Symbols are positioned in the environment to mark referents	I/P encourages and prompts the child to produce communication using the device	I/P provides vocabulary models to child by using the device; Symbols are positioned in the environment to mark referents; I/P encourages and prompts the child to produce communication using the device
Parent Coaching	I provides resource and coaching for P	I provides resource and coaching for P	I provides resource and coaching for P	I provides resource and coaching for P

Note. SC: Spoken Communication; AC-I: Augmented Communication-Input; AC-O: Augmented Communication-Output; AC-I/O: Augmented Communication-Input/ Output; I: Interventionist; P: Parent; I/P: Interventionist or Parent.

All interventions were delivered using the same protocol. Interventions consisted of 24 sessions implemented over an average of 16 weeks. Each session lasted 30 minutes, and consisted of three 10 minute activities: play, book, and snack. The first 18 intervention sessions were conducted in the Toddler Language Intervention Project Lab at Georgia State University. The final 6 sessions were conducted in the child's home. Target vocabulary words for each child were chosen collaboratively by the parent and the project's speech-language pathologist. When a child mastered the use of their target vocabulary set, additional words were added to it.

Over the course of the 24 sessions, parents were taught the intervention and gradually became more involved in its implementation. For the first 8 sessions, the project's interventionist implemented the intervention while the speech-language pathologist explained the techniques to the parent and answered his or her questions. For sessions 9-10, the parent implemented the intervention during the last 10 minutes, or snack. For sessions 11-12, the parent implemented the intervention during the last 20 minutes, or book and snack. Beginning in session 13, the parent implemented the entire 30 minute session, including all three activities. The interventionist continued to coach the parent as needed throughout the all of the sessions.

2.3 Measures

Six measures were used in order to answer the research questions of this study. These included five standardized measures and one non-standardized clinical assessment.

2.3.1 MSEL Receptive Language Scale

The Mullen Scales of Early Learning (MSEL), developed by Eileen Mullen (1995), is a comprehensive measure of cognitive development in infants and children, from birth to 68 months of age (Mullen, 1995). The MSEL is administered in a direct testing format. Most

individual items are rated either 0 or 1 by the examiner, according to whether or not a child is able to accomplish a particular task.

The normative sample included 1849 children. Children with known disabilities were excluded. During test development, a Rasch IRT model was applied to the items from an earlier edition, and items with poor psychometric properties were eliminated. The remaining items were arranged ordinally by difficulty in the measure. In terms of reliability, median internal consistency of the Receptive Language scale across age groups using a split-half method was .80. Test-retest reliability for a subset of 97 children was .82. Construct validity is supported by steady increases in scores from birth to 68 months. Concurrent validity is supported by a correlation of .54 with the Bayley Scales of Infant Development (BSID) Mental Development Index for a subset of 103 children and by a correlation of .85 with the Preschool Language Assessment (PLA) Auditory Comprehension for a subset of 65 children.

For the purpose of this study, we examined the thirty-three items that comprise the Receptive Language Scale. These items include demonstration of comprehension of a wide variety of language concepts, as well as memory for language. They were administered at the pre-intervention time point only.

2.3.2 VABS (II) Receptive Subdomain

The Vineland Adaptive Behavior Scales (VABS) and the Vineland Adaptive Behavior Scales- Second Editions (VABS-II; Sparrow et al., 1984, 2005) are measures of personal and social skills needed for daily living, from birth to 90 years of age. The VABS and VABS-II are both administered in a parent interview format. Items are rated 0 (never/not at all), 1 (sometimes/partially), or 2 (usually/completely) by the parent, according to the extent to which the child exhibits particular behaviors in everyday contexts. Because of the timing of the

updated edition of this measure, we used VABS in the first study and VABS-II in the second study. Before changing to the second edition, we administered both editions to a subset of 12 families and verified that results across editions were consistent.

The normative sample for the VABS consisted of 3,000 participants, including 600 between 2 and 4 years old. No exclusion criteria were used for the VABS. Three percent of school-age children in the normative sample received special education services. Among 2 to 4-year-old children, internal consistency, in the form of split-half reliability coefficients, ranged from .31 to .82 for the Receptive language subdomain. Among a subset of children this age (n = 144), test-retest reliability was reported as .91 to .98 for the Communication domain. Within the entire sample, inter-rater reliability was .99 for the Communication domain. Construct validity of the VABS is supported by steady increases in scores with age. Content validity is support by development of an item pool based on literature on child development, as well as established measures of adaptive behavior. Rasch-Wright analyses were applied to the item pool to eliminate poorly performing items and to arrange items ordinally by difficulty. Concurrent validity is support by correlations between the VABS Communication domain and Kaufman Assessment Battery for Children (K-ABC) ranging from .36 to .53.

The normative sample for the VABS-II consisted of 3,695 participants, including 615 between 2 and 4 years old. No exclusion criteria were used for the VABS-II. The proportion of school-age children in the normative sample receiving special education services varied from 6.7% to 15.4%, depending on the age bracket. Among 2 to 4-year-old children, internal consistency, in the form of split-half reliability coefficients, ranged from .74 to .83 for the Receptive language subdomain. Among a subset of children this age (n = 86), test-retest reliability was reported as .90 to .95 for the Communication domain. Inter-rater reliability for a

subset of 39 children was .67 for the Communication domain. Construct validity of the VABS is supported by careful justification of items using a theoretical framework and steady increases in scores with age. Content validity is support by development of an item pool based on the original VABS, as well as a reconsideration of the competencies that comprise each adaptive behavior domain. Rasch analyses were again applied to the item pool to eliminate poorly performing items, including those that were biased against various demographic groups, and to arrange items ordinally by difficulty. Special attention was also given to how well individual items differentiated between people with and without developmental disabilities. Concurrent validity is support by correlations between the VABS II Communication domain and ABAS Communication domain of .54 for children from birth to 5 years of age.

For the purpose of this study, we analyzed data from 20 Receptive Subdomain items on the VABS II. Nine of these items are identical or nearly identical across the two editions, and thus were administered to both samples. The remaining 11 items were unique to the VABS II, and thus were only administered to the 51 participants in the second study. The items inquire about a variety of behaviors related to language comprehension that the child may exhibit in everyday life. They were administered at the pre-intervention and twelve month follow-up time points.

2.3.3 SICD-R Receptive Scale

The Sequenced Inventory of Communication Development-Revised (SICD-R; Hedrick et al., 1984) is a comprehensive measure of early communication development for ages 4 to 48 months. The SICD-R includes a combination of both parent report and direct testing. Many items include multiple subparts, and children receive different ratings on an age-equivalent scale depending on which subparts they are able to complete.

The normative sample included 252 children. Children were typically developing per parent report. Mean inter-examiner reliability was 96% for a subset of 16 children. Mean test-retest reliability was 92.8% for a subset of 10 children. In terms of construct validity, items were designed to sample a variety of important communication milestones from the literature, and some items were adapted from established measures. Concurrent validity of the SICD-R Receptive Scale is supported by the finding of a Pearson correlation of $r = .81$ with the PPVT for the normative sample.

For the purpose of this study, we examined the thirty-five items that comprise the Receptive Scale. These items include demonstration of comprehension of a wide variety of language concepts. They were administered at all five time points.

2.3.4 CDI Words and Gestures Sections A, B, & D

The MacArthur-Bates Communicative Development Inventories (CDI) Words and Gestures (Fenson et al., 2006) is a measure of early language development for children ages 8 to 18 months. The CDI manual also encourages using the measure for children with developmental delays who are older than 18 months. The CDI is administered in a parent questionnaire format. On parts of the measure related to language comprehension, the parent is asked to respond by checking “yes” or “no” to indicate which words and phrases they believe the child understands.

The normative sample included 1099 children. The number of children in each one-month age bracket varied from 56 to 157. Children were excluded if they had Down syndrome, preterm birth (< 34 weeks gestation), or other serious medical problems. In terms of reliability, Cronbach’s alpha for internal consistency was .95 for the number of words understood. Test-retest correlations were in the mid .80s for all but one age bracket. In terms of construct validity, the authors note that the CDI was designed to sample all major domains of communication in

accordance with the developmental literature, with the exception of phonology. During test development, items that failed to show a pattern of steadily increasing probability of endorsement across child ages were eliminated. Convergent validity is supported by the fact that data from the normative sample was highly consistent with reports in the literature regarding the ages at which milestones in comprehension typically occur. Fenson et al. (2006) report a total of six different investigations of the concurrent validity of the Words Understood total on the CDI. Correlations with other measures of communicative development, including the Reynell Developmental Language Scales (RDLS; Reynell, 1990) Receptive range from .51 to .87.

For the purpose of this study, we examined the 427 items that comprise sections A, B, and D. These items inquire about the child's comprehension of common words and phrases. They were administered at the pre-intervention and post-intervention time points. Data from later time points could not be used, as the more advanced CDI form, Words and Sentences, was administered to participants for whom it was more appropriate at that time.

2.3.5 PPVT-III (4)

The Peabody Picture Vocabulary Test-Third Edition (PPVT-III) and Peabody Picture Vocabulary Test-Fourth Edition (PPVT-4; Dunn & Dunn, 2007, 1997) are measures of receptive vocabulary for ages 2.5 to 90 years. The PPVT-III and PPVT-4 are both administered in a direct testing format. All individual items are rated either 0 or 1 by the examiner, according to whether or not the child points to the correct picture on a 2 x 2 array when presented orally with a word. Because of the timing of the updated edition of this measure, we used PPVT-III in the first study and PPVT-4 in the second study. The PPVT-4 manual describes a study of the relationship between the PPVT-III and PPVT-4 that revealed correlations of .79 to .83 for children, indicating relatively high consistency between the two versions.

The normative sample of the PPVT-III consisted of 3726 participants, including 686 children ages 2 to 4 years. The only exclusion criteria was limited exposure to English. The proportion of school-age children who qualified for special education services ranged from 0.1% to 5.5% under various diagnostic categories. For both editions, during item selection a Rasch IRT model was applied to items from earlier editions, as well as new items selected by a panel of experts. Items with poor psychometric properties were eliminated and the remaining items were arranged ordinally by difficulty. Internal consistency of the PPVT-III, as measured from co-efficient alpha, ranged from .93 to .98 across age groups. Test-retest reliability on a subset of 226 participants ranged from r = .85 to r = .91. Construct validity of the PPVT-4 is supported by steady increases in scores from early childhood to young adulthood. Concurrent validity is supported by correlations of .52 to .70 with the Oral and Written Language Scales (OWLS) Listening Comprehension (Carrow-Woolfolk, 1995).

The normative sample of the PPVT-4 included 3540 participants, including 500 children ages 2 to 4 years. Internal consistency of the PPVT-4, as measured from co-efficient alpha, ranged from .94 to .98 across age groups. Test-retest reliability on a subset of 340 participants ranged from r = .91 to r = .94. Construct validity of the PPVT-4 is supported by steady increases in scores from early childhood to young adulthood. Concurrent validity is supported by correlations of .41 to .77 with the Comprehensive Assessment of Spoken Language (CASL; Carrow-Woolfolk, 1999) and correlations of .67 to .75 with the Clinical Evaluation of Language Fundamentals-Fourth Edition (CELF-4; Semel, Wiig, & Secord, 2003)

For the purpose of this study, we examined the subsets of children who were and were not able to attain basal scores on the PPVT-III or 4. We are not able to examine individual items

due the very limited overlap in items between PPVT-III and 4 at the age level of the children in our sample.

2.3.6 CALC Emerging Language & Developing Language

The Clinical Assessment of Language Comprehension (CALC; Miller & Paul, 1995) is a non-standardized clinical assessment designed to provide information about the language comprehension of children who cannot meet the perceptual, motor, or behavioral requirements necessary for valid test results from standardized measures. The distinguishing feature of the CALC is its flexibility; examiners may use their judgment in selecting the vocabulary, materials, response modalities, and communication partners that are most likely to allow the child to demonstrate the language comprehension concept being tested (ex. response to joint attention). Items can also be repeated according the examiner's judgment. Items are rated as passed or failed by the examiner, according to his or her perception of whether or not the child demonstrates competence on the language comprehension concept.

The CALC was not developed with the use of a normative sample. Empirical studies of validity and reliability of the CALC are not available. Construct validity is supported by the fact that the items of the CALC are based upon the language comprehension development literature. Additionally, the authors of the CALC assert that its design is meant to minimize the common threats to validity that occur when attempting to apply standardized measures to young children with complicated presentations.

For the purpose of this study, we examined nine items of the CALC; seven that comprise the Emerging Language section and two additional items from the Developing Language section. These items include demonstration of comprehension of a wide variety of language comprehension concepts. They were administered at all five time points.

2.4 Data Analysis

We performed preliminary, descriptive analyses in SPSS before proceeding with hypothesis testing. Specifically, we computed and examined means and variance for both raw and standardized scores on the MSEL, VABS (II), SICD-R, and CDI. For the CALC, we examined frequencies of participants passing individual items instead. Additionally, we systematically classified items in order to examine patterns of language comprehension content across measures.

IRT analyses were carried out using Winsteps (Linacre, 2016). In light of the anticipated unidimensional structure of the data and the relatively small sample size, we selected the Rasch model, which includes one parameter. We constructed two separate Rasch models: one for the CDI and one for all other measures combined (MSEL, VABS (II), SICD-R, and CALC). It was necessary to analyze the CDI separately because the number of items it contains (427) was much larger than all other measures (m = 50). As a result, analyzing all data together caused distortions in item parameters for non-CDI items.

2.4.1 Research Aim 1

Evaluate the psychometric properties of several language comprehension measures in a sample of children with significant developmental delays by examining the extent to which they meet the assumptions of IRT and appropriately fit an IRT model.

- We tested the assumption of the aggregate data fitting an IRT model using the separation and reliabilities of separation indices.
- We tested the assumption of individual items fitting the IRT model by both visually examining ICCs for monotonicity and using infit and outfit meansquare (MNSQ) statistics.

- We tested the assumption of unidimensionality using a principle components analysis (PCA) of the residuals after a Rasch dichotomous model was fitted to the data.
- We tested the assumption of local independence using the *Q3* statistic, which is equivalent to residual correlations.

2.4.2 Research Aim 2

Determine the appropriateness of the measure for the sample.

- We plotted ICCs for all items on the same graph. We then looked for indications of redundancy (too many items at a similar difficulty level) and gaps (a lack of items at a difficulty level). We also developed person-item maps, displaying the range of theta in the sample and the range of item difficulty levels on the same scale.
- We calculated and plotted information across the range of theta. This indicated the precision of measurement at varying levels of the latent variable.
- We examined the proportion of children who were and were not able to attain basal scores on the PPVT-III or 4 and compared this to evidence regarding the appropriateness of the other measures derived from the analyses described above.

2.4.3 Research Aim 3

Examine the development of comprehension over time.

- We calculated difficulty parameters for all individual items at the baseline time point. Then, we ordered all items from lowest to highest difficulty. We described patterns in the relationship between item content and difficulty.

- We highlighted similarities and differences between the rank order of MSEL Receptive Language scale items by difficulty derived for our sample compared to the normative sample.
- We examined change in participant theta levels (ability estimates) across the time points.
- We examined the probability of a change in item score (incorrect to correct) for each item from one time point to the next. In other words, we identified items that were the most likely to display change over time.

3 RESULTS

3.1 Descriptive Statistics

Descriptive statistics, including means, variances, and ranges for raw and standardized scores for the MSEL, VABS (II), and PPVT-III (4) are contained in Table 3.1, Table 3.2,Table 3.3. Descriptive statistics for raw CDI Words Understood totals and SICD-R Receptive Language age-equivalent levels are contained in Table 3.4 andTable 3.5. Due to the criterion-based format of the CALC, we calculated frequency statistics, and percentages and numbers of children who successfully completed each item are contained in Table 3.6. We defined floor effect as > 1 participant receiving the minimum raw score possible on a measure. Following this definition, we did not observe floor effects in any of the language comprehension measures, with the exception of the PPVT-III (4). We identified one high outlier using the interquartile range rule (IQR > 1.5) (Field, 2013) in the SICD-R Receptive data at the pre-intervention, post-intervention, and three-month follow-up time points. At each time point, the outlier was the result of data from the same participant, who displayed unusually advanced language comprehension for our sample. We did not observe any other outliers.

Table 3.1 Descriptives for MSEL-Receptive Language Scale

	Raw Score		T-Score	
	Range	m (SD)	Range	m (SD)
Pre-Intervention	4 - 35	17.9 (6.4)	19 - 59	27.6 (11.4)

Note: n = 113. T-score: m = 50; sd = 10.

Table 3.2 Descriptives for VABS (II) Receptive Communication Subdomain

	Raw Score				V-Scale Scores			
	VABS		VABS II		VABS		VABS II	
	Range	m (SD)	Range	m (SD)	Range	m (SD)	Range	m (SD)
Pre-Intervention	8 - 22	16.7 (3.8)	5 - 28	14.9 (6.2)	—	—	4 - 14	9.8 (2.4)
12 Months Post	14 - 24	20.9 (2.8)	10 - 31	23.9 (5.5)	—	—	2 - 15	10.9 (2.8)

Note: n = 62 for pre-intervention VABS, n = 51 for pre-intervention VABS II, n = 48 for 12 months post VABS, n = 44 for 12 months post VABS II. V-scale score: m = 15; sd = 3. VABS does not have standardized scores for subdomains, whereas VABS II does.

Table 3.3 Descriptives for PPVT-III and PPVT-4

	Raw Score				Standard Score			
	PPVT-III		PPVT-4		PPVT III		PPVT-4	
	Range	m (SD)	Range	m (SD)	Range	m (SD)	Range	m (SD)
12 Months Post	9 - 69	32.3 (17.9)	10 - 82	40.0 (19.1)	40 - 110	83.0 (19.6)	43 - 114	81.1 (16.2)

Note: n = 21 for PPVT-III, n = 31 for PPVT-4. Forty-two participants received the test, but were excluded from analyses (29 from PPVT III and 13 from PPVT-4) because they did not attain basal scores. Standard score: m = 100; sd = 15.

Table 3.4 Descriptives for MacArthur-Bates CDI-Words Understood

	Range	m (SD)
Pre-Intervention	2 – 395	131.5 (100.8)
Post-Intervention	3 – 395	192.4 (113.0)

Note: n = 113 for pre-intervention, n = 103 for post-intervention.

Table 3.5 Descriptives for SICD-R Receptive Scale

	Age-Equivalent in Months	
	Range	m (SD)
Pre-Intervention	4 - 40	18.0 (6.6)
Post-Intervention	4 - 49	22.1 (8.1)
3 Months Post	10 - 49	24.3 (8.1)
6 Months Post	12 - 49	26.8 (9.0)
12 Months Post	12 - 63	29.6 (10.1)

Note: n = 113 for pre-intervention, n = 111 for post-intervention, n = 99 for 3 months post, n = 91 for 6 months post, n = 95 for 12 months post. SICD does not use raw and standardized scores in conventional ways. Rather, individuals items are categorized according to the age level at which they are typically mastered. A child's score is the age level at which he or she correctly responds to at least 80% of items.

Table 3.6 Frequencies of Passing Scores on CALC Items

Item	Pre-Intervention	Post-Intervention	3 Months Post	6 Months Post	12 Months Post
	n (%)	n (%)	n (%)	n (%)	n (%)
Familiar Routines	101 (89.4)	107 (96.4)	99 (99.0)	87 (95.6)	92(96.8)
Joint Reference Activity	94 (83.2)	96 (86.5)	91 (91.0)	78 (85.7)	86 (90.5)
Object and Person Names	58 (51.3)	72 (64.9)	82 (82.0)	68 (74.7)	79 (83.2)
Action Words	32 (28.3)	52 (48.6)	62 (62.0)	64 (70.3)	74 (77.9)
Words for Absent Persons and Objects	26 (23.0)	49 (44.1)	52 (52.0)	58 (63.7)	69 (72.6)
Early Two-Word Relations	17 (15.0)	41 (36.9)	49 (49.0)	53 (58.2)	63 (66.3)
Turn-Taking in Discourse	11 (9.7)	27 (24.3)	39 (39.0)	37 (40.7)	47 (49.5)
Two- and Three-Word Relations	7 (6.3)	9 (8.1)	12 (12.0)	14 (15.4)	27 (28.4)
Comprehension of Word Order	0 (0.0)	2 (1.8)	1 (1.0)	3 (3.3)	6 (6.4)

Note: n = 113 for pre-intervention, n = 111 for post-intervention, n = 100 for 3 months post, n = 91 for 6 months post, n = 95 for 12 months post.

3.2 Item-level Data Entry and Processing.

We entered item-level data for all measures at all time points. For the measures with basals and ceilings (SICD-R, MSEL, and VABS (II)), we counted items below the basal as correct and items above the ceiling as incorrect, in accordance with the assumptions of the measure design. The primary investigator and an undergraduate research assistant worked together to complete data entry. First, the primary investigator entered all item-level data. Next the undergraduate research assistant completed double entry of 75% of the data. A comparison of the data indicated that consistency was high (99.7%). We noted a few instances of unclear documentation (e.g. parent placed a check mark between "yes" and "no"), which were resolved via consensus.

3.2.1 Classification of Items

In order to facilitate the interpretation of item-level results, both the primary investigator and an undergraduate research assistant independently classified items on a variety of characteristics. See Appendix A for the results of item classification. The characteristics included linguistic content, administration format, response format, materials, cues, flexibility, field size, and scoring. A comparison of the classification indicated that agreement was high (97.1%). We resolved disagreements via discussion and additional review of test manuals. Additionally, we grouped and highlighted items that were extremely similar across measures in order to explore consistency in results across items in our item-level analyses. See Table 3.7 for descriptions of these groups.

Table 3.7 Similar Item Types Across Measures

	MSEL	VABS (II)	SICD-R	CALC	CDI
Response to sound other than voice	1, 2	1	1, 4, 5, 7		
Response to voice	3, 5, 6	2			
Response to child's own name	10	3	2, 3		A1, D11
Response to inhibitory command ("no")	11	4	9		A2
Response to encouragement ("yes")		5			D12
Response to indication child will be picked up	8		6		B, D18
Comprehension of any words, parent selected	9		8	2.3, 2.5	
Gestures to show comprehension of routine	12		18	2.1	B, D12
Response to "give me" with gesture	13		10		B, D13
Response to "give me" without gesture	15, 17		12C		B, D13
Selects concrete noun from a field of ≥2	14, 19, 21	8	11, 15		
Comprehension of surroundings (e.g. "door")	16				D8
Comprehension of familiar person names			8B, 8E, 16		A3, D11
Identification of body parts	18	7, 11	13		D7
Follows one-step command	20	10, 17	12A&B, 14, 20A	2.4, 2.6	B
Comprehension of locations	22		17		D18
Comprehension of verbs	23		8D		D13
Identification of object function	24		21		
Follows two-step command	25	12	27		
Comprehension of size words	26, 28		23		D15
Comprehension of other adjectives & adverbs			8G, 20B, 20C, 30, 31		D15
Comprehension of comparative concepts	29				D8
Identification of colors	27		25		D15
Follows three-step commands	31	16	33		
Numeracy concepts & quantifiers	32		19, 34		D8
Identification of coins			32		D9
Comprehension of pronouns			8I		D16
Listening skills		6, 14, 15, 19, 20			
Auditory discrimination			24, 29		
Comprehension of word order				3.3, 3.7	

Note. Contents of cells in CDI column refer to questionnaire sections in which multiple examples can be found, not individual items

3.3 Item-Level Analyses

3.3.1 Research Aim 1

Examine the extent to which the measures meet the assumptions of IRT and appropriately fit an IRT model.

3.3.1.1 Reliability

We assessed overall model consistency using separation indices and reliability of separation statistics. Separation indices reflect the number of distinct levels of difficulty (for items) or ability (for participants) that emerge in the data. Separation indices smaller than two are undesirable because they suggest that the items do not adequately cover the continuum of ability and the test is not able to distinguish between participants of different ability levels (Linacre, 2012). Reliability of separation refers to the reproducibility of item parameters, and is interpreted in an analogous manner to Cronbach's alpha in Classical Test Theory, with values ≥ .9 suggesting excellent reliability (Bond, Fox, & Bond, 2007).

Within the non-CDI model, the item separation index for was 5.05, indicating approximately five different levels of difficulty among the items. The reliability of separation index for items was .96, indicating excellent reliability. The person separation index was 8.25, indicating approximately eight different levels of language comprehension ability in our sample. The reliability of separation index for persons was .99, also indicating excellent reliability.

Within the CDI model, the item separation index for was 5.50, indicating approximately six different levels of difficulty among the items. The reliability of separation index for items was .97, indicating excellent reliability. The person separation index was 11.60, indicating approximately 12 different levels of language comprehension ability in our sample. The reliability of separation index for persons was .99, also indicating excellent reliability.

3.3.1.2 Item Fit

We used monotonicity and infit/outfit analyses to evaluate item fit to the Rasch model. Appendix B contains all item fit analyses and Table 3.8 contains a short summary of item fit. We examined empirical ICCs in order to evaluate monotonicity of each item. Each empirical ICC was compared to both its model ICC and the 95% confidence intervals of the empirical ICC. We then classified ICCs into several pre-defined categories: 1) monotonic curve, 2) approximately monotonic curve with one deviation from the 95% confidence interval of the empirical ICC, and 3) non-monotonic curve. See Figure 3.1. Classification of Curves by MonotonicityFigure 3.1 for examples of each of these. Items with no variability in response in our sample (e.g. 100% of participants were either responded correctly or incorrectly) were automatically classified as non-monotonic.

We found that 105 out the 200 items in the non-CDI model (53%) were non-monotonic. Non monotonic items included 27 MSEL items, 29 VABS (II) items, 39 SICD-R items and 10 CALC items. We found that 128 out the 427 items in the CDI model (30%) were non-monotonic. The proportions of non-monotonic items varied among the 20 CDI sections. We observed the lowest proportions in D6 (Clothing; 1 item, 5%), D8 (Furniture and Rooms; 2 items, 8%), and D13 (Action Words; 5 items, 9%). We observed the highest proportions in B (Early Phases; 17 items, 61%), D19 (Quantifiers; 5 items, 63%), and D16 (Pronouns; 9 items, 82%).

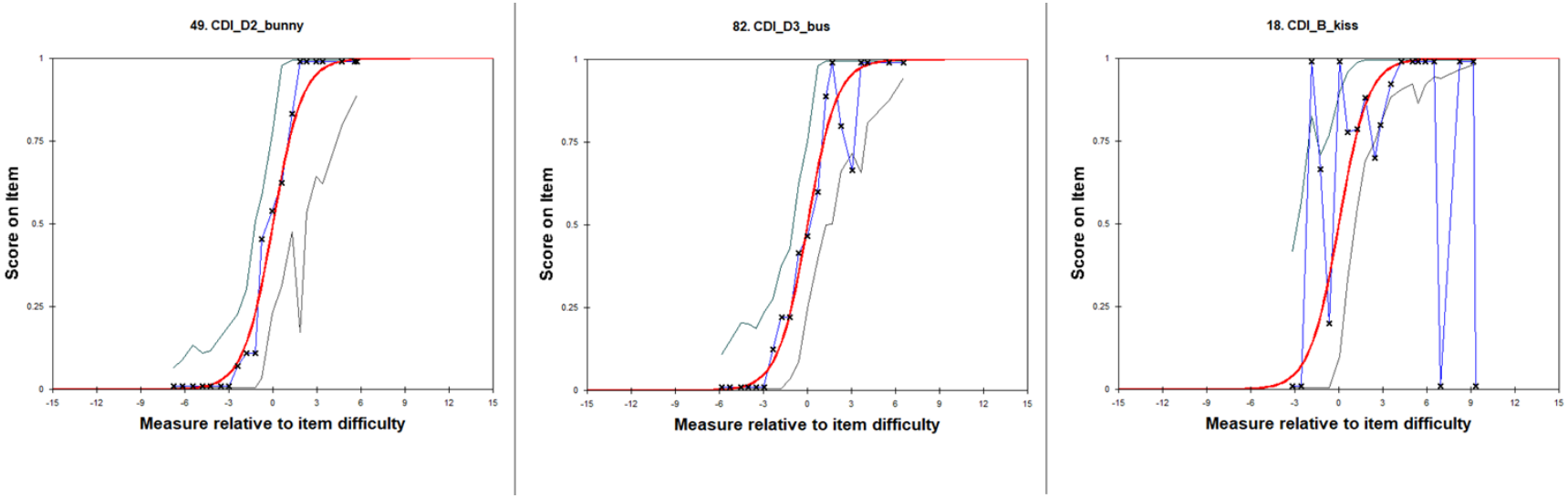
Infit meansquare (MNSQ) and outfit meansquare (MNSQ) are measures of item fit to the Rasch model. Both are sensitive to misfit, or unexpected participant responses. We used the procedure suggested by Linacre (2012) to assess infit MNSQ and outfit MNSQ. Specifically, we first identified a subset of items with infit MNSQ or outfit MNSQ values outside the

recommended range of .5 to 1.5. Next, we examined the z-standardized scores for this subset. We classified items with z-standardized scores > 1.96 or < -1.96 as not fitting the Rasch model.

Table 3.8 Summary of Poorly Fitting Items

Measure or Section	Total Number of Items	Non-Monotonic Items (% of total)	Misfitting Items (% of total)
MSEL	46	27 (59%)	3 (7%)
VABS (II)	40	29 (73%)	10 (25%)
SICD-R	96	39 (41%)	13 (14%)
CALC	18	10 (56%)	4 (22%)
CDI	427	128 (30%)	33 (8%)
B Early Phrases	28	17 (61%)	11 (39%)
D1 Sound Effects	12	7 (58%)	4 (33%)
D2 Animal Names	36	5 (14%)	0 (0%)
D3 Vehicles	9	2 (22%	1 (11%)
D4 Toys	8	1 (13%)	0 (0%)
D5 Food & Drink	30	13 (43%)	0 (0%)
D6 Clothing	19	1 (5%)	0 (0%)
D7 Body Parts	20	5 (25%)	3 (15%)
D8 Furniture & Rooms	24	2 (8%)	0 (0%)
D9 Small Household Items	36	6 (17%)	1 (3%)
D10 Outside & Places	27	3 (11%)	0 (0%)
D11 People	20	10 (50%)	8 (40%)
D12 Games & Routines	19	7 (37%)	0 (0%)
D13 Verbs	55	5 (9%)	0 (0%)
D14 Time Words	8	3 (38%)	0 (0%)
D15 Adjectives	37	18 (49%)	2 (5%)
D16 Pronouns	11	9 (82%)	0 (0%)
D17 Question Words	6	1 (17%)	0 (0%)
D 18 Prepositions	11	6 (55%)	2 (18%)
D19 Quantifiers	8	5 (63%)	0 (0%)

Note. Misfitting refers to items with infit or outfit values outside of specified range (0.5-1.5).



a. Monotonic Curve b. Monotonic with one deviation c. Non-monotonic

Figure 3.1. Classification of Curves by Monotonicity

Note. Blue lines are empirical ICCs, red lines are model ICCs, green lines are 95% confidence intervals

We found that 5 items out of the 200 items in the non-CDI model (3%) demonstrated infit MNSQ problems. These included four VABS (II) items and one SICD-R item. Twenty-five items (13%) demonstrated outfit MNSQ problems. These included three MSEL items, six VABS (II) items, 12 SICD-R items, and four CALC items. We found that 5 out the 427 items in the CDI model (1%) demonstrated infit MNSQ problems. Twenty-eight items (7%) demonstrated outfit MNSQ problems. The proportions of misfitting items varied among the 20 CDI sections, with 12 sections having zero misfitting items. We observed the largest proportions in sections D11 (People; 8 items, 40%), B (Early Phrases, 11 items, 39%) and D1 (Sound Effects; 4 items, 33%).

We were particularly interested in items that displayed extremely poor fit, despite having appropriate difficulty levels for our sample. We defined extremely poor fit as > 2.0 infit MNSQ or outfit MNSQ values, as Linacre (2012) notes that such items likely distort or degrade the measurement system. Items displaying extremely poor fit contain outlier responses, or unexpected cases in which either participants of low ability responded correctly or participants of high ability responded incorrectly. Appropriate difficulty was defined as items to which > 5% of the sample responded correctly and incorrectly. Table 3.9 shows items with extremely poor fit.

We observed several patterns in the extremely poorly fitting items. First, two items (VABS (II) 6 and SICD-R 9) relate to the child's response to an inhibitory command ("no" and "don't touch"). Second, several items related to listening from the VABS (II) displayed extremely poor fit, though only one also met our criteria of appropriate difficulty. Third, many of the items displaying extremely poor fit were from the CDI, which is surprising given that the overall proportion of misfitting items on the CDI was low. In particular, there were many CDI items from sections B (Early Phrases; n = 10) and D11 (People; n = 7).

Table 3.9 Items Displaying Extreme Misfit

Item Description	Item Parameters		Item Fit			Baseline Correct (Prop.)
	Item Difficulty	Difficulty S.E.	Infit MNSQ	Outfit MNSQ	Monotonicity	
VABS (II)						
4. Understands no	-5.15	0.38	1.28	2.56	No	102/113 (0.9)
6. Listens to a story ≥5 min	-0.3	0.29	2.17	2.84	No	46/113 (0.41)
SICD-R						
9. Response to intonation	-4.52	0.33	1.53	6.78	No	97/113 (0.86)
CDI						
Early Phrases						
Change diaper	-2.15	0.25	1.04	9.90	Yes	72/113 (0.64)
Get up	-2.09	0.25	1.39	2.19	No	71/113 (0.63)
Give a hug	-2.73	0.26	1.19	5.27	No	81/113 (0.72)
Give a kiss	-3.48	0.29	1.51	9.90	No	91/113 (0.81)
Good boy/girl	-2.46	0.25	1.14	2.58	No	77/113 (0.68)
Look at this	-1.91	0.24	0.93	2.62	No	68/113 (0.6)
Open your mouth	-0.85	0.24	1.15	7.02	No	50/113 (0.44)
Sit down	-3.01	0.27	1.49	2.04	No	85/113 (0.75)
Spit it out	-0.24	0.25	1.49	5.41	No	40/113 (0.35)
This little piggy	0.46	0.28	1.31	2.98	No	30/113 (0.27)
D1 Sound Effects						
grr	-0.36	0.25	1.41	5.82	No	42/113 (0.37)
uh-oh	-2.27	0.25	1.34	4.75	No	74/113 (0.65)
woof	-1.38	0.24	1.20	2.31	No	59/113 (0.52)
D3 Vehicles						
car	-3.24	0.28	0.81	7.76	Yes	88/113 (0.78)
D7 Body Parts						
mouth	-2.27	0.25	1.09	2.50	No	74/113 (0.65)
nose	-3.24	0.28	1.02	4.57	No	88/113 (0.78)
D9 Small Household Items						

trash	-0.55	0.25	0.90	2.48	No	45/113 (0.4)
D11 People						
babysitter	2.70	0.44	1.18	4.85	No	10/113 (0.09)
babysitter's name	2.19	0.39	1.43	2.29	No	13/113 (0.12)
brother	0.24	0.27	1.82	3.92	No	33/113 (0.29)
daddy	-3.66	0.3	1.30	7.07	No	93/113 (0.82)
grandpa	-0.79	0.24	1.22	3.31	No	49/113 (0.43)
child's own name	-2.94	0.27	1.32	9.90	Yes	84/113 (0.74)
sister	0.03	0.26	1.56	2.64	No	36/113 (0.32)
D15 Adjectives						
dark	2.19	0.39	1.27	2.47	No	13/113 (0.12)
pretty	2.05	0.37	1.56	2.53	No	14/113 (0.12)
soft	1.91	0.36	1.29	3.12	No	15/113 (0.13)
D16 Pronouns						
I	1.34	0.32	1.24	2.36	No	20/113 (0.18)
D17 Question Words						
how	3.12	0.49	0.99	2.17	No	8/113 (0.07)
D18 Prepositions & Locations						
in	-1.03	0.24	1.40	2.06	No	53/113 (0.47)
out	-0.67	0.25	1.49	2.21	No	47/113 (0.42)

3.3.1.3 Unidimensionality

In order to assess the dimensionality of the data, we conducted principle components analyses (PCA) of residuals. Linacre (2012) suggests several guidelines for interpreting the results of this analysis and determining the presence of unidimensionality. First, in unidimensional measures, the observed and expected variance explained by the measures should be roughly similar. Second, the variance explained by measures should be ≥40%. Third, <5% of the remaining variance should be explained by contrasts, or residual components derived from the correlation matrix. Fourth, the strength of contrasts in eigenvalue units should be <2.0. In cases in which the results of the PCA do not fulfill these criteria, the investigators should examine the items contained in the clusters and consider the possibility of multidimensionality. However, the investigators must apply their content knowledge of the latent trait being assessed to make a judgement regarding whether or not the contrasts are truly indicative of multidimensionality.

For the non-CDI data, the observed variance explained by measures was 73%, which was highly similar to the expected variance explained of 72%. This value also exceeds our criteria of 40%. The total variance explained by all contrasts was 7%, with the first and largest contrast explaining 2%. The strength of the first contrast was 11.9 eigenvalue units. Table 3.10 contains the items that comprise the contrast. An examination of the items contained in the clusters revealed that cluster 1 consists entirely of items with high difficulty levels that were correctly completed by only 1 participant in our sample. Cluster 2 consists of preponderance of items related to identification of body parts (6/11), and a variety of other items of varying difficulty levels.

Table 3.10 Dimensionality Analyses, non-CDI measures

First Contrast from Non-CDI Measure Analyses, Outlier Included			
Item Cluster 1		Item Cluster 2	
MSEL 28	Length concepts	MSEL18B	Recognizes ≥1 body part
MSEL 29A	Comparative concepts, ≥3 concepts	MSEL 18C	Recognizes ≥4 body parts
SICD-R 27A	Responds to two action commands, ex. 1	MSEL 19	Comprehends questions II
SICD-R 27C	Responds to two action commands, ex. 3	MSEL 23A	Comprehends ≥1 action word
SICD-R 28A	Understands plurals, example 1	VABS 13A	Follows if-then instructions, partial
SICD-R 28B	Understands plurals, example 2	SICD-R 13 Aa	Body part comprehension, ears
SICD-R 29A	Sound discrimination, high	SICD-R 13 Ab	Body part comprehension, eyes
SICD-R 30	Identification of hard and soft	SICD-R 13 Ac	Body part comprehension, hair
SICD-R 31	Identification of rough and smooth	SICD-R 13 Ad	Body part comprehension, mouth
SICD-R 34A	Understanding of numbers, ex. 1	SICD-R 29B	Sound discrimination, low
SICD-R 34B	Understanding of numbers, ex. 2	CALC 3.7A	Word order, full credit
First Contrast from Non-CDI Measure Analyses, Outlier Removed			
Item Cluster 1		Item Cluster 2	
SICD-R 8C	Understands ≥2 words for clothing [a]	SICD-R 25A	Identification of colors, orange
SICD-R 8E	Understands ≥2 names of acquaintances [a]	SICD-R 25B	Identification of colors, purple
SICD-R 8G	Understands ≥2 adjectives [a]	SICD-R 25C	Identification of colors, red
SICD-R 8H	Understands ≥2 words for household tools [a]	SICD-R 25D	Identification of colors, yellow
SICD-R 8I	Understands ≥2 pronouns [a]	SICD-R 25E	Identification of colors, green
SICD-R 8J	Understands ≥2 words for places [a]	SICD-R 25F	Identification of colors, blue
SICD-R 9	Response to intonation	SICD-R 13 Ab	Body part comprehension, eyes
VABS 9B	Listens to instructions, full credit [a]	SICD-R 13 Bb	Body part comprehension, eyes [a]
VABS 10B	Follows 1-step instructions, full credit [a]	SICD-R 26A	Responds to two object commands, ex. 1
CALC 2.3B	Object/person names, full credit	MSEL 18C	Recognizes ≥6 body parts

[a] Item is parent report only

Because the outlier participant likely impacted the dimensionality analysis, we reran the analysis a second time with this participant removed. In our second analysis, the observed variance explained by measures was 71%, which was highly similar to the expected variance explained of 70%. This value also exceeds our criteria of 40%. The total variance explained by all contrasts was 7%, with the first and largest contrast explaining 2%. The strength of the first contrast was 10.2 eigenvalue units. Table 14 also contains the items that comprise the contrast. An examination of the items contained in the clusters revealed that cluster 1 consists predominantly of parent report items (8/10), especially the subparts of SICD-R 8, in which parents are asked to name examples of words that children understand from a variety of categories (6/10). Cluster 2 consists predominantly of direct assessment items (9/10), especially color identification from the SICD-R (6/10).

The pattern of the second analysis may provide some evidence of multidimensionality in the non-CDI data. However, this should be interpreted cautiously in light of relatively small amount of variance explained by even the largest contrast (2%). Specially, the item clusters in this contrast may indicate that parent report and direct assessment constitute two separate dimensions within language comprehension.

For the CDI data, the observed variance explained by measures was 51%, and the expected variance explained was 51%. This value also exceeds our criteria of 40%. The total variance explained by all contrasts was 8%, with the first and largest contrast explaining 2%. The strength of the first contrast was 18.1 eigenvalue units. Table 3.11 contains the items that comprise the contrast. An examination of these items contained in the clusters revealed that cluster 1 includes adjectives (6/18), time words (3/18), small household items (2/18), verbs (2/18), quantifiers (2/18), and a few other category members. The items tended to have high

difficulty, and most (11/18) displayed poor fit in our infit and outfit MNSQ analyses. The number of parents who endorsed each vocabulary word in this group ranged from 4 to 43. Cluster 2 includes a preponderance of animal names (12/18) and a few words from other categories. The items tended to have low difficulty, and most (17/18) displayed good fit in our infit and outfit MNSQ analyses.

Table 3.11 First Contrast from CDI Measure Analyses

Item Cluster 1		Item Cluster 2	
D9	dish	D1	moo
D9	plant	D2	fish
D10	rock	D2	pig
D13	smile	D2	cow
D13	finish	D2	horse
D14	night	D2	cat
D14	today	D2	bunny
D14	later	D2	elephant
D15	dark	D2	butterfly
D15	pretty	D2	tiger
D15	big	D2	frog
D15	hurt	D2	sheep
D15	soft	D2	teddy
D15	broken	D3	train
D16	me	D4	balloon
D18	away	D5	apple
D19	other	D7	tongue
D19	some	D12	bye

This pattern may provide some evidence of multidimensionality in the CDI data. However, this should be interpreted cautiously in light of relatively small amount of variance explained by even the largest contrast (2%). Specially, the item clusters in this contrast may indicate that animal names constitute a separate dimension within language comprehension, as this was the most consistent observation in item content across the two clusters.

3.3.1.4 Local Independence.

Local independence refers to the assumption that items should be related to one another via the influence of latent variable and no other variables. The Q3 statistic is one method of examining local independence, and is calculated using item residual correlations. Q3 values > .3 are concerning for local dependence (Yen, 1993). Yen (1993) outlines the possible causes of local independence, which include both issues that threaten measurement validity (e.g. external assistance or interference) and issues that may be unavoidable when measuring certain latent variable. For example, on any measure in which items could be placed into subgroups of more specific content areas (e.g. addition problems and subtraction problems on an arithmetic test) one might expect each item to be locally dependent with other items in the same subgroup. Thus, as in the case of interpreting dimensionality analyses, the investigators must examine the content of locally dependent items and consider hypotheses regarding its cause in order to determine whether local dependence poses a serious problem in the data.

Before proceeding with analyses, we removed items to which < 5% of the sample responded either correctly or incorrectly. Such items automatically displayed high residual correlations with one another due to their limited variability in our sample. Next, we examined item pairs with Q3 > .3. For both the non-CDI and CDI analyses, many item pairs had met this criteria (> 700 item pairs in across both analyses). Therefore, we decided to closely examine the 20 item pairs with the highest Q3 values in each analysis in order to capture general patterns.

Table 3.12 contains the 20 item pairs with highest Q3 values in the non-CDI analysis. We made several observations regarding the contents of these item pairs. First, item pairs are consistently from the same measure (20/20). Second, item pairs are usually in the same content area (e.g. color identification, body part identification, adverb comprehension) (15/20). In

particular, many of the item pairs consist of color identification items on the SICD-R. Third, item pairs that are adjacent parent report items display high levels of consistency with one another, even when the content areas vary. This was noted especially with regard to items that comprise SICD-R 8, in which the parent reports whether the child knows any examples of words from diverse categories (pronouns, places, adjectives, and tools).

Overall the local independence analyses of the non-CDI data are encouraging with regard to validity. Most item pairs with high Q3 values appear to be explained by the fact that they are similar in content. The high Q3 values from SICD-R 8, however, are somewhat more concerning because they suggest that item position within the questionnaire and/or response format could also be driving local dependence among items.

Table 3.12 Locally Dependent Items in Non-CDI Analyses

Pair Member 1		Pair Member 2		Q3
VABS 11A	Points to ≥5 body parts, partial [a]	VABS 11B	Points to ≥5 body parts, full credit [a]	1
SICD-R 25A	Identification of colors, orange	SICD-R 25C	Identification of colors, red	.95
SICD-R 25A	Identification of colors, orange	SICD-R 25E	Identification of colors, green	.92
SICD-R 25B	Identification of colors, purple	SICD-R 25E	Identification of colors, green	.90
SICD-R 25C	Identification of colors, red	SICD-R 25E	Identification of colors, green	.88
SICD-R 25B	Identification of colors, purple	SICD-R 25F	Identification of colors, blue	.86
SICD-R 25D	Identification of colors, yellow	SICD-R 25F	Identification of colors, blue	.85
SICD-R 25B	Identification of colors, purple	SICD-R 25D	Identification of colors, yellow	.85
SICD-R 25A	Identification of colors, orange	SICD-R 25F	Identification of colors, blue	.83
SICD-R 25A	Identification of colors, orange	SICD-R 25B	Identification of colors, purple	.83
SICD-R 25A	Identification of colors, orange	SICD-R 25D	Identification of colors, yellow	.82
CALC 2.1A	Familiar routines, partial	CALC 2.2B	Joint reference activity, full credit	.81
SICD-R 8H	Understands ≥2 words for tools [a]	SICD-R 8I	Understands ≥2 pronouns [a]	.81
VABS 9A	Listens to instructions, partial [a]	VABS 10A	Follows 1-step instructions, partial [a]	.80
SICD-R 20B	Responds to commands, walk fast	SICD-R 20C	Responds to commands, walk slowly	.80
SICD-R 25C	Identification of colors, red	SICD-R 25F	Identification of colors, blue	.79
SICD-R 25B	Identification of colors, purple	SICD-R 25C	Identification of colors, red	.79
SICD-R 8G	Understands ≥2 adjectives [a]	SICD-R 8I	Understands ≥2 pronouns [a]	.79
SICD-R 8I	Understands ≥2 pronouns [a]	SICD-R 8J	Understands ≥2 words for places [a]	.78
SICD-R 25C	Identification of colors, red	SICD-R 25D	Identification of colors, yellow	.78

Note. [a] Item is parent report only

Table 3.13 contains the 20 item pairs with highest Q3 values in the CDI analysis. We made several observations regarding the contents of these item pairs as well. First, some item pairs were from the same section/content area (9/20). Second, most individual items in the pairs belonged to a subset of sections, including phrases (18/40) and body parts (9/40). Both of these sections appear early in the questionnaire. Despite the preponderance of items from two sections, relatively few item pairs showed close semantic relationships or both belonged to an obvious superordinate category (e.g. colors: red & blue; body parts: hand & nose) (4/20).

Table 3.13 Locally Dependent Items in CDI Analyses

Pair Member 1		Pair Member 2		Residual Correlation Q3
B	change diaper	B	give a kiss	.92
D15	blue	D15	red	.91
B	open your mouth	D11	child's own name	.87
D7	nose	D11	Daddy	.86
B	give a hug	D3	car	.84
B	look at this	D3	car	.81
B	change diaper	D1	grr	.80
B	change diaper	B	spit it out	.79
D7	hair	D7	nose	.79
B	give a hug	B	look at this	.78
B	give a kiss	D1	grr	.77
B	give a kiss	B	spit it out	.77
D9	dish	D15	pretty	.77
B	open your mouth	D7	nose	.76
B	change diaper	B	this little piggy	.76
D7	head	D7	nose	.75
B	open your mouth	D11	Daddy	.74
D7	hair	D11	Daddy	.74
B	spit it out	D1	grr	.73
D7	hand	D7	nose	.73

Overall, the local dependence data from the CDI are somewhat concerning in that many items pairs with high Q3 values are not explained by similar item content. Moreover, many pairs

are from sections that appear early in the questionnaire, suggesting that item position could be influencing parent responses.

3.3.2 Research Aim 2

Determine the appropriateness of the measure for the sample.

3.3.2.1 Multiple ICCs and item-person maps

We used both plots of multiple ICCs and item-person maps in order to evaluate the congruence between item difficulty and participant abilities in our sample. This process allows us to visually evaluate redundancy (too many items at a certain difficulty level) and gaps (too few items at a certain difficulty level).

Figure 1.1Figure 3.2 displays all ICCs from the non-CDI measures. Item difficulties can be estimated by identifying the point along the x axis at which a participant has a .5 probability of responding correct. The item difficulties in Figure 3 appear to thoroughly cover -8 to +4 logits. This suggests that the items are able to distinguish among participants of varying ability levels within this range. There appear to be many items with overlapping or extremely similar difficulty levels between -4 and 0, suggesting possible redundancy in this range. Gaps appear evident between +5 and +9 logits, which should be interpreted in the context of the item-person map.

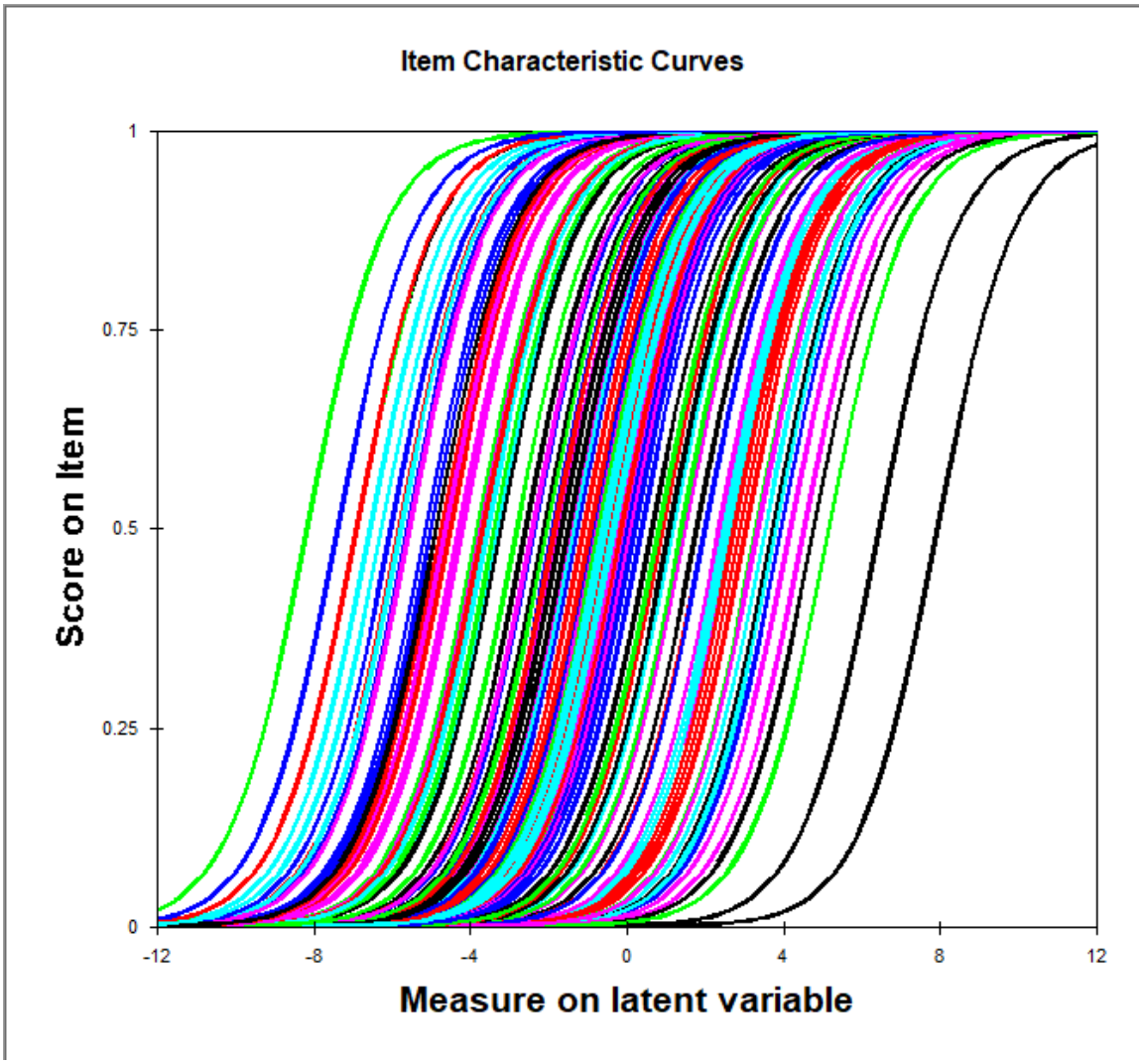


Figure 3.2 All ICCs for non-CDI measures

Note each line is the logistic function for an individual item. X-axis is participant ability in logits. Y-axis is probability of responding correctly.

The item-person map displays both items and participants on the same scale, with participant logits referring to language comprehension ability and item logits referring to item difficulty. When a participant and an item are in the same location on the vertical axis, it indicates that the participant has a .5 probability of responding correctly to the item. Figure 3.3 contains the item-person map for the non-CDI analyses. It shows that two participants have ability levels > 5, and therefore only these two participants are affected by the dearth of items

between +5 and +9 logits. Overall, the appearance of the item-person map suggests that most of the range of participant ability in our sample is well-covered by this item set, with the possible exception of relatively few items at -3 logits. Additionally, there are more problems than necessary at both the very highest (+9 logits) and lowest (< -7 logits) difficulty levels. These difficulty values correspond to items to which all participants responded either correctly or incorrectly.

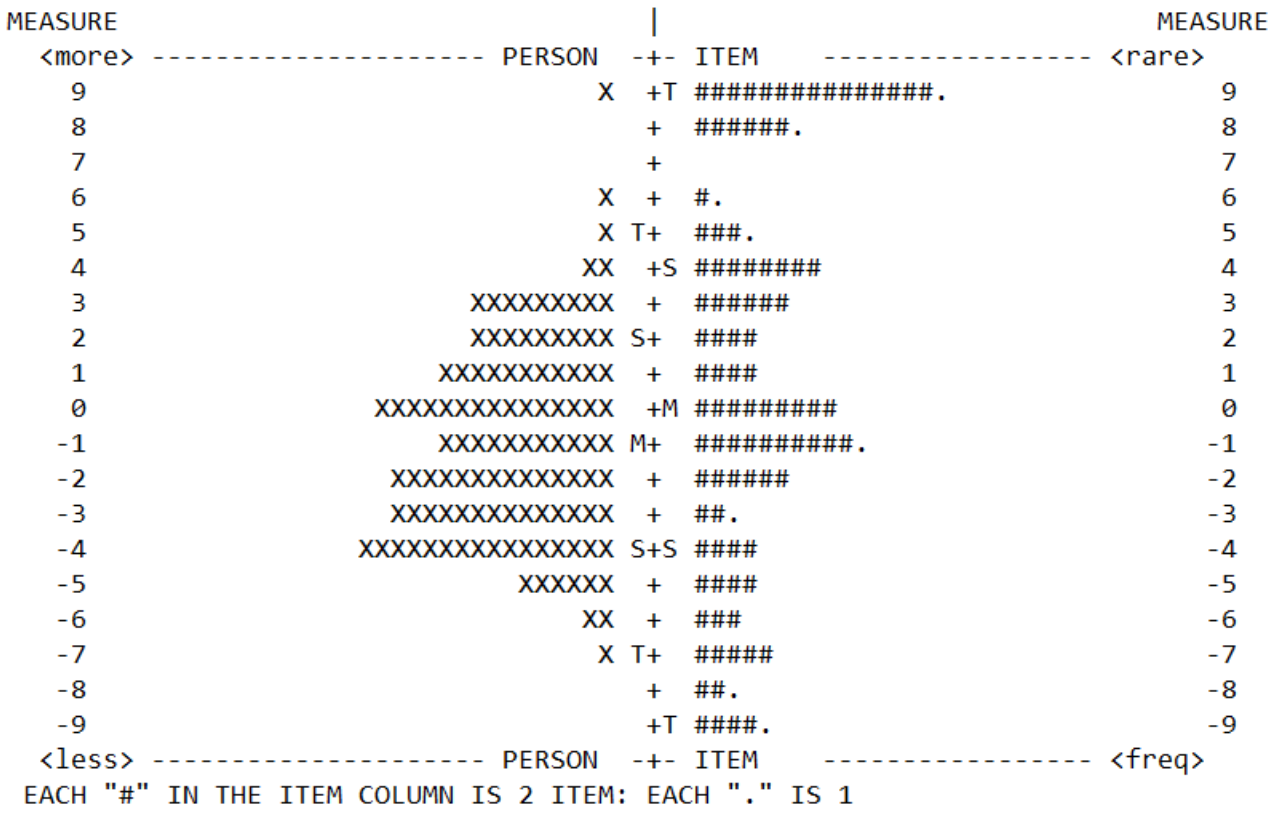


Figure 3.3 Item-person map for non-CDI measures.

Note the left side of the figure displays the distribution of participant language comprehension ability in logits, with each X representing 1 participant. The right side of the figure displays the distribution of item difficulty in logits, with each # representing 2 items and each . representing 1 item. On the center axis M = mean, S= ± 1 standard deviation, T = ± 2 standard deviation.

Figure 3.4 displays all ICCs from the CDI. The item difficulties appear to thoroughly cover -4 to +4 logits. This suggests that the items are able to distinguish among participants of

varying ability levels within this range. There appear to also be many items with overlapping or extremely similar difficulty levels between -4 and +4, suggesting possible redundancy in this range as well. Gaps appear evident for ability levels < -4 logits, which should be interpreted in the context of the item-person map.

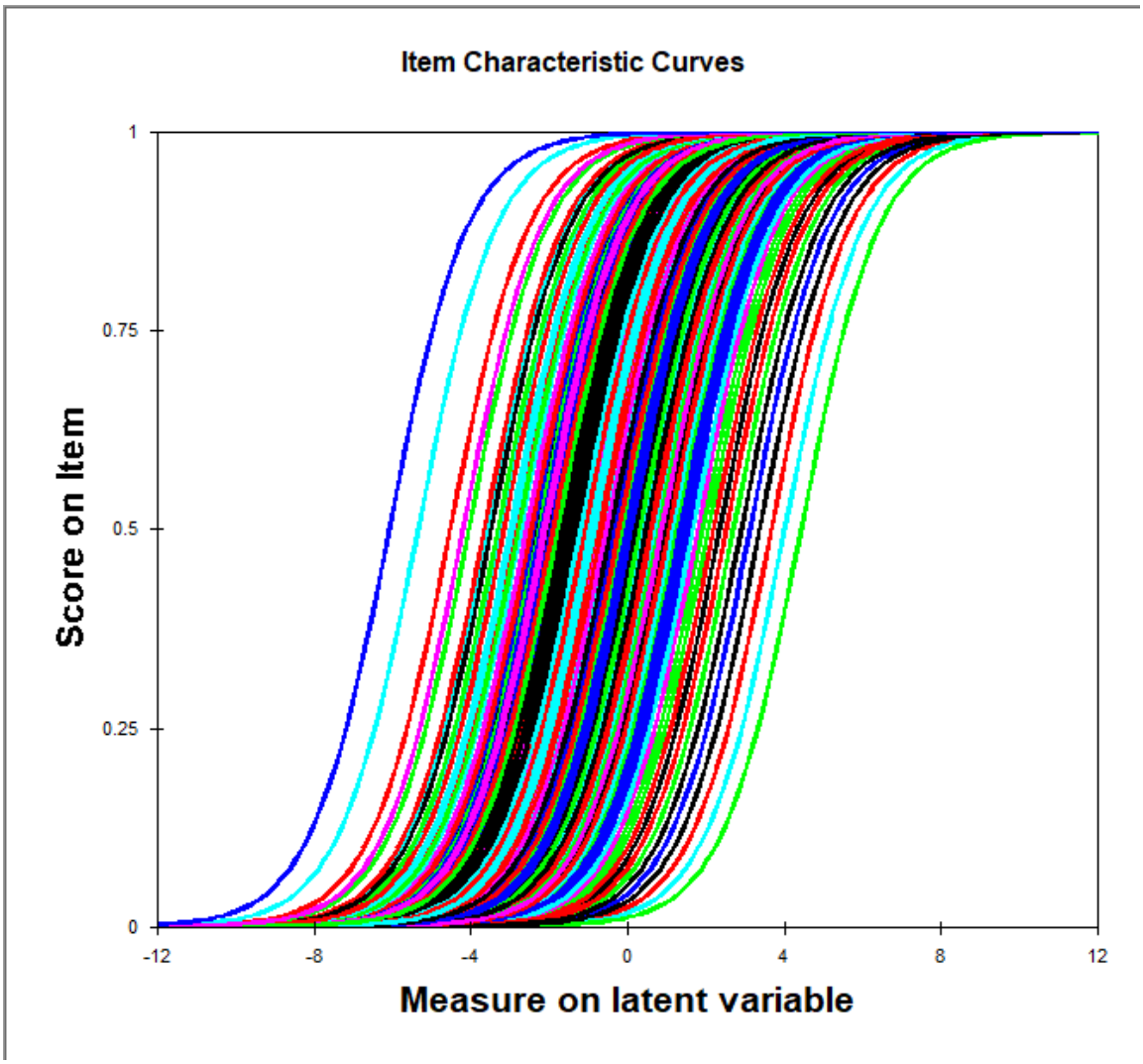


Figure 3.4 All ICCs for the CDI

Note each line is the logistic function for an individual item. X-axis is participant ability in logits. Y-axis is probability of responding correctly.

Figure 3.5 contains the item-person map for the non-CDI analyses. It shows that six participants have ability levels < -4, and therefore are affected by the dearth of items with difficulty levels < -4 logits. It is also evident that three participants have ability estimates higher than the difficulty level of the most difficult item on the CDI. Taken together, these findings suggest that measurement precisions may be impacted for participants with the very lowest and very highest ability levels on the CDI. At the same time, ability levels of most participants (-4 to +4) range, are well-covered by this item set.

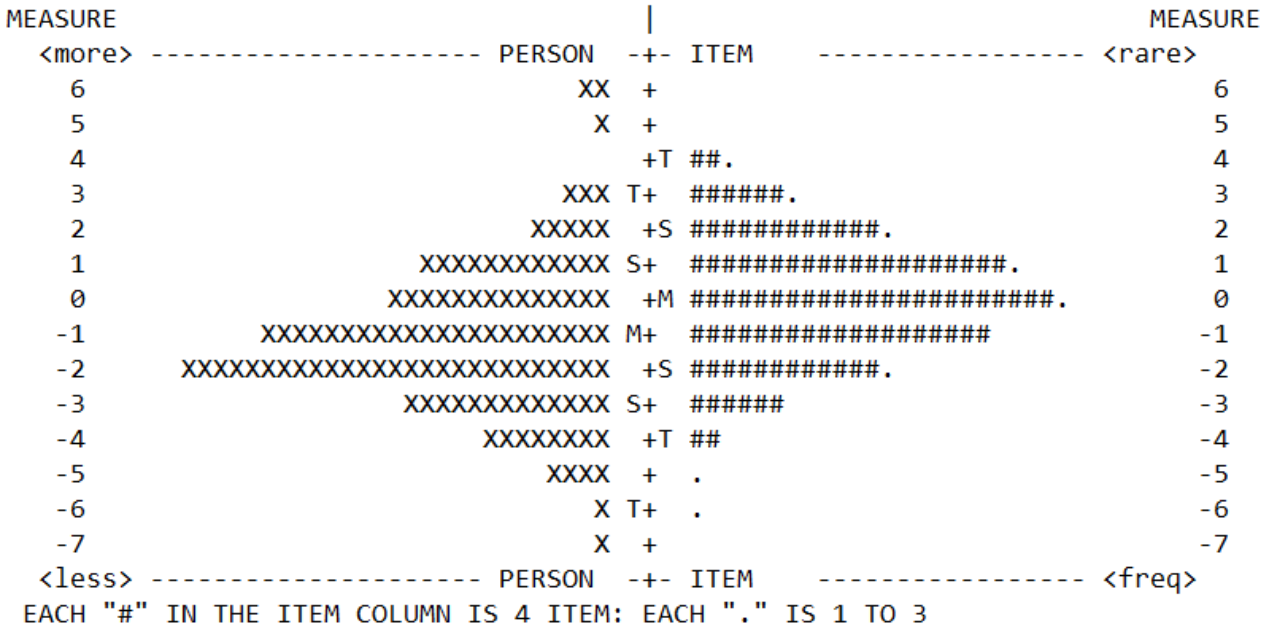


Figure 3.5 Item-person map for the CDI.

Note the left side of the figure displays the distribution of participant language comprehension ability in logits, with each X representing 1 participant. The right side of the figure displays the distribution of item difficulty in logits, with each # representing 2 items and each . representing 1 item. On the center axis M = mean, S= ± 1 standard deviation, T = ± 2 standard deviation.

3.3.2.2 Information and Standard Error

The test information function (TIF) displays the precision of measurement that the test achieves across varying levels of the latent variable, theta (θ). Information (I) is inversely related to standard error (SE) in IRT, such that $SE = 1/\sqrt{I(\theta)}$. Therefore, high information values suggest low SE and good test precision. Low information values suggest high SE and poor test precision. The height of the TIF is determined by the sum of information for individual items, and so it is affected by both measurement precision and the number of items in the test. Embretson and Reise (2000), suggest a guideline of a height of 10 in order to interpret a test as showing adequate information for a specific trait level. This height corresponds to an SE of 0.31 and a reliability coefficient of 0.90.

Figure 3.6 shows the TIF for the non-CDI analyses. The peak of the curve is located at approximately -1 logits. This indicates that this set of items shows maximum measurement precision for participants with language comprehension at this level, and poorer precision for participants with both lower and higher abilities. After applying the guidelines from Embretson and Reise (2000), we found that that the TIF suggests adequate information for participants with ability levels between -3 and +4 logits. This ability range included 68% of our sample (77 out of 113 participants). Also visible in the figure is the irregular or "bumpy" shape of the right-side slope. This suggests the presence of redundancies and/or gaps in the items at certain difficulty levels.

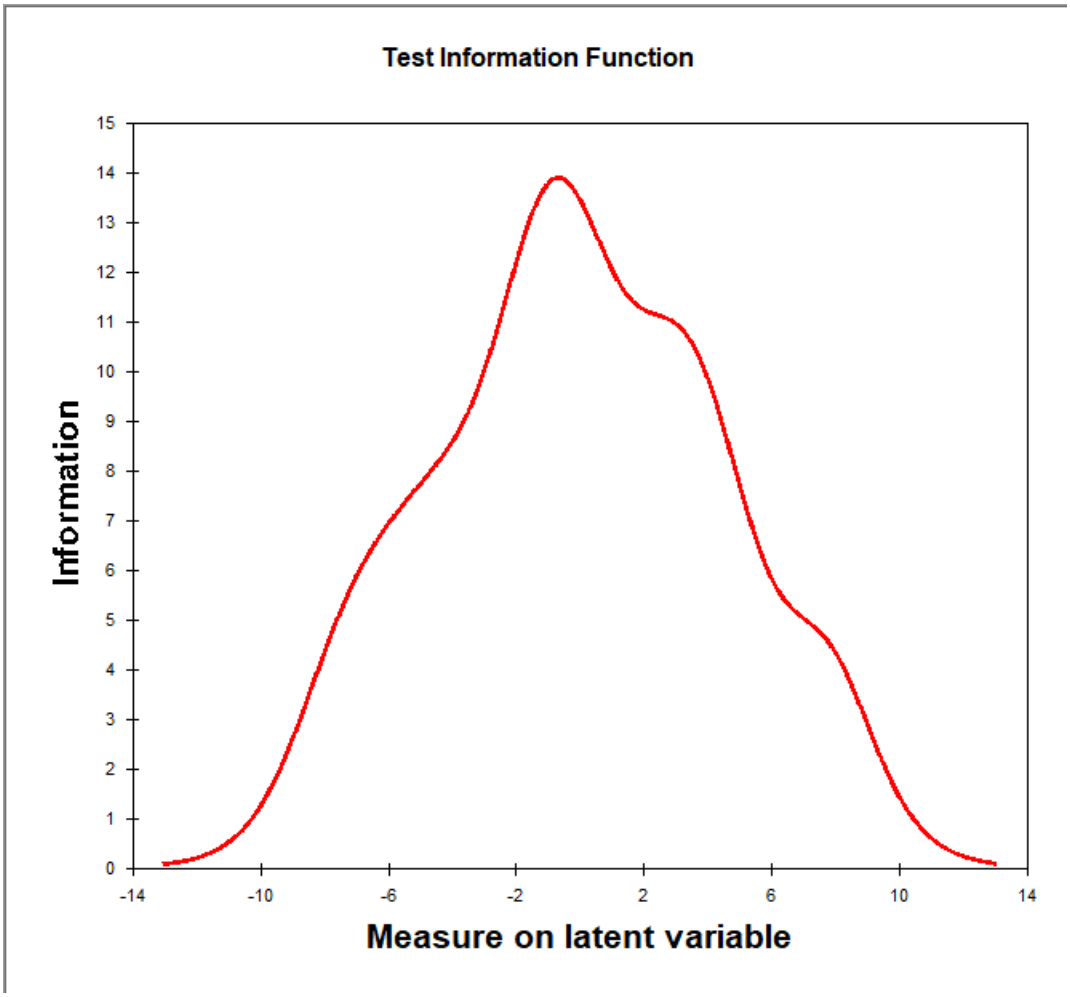


Figure 3.6 Test information function (TIF) for non-CDI measures

Note the Y-axis corresponds to information, which is calculated using the sum of individual item information, $I(\theta) = P_i(\theta)(1 - P_i(\theta))$, at each latent variable level. The X-axis corresponds to participant ability levels on the latent variable language comprehension in logits.

Figure 3.7 shows the TIF for the CDI analyses. The peak of the curve is located at approximately 0 logits. After applying the guidelines from Embretson and Reise (2000), we found that that the TIF suggests adequate information for participants with ability levels between -5 and +5 logits. This ability range included 96% of our sample (108 out of 113 participants).

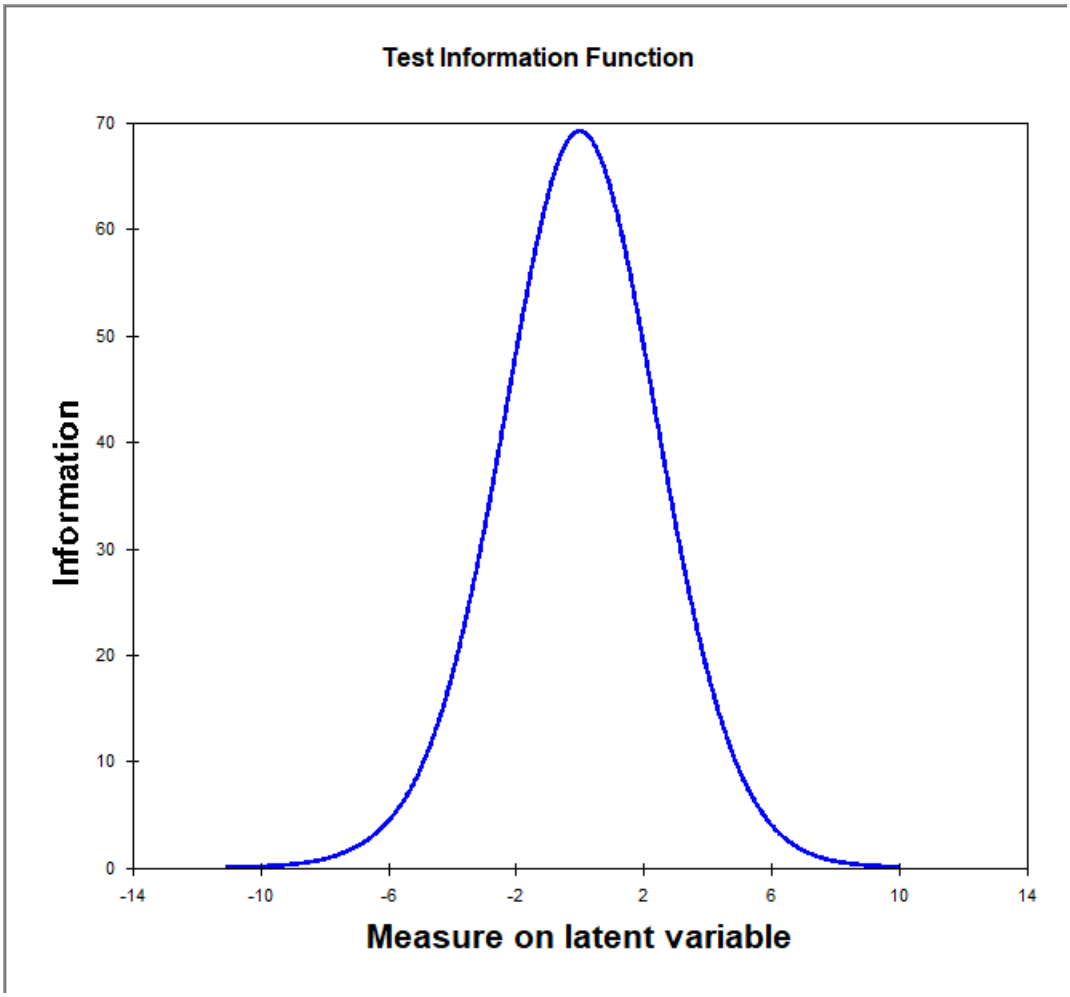


Figure 3.7 Test information function (TIF) for the CDI

Note the Y-axis corresponds to information, which is calculated using the sum of individual item information, $I(\theta) = P_i(\theta)(1 - P_i(\theta))$, at each latent variable level. The X-axis corresponds to participant ability levels on the latent variable language comprehension in logits.

When comparing the CDI TIF to the non-CDI TIF, one notes several things. First, the CDI TIF peak is much higher, which is to be expected given the number of items in this measure. Second, the CDI TIF is symmetrical, suggesting a more even distribution of number of items by difficulty level.

3.3.2.3 Comparison to the PPVT-III (4)

We examined the results of the PPVT-III (4) in the context of our information analyses from other measures in order to characterize the performance of the PPVT-III (4) in measuring language comprehension in our sample. Forty-two of the 94 participants (45%) who were administered the PPVT-III (4) did not attain valid basal scores. Therefore, the data in the Table 3.3 reflect a subset of participants from our sample who likely have relatively stronger language comprehension skills compared to participants who did not attain basal scores. Among participants who did not attain basal scores, the mean raw score on the PPVT III (4) was 9.2 (sd = 16.7), whereas among participants who did attain basal scores, the mean raw score was 64.4 (sd = 36.4). As expected, participants who did not attain basals showed substantially lower Rasch ability estimates (m = -2.8; sd = 1.8) compared to participants who did attain basals (m =0.6; sd = 2.9). Additionally, participants who did not attain basals showed lower baseline MSEL Early Learning Composite standard scores (m = 52.6; sd =6.9) compared to participants who did attain basals (m = 65.5, sd = 13.4).

Among the 42 participants who did not attain basals, 18 had ability estimates < -3 logits. Therefore, in light of our information analyses, their language comprehension skills also would also have lacked adequate measurement precisions using the non-CDI measures. Three participants had ability estimates < -5 logits. Therefore, in light of our information analyses, their language comprehension skills also would have lacked adequate measurement precisions using the CDI.

3.3.3 Research Aim 3

Examine the development of comprehension over time.

3.3.3.1 Item Difficulty at Baseline

Appendix B contains difficulty parameters for all items. For the non-CDI analyses, item difficulty ranged from -9.4 to +9.9 logits (m=0.93; sd = 5.50). The highest difficulty values reflect items to which no participants responded to correctly. The lowest reflect items to which all participants responded correctly. See Table 3.14 for a list of items with the lowest and highest difficulty. The contents of the lowest difficulty items included items that test for responsiveness to sounds or intact hearing. In addition, they included a few items involving response to name, comprehension of "no", joint attention, and interest in mirrors. The contents of the highest difficulty items included comprehension of comparative concepts, responses to general knowledge questions, following three-step commands, identification of coins, early numeracy skills, and comprehension of word-order. Both the lowest and highest difficulty lists contained items with similar content across measures, indicating that participants responded to this content in consistent ways. Table 3.15 contains item difficulty means for each measure. Overall, the CALC displayed the lowest difficulty and the MSEL displayed the highest difficulty.

For the CDI analyses, item difficulty ranged from -6.1 to +4.2 logits (m=0; sd = 1.81). Four participants responded correctly to the highest difficulty items. One hundred and six participants responded correctly to the low difficulty items. See Table 3.16 for a list of items with the lowest and highest difficulty parameters. The contents of the lowest difficulty items included phrases (4), games and routines (3), early signs of understanding (3), toys (2), people (2), verbs (2), and others. The contents of the highest difficulty items included time words (5), adjectives (5), question words (3), quantifiers (3), pronouns (2), and others. Overall, the sections of the CDI containing toys and early phrases displayed the lowest difficulty and the sections containing time words and questions words displayed the highest difficulty.

Table 3.14 Sample of Item Difficulty for Non-CDI Analyses

Lowest Difficulty			Highest Difficulty		
Item		Difficulty	Item		Difficulty
MSEL 1	Reacts to a loud noise	-9.43	MSEL 29B	Comparative concepts, ≥4 concepts	9.91
MSEL 2	Alerts to sound	-9.43	MSEL 29C	Comparative concepts, ≥5 concepts	9.91
MSEL 3	Responds to voice and face (smiling)	-9.43	MSEL 29D	Comparative concepts, ≥6 concepts	9.91
MSEL 4	Coordinates listening and turning	-9.43	MSEL 30A	General knowledge questions, ≥6 correct	9.91
VABS 1A	Turns eyes and head toward sound, partial	-9.43	MSEL 30B	General knowledge questions, ≥7 correct	9.91
SICD-R 1	Responds to sounds [a]	-9.43	MSEL 30C	General knowledge questions, ≥8 correct	9.91
SICD-R 4B	Turns to localize, 135 right side, 1st trial	-9.43	MSEL 30D	General knowledge questions, ≥9 correct	9.91
SICD-R 4B	Turns to localize, 135 right side, 2nd trial	-9.43	MSEL 30E	General knowledge questions, ≥10 correct	9.91
VABS 2A	Looks toward parent, partial [a]	-9.08	MSEL 31	Follows 3 unrelated commands	9.91
MSEL 8	Attends to words and movement	-8.17	MSEL 32	Has concept of numbers, ≥1 correct	9.91
VABS 4A	Understands no, partial [a]	-8.17	MSEL 33	Has concept of numbers, 2 correct	9.91
SICD-R 4A	Turns to localize, 135 left side, 1st trial	-8.17	SICD-R 29C	Sound discrimination, medium	9.91
SICD-R 4C	Turns to localize, 135 left side, 2nd trial	-8.17	SICD-R 32A	Identification of coins, penny	9.91
SICD-R 6A	Responds to come with movement [a]	-8.17	SICD-R 32B	Identification of coins, dime	9.91
MSEL 5	Responds to voice and face (vocalizing)	-7.4	SICD-R 32C	Identification of coins, nickel	9.91
MSEL 6	Coordinates listening and looking	-7.4	SICD-R 33A	Response to 3-step commands, ex. 1	9.91
MSEL 7	Enjoys self/mirror interaction	-7.4	SICD-R 33B	Response to 3-step commands, ex. 2	9.91
VABS 1B	Turns head toward sound, full credit [a]	-6.92	SICD-R 33C	Response to 3-step commands, ex. 3	9.91
CALC 2.2A	Joint reference activity, partial	-6.92	SICD-R 34C	Understanding of numbers, ex. 3	9.91
VABS 3A	Responds to name, partial [a]	-6.91	CALC 3.7B	Word order, full credit	9.91

Note. [a] Item is parent report only

Table 3.15 Measures and CDI Sections: Lowest to Highest Difficulty

Measure		Mean Difficulty (SD)
CALC		0.15 (4.50)
VABS		0.64 (5.16)
SICD-R		1.00 (5.24)
MSEL		1.32 (6.69)
CDI		0 (1.81)
D4	Toys	-1.98 (1.89)
B	Early Phrases	-1.97 (1.31)
D12	Games & Routines	-1.51 (1.34)
D1	Sound Effects	-0.98 (0.77)
D7	Body Parts	-0.89 (1.16)
D3	Vehicles	-0.78 (1.33)
D13	Verbs	-0.52 (1.24)
D8	Furniture & Rooms	-0.19 (0.10)
D6	Clothes	-0.08 (1.65)
D18	Locations & Prepositions	0.05 (1.36)
D9	Small Household Items	0.10 (1.28)
D5	Food & Drink	0.11 (1.32)
D11	People	0.47 (2.29)
D2	Animal Names	0.56 (1.33)
D10	Outside & Places to Go	0.62 (1.42)
D15	Adjectives	1.37 (1.59)
D16	Pronouns	1.89 (1.09)
D19	Quantifiers	2.06 (1.69)
D17	Question Words	2.27 (1.50)
D14	Time Words	3.12 (1.07)

Table 3.16 Sample of Item Difficulty for CDI Analyses

Lowest Difficulty			Highest Difficulty		
Item		Difficulty	Item		Difficulty
A	Responds to his or her name	-6.09	D14	Later	2.9
A	Responds to no	-6.09	D15	Sick	2.9
D4	Ball	-5.35	D19	None	2.9
D11	Mommy	-4.53	D14	Morning	3.12
A	Responds to mommy/daddy	-4.21	D17	How	3.12
B	Come here	-4.05	D19	Other	3.12
D4	Book	-3.66	D8	Playpen	3.37
D11	Daddy	-3.66	D9	Penny	3.37
D12	Bath	-3.66	D16	His	3.37
D12	No	-3.66	D19	Not	3.37
D13	Kiss	-3.66	D16	Her	3.66
D12	Bye	-3.57	D17	Why	3.66
B	Give a kiss	-3.48	D14	Tomorrow	4.01
B	Want more?	-3.48	D17	When	4.01
D6	Shoe	-3.32	D14	Today	4.42
B	Bye-bye	-3.32	D14	Tonight	4.42
D3	Car	-3.24	D15	Fine	4.42
D7	Nose	-3.24	D15	Hard	4.42
D15	All gone	-3.24	D15	Naughty	4.42
D13	Eat	-3.16	D15	Old	4.42

Additionally, we conducted a close examination of the rank order of item difficulty for the MSEL because it was developed using Rasch analysis and arranging items ordinally by difficulty level (Mullen, 1995). Thus comparing our rank order of item difficulty to the order in the measures allows us to comment on possible differences in the difficulty of particular items between our sample and the normative sample. We noted a deviation from the anticipated item order only in cases where this deviation is driven by a difference of > 5 participant responses, due to possibility that small deviations may result from chance. The results of our analysis of the

MSEL indicated that three items displayed higher difficulty than anticipated: MSEL 10 (Response to Name), MSEL 17 (Follows Directions), and MSEL 20 (Follows Related Commands).

3.3.3.2 Change in Ability Estimates

In order to explore growth in language comprehension over time, we examined change in Rasch ability estimates in our sample across the five assessment time points. The data used in this analysis included only the two measures that were administered at all time points: the SICD-R and CALC. Like the difficulty parameter, Rasch ability estimates are also expressed in logits, with higher numbers indicating higher abilities and lower numbers indicating lower abilities.

Table 3.17 contains a summary of the Rasch ability data across the five time points. As expected, Rasch ability estimates consistently increased from one time point to the next, indicating growth in language comprehension over time. The degree of growth appeared relatively steady, with the largest increase appearing between baseline and post-intervention (1.55 logits), and slightly smaller increases following (Post to three-months = 1.15, three months to six months = 1.14, six month to twelve months = 1.15).

Table 3.17 Rasch Ability Estimates Based on SICD-R and CALC

Time Point	M (SD)
Baseline	-19.74 (2.96)
Post-Intervention	-18.19 (3.48)
Three Months Post	-17.04 (3.43)
Six Months Post	-15.90 (3.80)
Twelve Months Post	-14.75 (3.77)

Note. n = 113 at baseline, n = 112 at post-intervention, n = 100 at three months post, n = 91 at six months post, and n = 95 at twelve months post.

3.3.3.3 Probability of Change for Items.

In order to identify which items were more and less sensitive to change over time, we calculated the probability of the combination of an incorrect response at baseline and a correct response at post-intervention. We conducted this analysis for all items from the three measures administered at both baseline and post-intervention: the SICD-R, CALC, and CDI.

Table 3.18, Table 3.19, and Table 3.20 and contain the results. Broadly, items from each measure that showed the highest rates of change tended to be moderate in difficulty, with both easy and hard items showing lower rates of change. On the CALC, items showing the highest rates of change included turn-taking, two-word relations, absent persons/objects, and action words. On the SICD-R, item showing the highest rates of change included body part identification (n = 4), parent report of the child understanding of any words from certain categories (n = 4), one-step commands (n = 3), object functions (n = 2), numeracy concepts (n = 2), object identification (n = 2), and turn-taking (n = 1). Overall, the items included equal proportions of both direct assessment and parent report items. However, it is worth noting that in instances in which there were analogous direct testing and parent report versions of the same items, the parent report versions showed higher rates of change (SICD-R 13). On the CDI, items showing the highest rates of change included verbs (n = 8), adjectives (n = 6), prepositions & locations (n = 2), animals (n = 2), and a variety of single items from additional categories. In reviewing the content of items, we also observed semantically related items that appeared across categories, specifically the words bus, teacher, and school.

Table 3.18 Items with Highest Proportions of Change on the CALC

Item		Increase in Proportion Correct
CALC 2.7A	Turn-taking, partial	.31
CALC 2.6B	Two-word relations, full credit	.23
CALC 2.5B	Absent persons/objects, full credit	.22
CALC 2.5A	Absent persons/objects, partial	.21
CALC 2.4B	Action words, full credit	.20
CALC 2.6A	Two-word relations, partial	.20
CALC 2.4A	Action words, partial	.20

Note. Shading intended to highlight proportions that are identical. All items with proportion ≥.20 included.

Table 3.19 Items with Highest Proportions of Change on the SICD-R

Item		Increase in Proportion Correct
SICD-R 14B	Response to stand up and sit down [a]	.22
SICD-R 19B	Responds to number concepts, all	.21
SICD-R 22	Takes turns [a]	.21
SICD-R 12B	Response to situational commands, put down	.21
SICD-R18	Responds to bye-bye	.21
SICD-R 15E	Speech discrimination, box	.21
SICD-R 8G	Understands ≥2 adjectives [a]	.21
SICD-R 13Bb	Body part comprehension, eyes [a]	.21
SICD-R 11A	Response to specific word	.21
SICD-R 21B	Understands object function, shoe	.21
SICD-R 21C	Understands object function, book	.21
SICD-R 12C	Response to situational commands, give	.21
SICD-R 8J	Understands ≥2 words for places [a]	.21
SICD-R 8H	Understands ≥2 words for household tools [a]	.21
SICD-R 13Bc	Body part comprehension, hair [a]	.21
SICD-R 19A	Responds to number concepts, one	.20
SICD-R 8I	Understands ≥2 pronouns [a]	.20
SICD-R 13 Ba	Body part comprehension, ears [a]	.20
SICD-R 13 Bd	Body part comprehension, mouth [a]	.20

Note. [a] Item is parent report only. Shading intended to highlight proportions that are identical. All items with proportion ≥.20 included.

Table 3.20 Items with Highest Proportions of Change on the CDI

Section	Item	Increase in Proportion Correct
D13	Read	0.32
D1	Yum	0.30
D11	Teacher	0.30
D15	Clean	0.30
D19	More	0.30
D3	Bus	0.29
D8	Bathroom	0.29
D10	School	0.29
D13	Drink	0.29
D13	Help	0.29
D13	Look	0.29
D16	Mine	0.29
B	Open your mouth	0.28
D13	Jump	0.28
D13	Love	0.28
D13	Play	0.28
D15	Asleep	0.28
D15	Good	0.28
D18	Out	0.28
D2	Horse	0.27
D2	Monkey	0.27
D5	Drink	0.27
D9	Bowl	0.27
D9	Plate	0.27
D12	Want	0.27
D13	Ride	0.27
D15	Cold	0.27
D15	Empty	0.27
D15	Hungry	0.27
D18	Inside	0.27

Note. Shading intended to highlight proportions that are identical. All items with proportion ≥.27 included.

4 DISCUSSION

The present study investigated language comprehension in a sample of toddlers with significant developmental delays using IRT methods. We found that the aggregate data

adequately fit the Rasch model, though each measure also contained individual items with poor fit. Dimensionality analyses provided some evidence of multidimensionality in both models, though this should be interpreted cautiously because the variance explained by contrasts was low. Local independence analyses indicated that many item pairs were correlated beyond the extent that would be expected based on the latent variable.

Analyses related to the correspondence between item difficulties and participant abilities generally supported the appropriateness of the measures for our sample, and indicated acceptable measurement precision for the majority of participants. At the same time, these analyses also revealed several areas where minor improvements are possible. For example, adding a few items of moderately low difficulty would improve measurement precision for participants of moderately low language comprehension abilities.

Examination of the relative difficulty of items indicated patterns that were largely consistent with the literature on typically developing children, with a few exceptions. Participant Rasch ability estimates consistently increased from each assessment to the next, indicating growth in language comprehension ability over time. Investigation of individual items showing the highest proportions of change in our sample indicated that parent-report items of moderate difficulty were most likely to reflect language comprehension improvement.

4.1 Descriptive Statistics

Standardized scores indicated significant delays in receptive language, as expected given our study inclusion criteria. Nonetheless, there were no floor or ceiling effects on MSEL, VABS (II), SICD-R, CALC, or CDI, which provided preliminary support for the appropriateness of these measures for our sample before analyzing the data using IRT methods. Additionally, our examination of raw scores for the SICD-R and CALC at multiple time points revealed steady

increases, suggesting these measures are able to capture improvements in language comprehension over time in our sample of young children with significant developmental delays.

4.2 Classification of Items

Our classification of items indicated both areas of overlap and divergence among the measures in terms of the specific skills tested. Areas of overlap were consistent with extant knowledge about typical language comprehension development. For example, precursors to language comprehension, such as intact hearing and interest in voices, were tested in items that appeared in the beginning of several measures. These items were often followed by recognition of the child's own name, inhibitory commands, and common routines that the parent reports are familiar to the child. Additionally, several measures included items that test comprehension of word-gesture combinations, followed by analogous items that test comprehension of words only. Such combinations clearly seek to explore the child's ability to use non-linguistic or social cognition-related comprehension strategies verses pure linguistic comprehension. Inclusion of progressively longer single and multi-step commands was also common across measures. Such items incorporate both language comprehension and working memory. Additionally, comprehension of specific material frequently taught to toddlers, such as body parts and colors, was explored in several measures. Finally, more complex, preschool-appropriate content, including comparative concepts, numeracy concepts/quantifiers, and various descriptors, often appeared toward the end of measures.

In addition to their commonalities, each measure also included content of unique importance to the particular test developers. For example, the MSEL's advanced items included "why" questions (e.g. "why do we have refrigerators?"). Such items both require sophisticated expressive language and relate a child's verbal reasoning ability. The VABS (II) included

several items related to listening skills. This feature may relate to its role as an adaptive functioning measure and the importance of listening skills to adaptive functioning for older children, especially in school environments. The SICD-R was unique in its inclusion of auditory discrimination items. This interest in auditory discrimination is the result of the theoretical perspective of its authors, who view receptive language as being comprised of three components: awareness, discrimination, and understanding (Hedrick et al., 1984). The CALC included a focus on comprehension of word-order, or syntax. The authors of the CALC view comprehension of syntax as a pivotal skill needed to move young children from the "developing language" to "using language for learning" stages (Miller & Paul, 1995). The CDI included time words. A child's comprehension of such words may be difficult to explore in a direct testing context, especially without requiring the use of expressive language. Thus, a parent-completed vocabulary inventory may be both a valid and efficient way to gain information regarding the child's comprehension of the time words in everyday contexts.

4.3 Item-Level Analyses

4.3.1 Research Aim 1

Examine the extent to which the measures meet the assumptions of IRT and appropriately fit an IRT model.

4.3.1.1 Reliability.

Consistent with our hypotheses, we found that the separation indices and reliability of separation statistics support the efficacy the measures. Across our two Rasch models, the results indicated that approximately five-to-six different levels item difficulty and eight-to-eleven different levels of participant language comprehension ability emerged. Thus, our measures are successfully able to differentiate among participants' ability levels.

4.3.1.2 Item Fit.

Despite the promising data regarding the overall model, we also identified many misfitting items using our monotonicity and infit/outfit analyses. In general, the monotonicity analyses identified misfitting items that displayed too little variability in our sample. In other words, items that lacked monotonicity tended to be either too easy or too difficult for our sample; the vast majority of participants responded either correctly or incorrectly to them. Consistent with this explanation, the VABS (II) had the highest proportion of non-monotonic items. This measure contains a wide range of item difficulty, which is necessary to assess language comprehension across the lifespan. Conversely, the CDI displayed the lowest proportion of non-monotonic items. This measure contains a more limited range of item difficulty, and was intended for children of developmental levels from 8 to 18 months.

Infit/outfit analyses identified items on which participants displayed unpredictable performance. In other words, participants made responses that were unexpected based on their ability level, which suggests that the item may not validly assess the latent variable language comprehension. Relatively lower proportions of items with infit/outfit problems were observed on the MSEL and CDI, whereas higher proportions were observed on the VABS (II), SICD-R, and CALC. The low proportion of misfitting items on the MSEL may relate to the fact that this measure was developed using Rasch analysis, and poorly fitting items were eliminated before its publication. This process was not applied in the development of the CALC or SICD-R, though it was for the VABS (II). The mediocre performance of the CALC in the item fit analyses was noteworthy given its unique design as a highly flexible measure that was expressly created for assessing young children with disabilities. The infit/outfit analyses do not indicate that flexible

procedures had any advantage over standardized procedures in measuring language comprehension in our sample.

The pattern of infit/outfit results for the CDI was particularly interesting. Overall the CDI had a low proportion of misfitting items, but misfitting CDI items tended to show extreme misfit. This makes sense when considering the specific vocabulary involved. For example, one might expect that children would have different degrees of exposure to the words "brother" and "sister" based on their family compositions, a factor that is presumably unrelated to language comprehension. This observation highlights larger questions about the role of environmental exposure in language comprehension measurement. Variability in exposure to specific vocabulary stems from a variety of sources, including sources that seem idiosyncratic or irrelevant to language comprehension (e.g. child is lactose intolerant, so food words referring to dairy are unfamiliar), and sources that suggest impoverished living conditions (e.g. child has very limited access to books and toys, so many words are unfamiliar). Studies suggest that the latter has a far-reaching negative impact on language development (Hart & Risley, 2003).

Additional items that showed extremely poor fit using infit/outfit statistics include listening skills on the VABS (II) and compliance with inhibitory commands on the SICD-R and VABS (II). Parent report of a toddler's listening skills may be better explained by other factors, such as attention span. Although attention is distinct from language comprehension, it is possible that the two constructs are related because attention supports growth in language comprehension via a child's ability to engage in the linguistic environment for longer periods of time (Peyre et al., 2016). This may explain how measures like the VABS (II) justify integrating listening skills items. Nonetheless, our findings provide support for listening skills being distinct from language comprehension in toddlers with significant developmental delays.

Similarly, a child's response to inhibitory commands likely depends on his or her past experience with them, including their frequency, the tone with which they were delivered, and associated consequences (e.g. parent removing objects). Consistent with this, parenting experts frequently advise using "no" sparingly because otherwise young children may become desensitized to it, causing it to become ineffective (Ricker, 1998). Thus, it is possible that compliance with inhibitory commands is a not a valid indicator of language comprehension because a child's response is dependent on the parenting practices that occur in his or her home.

4.3.1.3 Unidimensionality.

In contrast to our hypotheses, our analyses indicated possible multidimensionality in the non-CDI data. One interpretation of this is that parent report and direct testing formats comprise two separate dimensions within language comprehension. This finding is less surprising when considered in the context of the strengths and weaknesses of each format. Specifically, parent report is particularly helpful in assessing infrequent, emerging behaviors, but may be impacted by limitations in parent recall, response biases, or variability in item interpretation. Direct testing is more standardized in terms of both assessment procedures and scoring, but results are vulnerable to interference from temporary fluctuations in child mood and behavior. Thus, it is possible that parent report and direct testing formats each contribute distinct and valuable information to language comprehension assessment.

Our analyses also indicated possible multidimensionality in the CDI data, with one dimension being dominated by animal names and sounds. This suggests that multiple traits may influence performance on CDI items, with one trait being uniquely related to knowledge of animal names and sounds. This trait could be child exposure to or interest in animal names and sounds. In general, multidimensionality is undesirable for the purpose of modeling language

comprehension using IRT methods. This is especially true in cases where multidimensionality seems to be related to specific content knowledge, rather than theory-based language dimensions, such as vocabulary and grammar (Tomblin & Zhang, 2006). One possible remedy might be reducing the number of animal names and sound items, and retaining only those with the best psychometric properties and lowest residual correlations with one another.

4.3.1.4 Local Independence

Our local independence analyses identified many examples of item pairs with high residual correlations, which was not anticipated in our hypotheses. Within the non-CDI measures, most of these item pairs involved multiple items exploring the same content. For example, the SICD-R contains six items on color identification, which all displayed high residual correlations. As Yew (1993) explained, this type of local dependence is to be expected, and does not necessarily indicate broader measurement problems. However, it may suggest that the measure has redundant items, and items could be eliminated from the measure with little or no harm to precision. For example, it is possible that testing the child's knowledge of only two or three colors rather than six may provide almost identical information. Although this change may sound trivial, any efforts to minimize testing times may be helpful in an evaluation context.

The CDI displayed a more concerning pattern of local dependence, in that items pairs tended to be diverse in content but located in similar places on the form, often near the beginning. Given the very large number of items on the CDI, this should be interpreted with caution due to possibility of type 1 error resulting in spurious high residual correlations. At the same time, these high residual correlations may suggest that the location of an item on the inventory directly influences parent responses. This could be because parents get into a behavioral set or habit of consistently responding with "yes" or "no" to adjacent items. One

possible remedy might be “scrambling” the CDI items into a random order and examining whether such a change produces consistent item parameters to those observed in the present study.

4.3.2 Research Aim 2

Determine the appropriateness of the measure for the sample.

4.3.2.1 Multiple ICCs, item-person maps, and information

In support of our hypotheses, the results of our analyses using ICCs, item-person maps, and TIFs generally supported the appropriateness of the measures for our sample, but also highlighted several areas of potential improvement. On a basic level, within the non-CDI analyses, the difficulty levels of items appeared to adequately cover the ability range in our sample. On a more detailed level, the TIF suggested that increased precision of measurement might be achieved by adding items of moderately low difficulty, at approximately -3 logits.

Additionally, there were many more items of both very high and very low difficulty than were necessary. This is to be expected, given the age ranges for which the various measures were designed; from infancy to adulthood. In many cases, items of very high and very low difficulty were not administered to all participants, due to being below basals or above ceilings. Therefore they were not as detrimental to testing efficiency as one might suspect. Nonetheless, there are alternative measure designs that allow test administrators to more efficiently assess a latent variable by using an item set that was carefully calibrated to achieve fine distinctions within a specific the ability range. Such measures rely on item parameters to hone in on latent variable levels without the need for basals and ceilings. The Differential Ability Scales-Second Edition (DAS-II; Elliott, 2007) is one example of such a measure.

The CDI analyses showed evidence of a dearth of items at both the low and high ends of the ability range of our sample. The fact that this problem appeared on the high ability range is unsurprising, given that the CDI Words and Gestures was intended for children from 8 to 18 months, and the mean age of children in our sample was 30 months. Thus, the CDI may have had limited measurement precision for children in our sample with relatively mild developmental delays. The dearth of items in the low ability range is more difficult to explain. It is possible that some of the children in our sample truly did have ability levels that were too low to be appropriately assessed by the CDI. Alternatively, some form of response bias may have distorted the CDI data. For example, it is possible that a small minority of parents reported very low CDI Words Understood either due to social desirability or in an effort to ensure that his or her child would qualify for intervention.

Measurement precision for children with medium ability levels was excellent in the CDI analyses. While this is generally encouraging, precision should also be evaluated in the context of the resources that were spent in order to for it to be achieved. For the CDI, these resources include the time and effort that parents must put forth in order to complete a very lengthy questionnaire (427 items). The results of our item fit and ICC analyses indicate that many of these items could be removed with little or no harm to measurement precision because they are either poorly fitting or redundant.

4.3.2.2 Comparison to the PPVT-III (4)

Also consistent with our hypotheses, many of the children in our sample were not able to attain valid basal scores on the PPVT-III (4) at the 12-month follow-up assessment. This suggests that the PPVT-III (4) is limited in the information it can provide with regard to language comprehension for many preschool-age children with significant developmental delays. The

reason for this may relate to a variety of issues, such as the nature of the stimuli (two-dimensional illustrations only), the response format (pointing or verbally stating a number), or floor effects resulting from item difficulty. Regardless of the cause, the fact that this widely-used and well-established measure of language comprehension was so problematic for our participants highlights the importance of carefully considering the congruence between child and measure characteristics when selecting a clinical assessment battery.

4.3.3 Research Aim 3

Examine the development of comprehension over time.

4.3.3.1 Item Difficulty at Baseline

Our data on the relative difficulty of items at baseline was broadly consistent with extant knowledge about language comprehension development in typically developing children. Consistent with our hypotheses, items that tested responsiveness to noise showed the lowest difficulty. Items that tested comprehension of complex material, such as early numeracy skills, showed the highest difficulty. It was also interesting to note that difficulty levels tended to be identical or similar for analogous items across measures. This consistency indicates that child performance on these items were relatively reliable, which is promising.

The results of rank ordering MSEL items by difficulty from our sample indicated that three items were more difficult than anticipated. These included 1) response to name, 2) following both of the instructions "give the block to [parent]" and "give the car to me," and 3) following either of the instructions "stand up and get the ball" or "get the box and bring it to me". Despite our efforts to reduce the risk of type 1 error, it is possible that some of these differences arose due to chance. However, it is also possible that the increased difficulty of these items could be related to unique characteristics of our sample. For example, inconsistent or

absent response to name is regarded as an early sign of autism in infants and toddlers (Nadig et al., 2007). Although our study did not specifically investigate or assess autism-related symptoms, parents of 24 children reported that their child had been diagnosed with an ASD at follow-up appointments, indicating that a sizable portion of our sample likely met criteria for ASD at follow-up appointments (Romski et al., 2009). Thus it is possible that this item exhibited systematically different psychometric properties in our sample compared to the normative sample.

4.3.3.2 Change in ability estimates.

Ability estimates derived from the SICD-R and CALC increased from each time point to the next. This was consistent with our hypotheses, and it would be expected in any sample of young children followed from toddlerhood to preschool. The largest increase was observed from baseline to post-intervention. These increases are likely attributable to both natural maturation and the effects of intervention, though disentangling the effects of each is beyond the scope of this study.

4.3.3.3 Probability of change for items.

Our examination of the probability of baseline to post-intervention change for individual items revealed many items on which a substantial portion of our sample improved. In support of our hypotheses, items with the highest rates of improvement tended to be parent report rather than direct testing items. This was true both when comparing the CDI to direct testing items from other measures and also in specific cases where analogous parent report and directing testing items exploring the same content were examined. The observation of higher rates of change in parent report items could be attributable to the fact that parents are more likely to observe behaviors that are currently emerging, and thus occur infrequently, in young children. Another

possible explanation could be response bias in the form a desire to report improvement following the intervention program.

Additionally, we observed a high rate of improvement in turn-taking across two different measures. In typically developing children, it is common for turn-taking skills to improve during the preschool years, as play and conversation skills become more sophisticated (Hoff, 2013). In the context of our intervention, we highlighted turn-taking as a technique to encourage child communication. This was accomplished by teaching parents to pause and wait for children to respond before continuing the interaction. The effectiveness of this technique was supported by statistically significant increases in both child and parent turn-taking according to transcriptions of the baseline and final interventions sessions (Romski et al., 2010). Additionally, many children were encouraged to use target vocabulary related to turn taking (e.g. “my turn” and “your turn”) via either speech or SGD.

Finally, we observed high rates of improvement on vocabulary semantically related to school (school, teacher, and bus). This served as more evidence that specific child experiences and related exposure to vocabulary influence the development of language comprehension at the individual word level. It was common for children to begin attending preschool programs between study follow-up visits.

4.4 Clinical Implications

As noted in the introduction, language comprehension is critical to a wide variety of child outcomes, including academic success and emotional and social well-being. Effective intervention for language and communication disorders relies on valid, reliable language comprehension data to determine the intensity and techniques that are appropriate for an individual child. This study has the potential to both provide recommendations regarding the use

of currently available measures and to illuminate ways in which the field can improve in the development of future language comprehension measures.

Altogether, our impressions about the strengths and weaknesses of the measures support the importance of a multi-method approach in clinical assessment in order to gain a comprehensive and accurate understanding of a child's language comprehension profile, despite the shortcomings of each measure. Among the currently available measures, our analyses highlighted particularly high rates of misfitting items on the VABS (II) and CALC, calling into question the usefulness of these measures compared to others with more sound psychometric properties. However, there are also arguments that each of these measures is nonetheless an important tool in language comprehension assessment for young children with developmental disabilities. With regard to the VABS (II), adaptive behavior is a crucial construct in the field of developmental disabilities, both from a diagnosis and an intervention perspective (APA, 2013). With regard to the CALC, the authors' attempt to use flexibility to overcome measurement challenges in assessment of children with developmental disabilities may prove helpful in a subset of children with presentations even more complex than the majority of children in our sample. The SICD-R generally seemed to be a weaker measure, due to both its high rate of item misfit and its small, homogenous, and dated normative sample (252 Caucasian children, published in 1984). The MSEL and CDI displayed relatively stronger psychometric properties, although, as noted in more detail above, there are many opportunities for improving the CDI by eliminating misfitting and redundant items.

Our findings also highlight several ways in which IRT could be applied to improve measures of language comprehension that will be developed or updated in the future. First, in each of our analyses, we identified areas of the ability continuum in which our measures lacked

precision. This problem is likely more pronounced in assessment of atypical populations, such as the children in our sample, and highlights the important of calculating and examining the distribution of item difficulty levels at the time of test development. Additionally, our observation of redundant and misfitting items suggests possible improvements in testing efficiency by eliminating items. Shortening the length of evaluations may lessen the burden on families and improve access to care by reducing waitlists and evaluation costs. Combined, the goals of both improving measurement precision and testing efficiency highlight the merits of measure designs that involve well-calibrated item sets targeted to a child's hypothesized developmental level.

4.5 Limitations

Limitations of the present study include the sample size, which is relatively small among studies using IRT analyses. This may negatively impact the precision of our parameter estimates. Additionally, the fact that some participants did not attend follow-up appointments further reduced the quantity and completeness of our data with regard to language comprehension growth over time. Third, the distribution of maternal education suggests that our sample was predominantly of middle to high socioeconomic status (SES), and thus, our findings may not be generalizable to children of low SES. Finally, our assessment battery included two measures that were updated between the first and second studies (VABS and PPVT). This limited our ability to apply IRT analyses for these measures because the item-level content differed across the two versions.

4.6 Future Directions

One logical next step for our analyses would be to create a single language comprehension scale, or "super measure", by combining items with the best psychometric

properties from the six measures we examined. This would involve paring down items by removing those that showed either misfit or redundancy. We could further experiment in designing measures to maximize measurement precision within varying time limits.

Another follow-up study might explore child and parent characteristics that affect the relationship between item response and the latent variable language comprehension using DIF analyses. For example, it would be interesting to examine the extent to which child motor impairments impact responses to language comprehension items. Additionally, DIF might allow us to examine possible parent response bias by, for example, examining whether parent psychological distress impacts responses to parent report language comprehension items. Unfortunately, our present sample size is too small to support such analyses, and so this would only be possible in future studies.

Finally, a deeper investigation of the effects of intervention on language comprehension would also be extremely informative. For example, it is possible that the four intervention types in our study differed in the extent to which they promoted language comprehension growth. Romski and Sevcik (1993) hypothesize that interventions involving adult input using SGDs may be particularly beneficial to early comprehension development. Additionally, a few language comprehension items, especially on the CDI, overlap with individualized vocabulary targets from the intervention for some children. It would helpful to explore change over time in targeted items verses non-targeted items. This could distinguish between content that was taught to children and possible generalized effects of intervention. Finally, from a measurement perspective, it would interesting to examine the possibility that parent behavior as observers of child language comprehension changed as a result of participation in the intervention. The intervention itself created many opportunities for parents to both observe child language

comprehension and intervene to promote it in a carefully controlled, low-distraction environment. It is possible that this experience altered parent frames of reference when responding to questions about child language comprehension.

4.7 Conclusion

In conclusion, our investigation into the psychometric properties of language comprehension measures in a sample of toddlers with significant developmental delays revealed both strengths and weaknesses of the extant measures. We found that the aggregate data adequately fit the Rasch model, though each measure also contained individual items with poor fit. Items that displayed extremely poor fit included inhibitory commands, listening skills, and vocabulary such as early phrases and names of people. Our data suggest that these constructs may be unrelated to the latent variable language comprehension. Overall, the MSEL and CDI showed the strongest psychometric properties when examining individual item fit.

Dimensionality analyses provided some evidence of multidimensionality, which may be related to item format (parent report vs. direct testing). Both item fit and dimensionality analyses highlighted the influence of child exposure to specific material on language comprehension assessment. Local independence analyses indicated that many item pairs were correlated beyond the extent that would be expected based on the latent variable. In many cases, this was explained by the fact that item content was similar, though the CDI also indicated possible effects of item order.

Analyses related to the correspondence between item difficulties and participant abilities generally supported the appropriateness of the measures for our sample, and indicated acceptable measurement precision for the majority of participants. At the same time, these analyses also revealed several areas where minor improvements are possible. For example, adding a few items

of moderately low difficulty would improve measurement precision for participants of moderately low language comprehension abilities. Additionally, there was clear evidence of redundancy, or cases in which many items displayed similar difficulty. Such items could be eliminated in order to improve testing efficiency with little or no harm to measurement precision. Finally, the PPVT III (4) was extremely limited in its ability to measure language comprehension in our sample due to the fact that many participants did not attain basal scores.

Examination of the relative difficulty of items indicated patterns that were largely consistent with the literature on typically developing children. The few exceptions to this included child response to his or her name, which displayed unexpectedly high difficulty. Participant Rasch ability estimates consistently increased from each assessment to the next, indicating growth in language comprehension ability over time. Investigation of individual items showing the highest proportions of change in our sample indicated that parent-report items of moderate difficulty were most likely to reflect language comprehension improvement. Additionally, many participants improved on turn-taking and school-related vocabulary.

To our knowledge, this study is the first investigation of language comprehension measures in a sample of toddlers with significant developmental delays using IRT methods. This study has the potential to inform clinical practice, measure development, and knowledge about language comprehension development in atypical populations. With regard to clinical practice, our findings both identify measures with stronger psychometric properties and underscore measure limitations. With regard to measure development, our findings emphasize the benefits of integrating IRT methods in order to maximize both measurement precision and testing efficiency. With regard to knowledge about language comprehension development, our findings

provide information regarding the consistency of specific content difficulty between typically developing children and children with significant developmental delays.

REFERENCES

Abbeduto, L., Murphy, M. M., Cawthon, S. W., Richmond, E. K., Weissman, M. D., Karadottir, S., & O'Brien, A. (2003). Receptive language skills of adolescents and young adults with Down or fragile X syndrome. *American Journal on Mental Retardation, 108*(3), 149–160. https://doi.org/10.1352/0895-8017(2003)108<0149:RLSOAA>2.0.CO;2

Akshoomoff, N. (2006). Use of the Mullen Scales of Early Learning for the Assessment of Young Children with Autism Spectrum Disorders. *Child Neuropsychology : A Journal on Normal and Abnormal Development in Childhood and Adolescence, 12*(4–5), 269–277. https://doi.org/10.1080/09297040500473714

Anderson, J. D., Hess, R., & Richardson, K. (1980). Test–retest reliability of the test for auditory comprehension of language when it is used with mentally retarded children. *Journal of Speech & Hearing Disorders*, *45*(2), 195–199. https://doi.org/10.1044/jshd.4502.195

Barre, N., Morgan, A., Doyle, L. W., & Anderson, P. J. (2011). Language Abilities in Children Who Were Very Preterm and/or Very Low Birth Weight: A Meta-Analysis. *The Journal of Pediatrics*, *158*(5), 766–774.e1. https://doi.org/10.1016/j.jpeds.2010.10.032

Bond, T., Fox, C. M., & Bond, T. G. (2007). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences, Second Edition* (2 edition). Mahwah, N.J: Routledge.

Carrow-Woolfolk, E. (1995). *OWLS (oral and written language scales) manual: Listening comprehension and oral expression*. American Guidance Service.

Carrow-Woolfolk, E. (1999). *Comprehensive Assessment of Spoken Language (CASL)*. Torrence, CA: Western Psychological Services.

Carrow-Woolfolk, E. (2013). *Test for Auditory Comprehension of Language-Fourth Edition* (4th ed.). Austin, TX: Pro ed.

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

Davis, J. M. (1977). Reliability of hearing-impaired children's responses to oral and total presentations of the Test of Auditory Comprehension of Language. *Journal of Speech & Hearing Disorders*, *42*(4), 520–527. https://doi.org/10.1044/jshd.4204.520

de Ayala, R. J. (2008). *The Theory and Practice of Item Response Theory* (1 edition). New York: The Guilford Press.

Dunn, L. M., & Dunn, D. M. (2007). Peabody Picture Vocabulary Test--Fourth Edition. *PsycTESTS*. https://doi.org/10.1037/t15144-000

Dunn, L. M., & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test-Third Edition (PPVT-III)*. San Antonio, TX: Pearson.

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists* (1 edition). Mahwah, N.J: Psychology Press.

ETS. (n.d.). The GRE Tests. Retrieved February 12, 2017, from https://www.ets.org/gre

Fennell, E. B., & Dikel, T. N. (2001). Cognitive and neuropsychological functioning in children with cerebral palsy. *Journal of Child Neurology*, *16*(1), 58–63.

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

Field, A. (2013). *Discovering Statistics using IBM SPSS Statistics* (Fourth Edition edition). Los Angeles: SAGE Publications Ltd.

Geytenbeek, J. J. M., Heim, M. M. J., Vermeulen, R. J., & Oostrom, K. J. (2010). Assessing comprehension of spoken language in nonspeaking children with cerebral palsy: Application of a newly developed computer-based instrument. *Augmentative and Alternative Communication*, *26*(2), 97–107. https://doi.org/10.3109/07434618.2010.482445

Hambleton, R. K. (1991). *Fundamentals of Item Response Theory* (1 edition). Newbury Park, Calif: SAGE Publications, Inc.

Hansen, B. D., Wadsworth, J. P., Roberts, M. R., & Poole, T. N. (2014). Effects of naturalistic instruction on phonological awareness skills of children with intellectual and developmental disabilities. *Research in Developmental Disabilities*, *35*(11), 2790–2801. https://doi.org/10.1016/j.ridd.2014.07.011

Hedrick, D. L., Prather, E. M., & Tobin, A. (1984). *Sequenced Inventory of Communication Development-Revised (SICD-R)*. Western Psychological Services. Retrieved from Seattle, WA

Hobson, J. A., Hobson, R. P., Malik, S., Bargiota, K., & Calo, S. (2013). The relation between social engagement and pretend play in autism. *British Journal of Developmental Psychology*, *31*(1), 114–127.

Linacre, J. M. (2012). *Winsteps Rasch measurement computer program user's guide*. Beaverton, Oregon: Winsteps.com.

Linacre, J. M. (2016). *Winsteps Rasch Measurement Computer Progrm*. Beaverton, Oregon: Winsteps.com.

Liogier d'Ardhuy, X., Edgin, J. O., Bouis, C., de Sola, S., Goeldner, C., Kishnani, P., … Khwaja, O. (2015). Assessment of cognitive scales to examine memory, executive function and language in individuals with Down syndrome: Implications of a 6-month observational study. *Frontiers In Behavioral Neuroscience, 9*, 300–300. https://doi.org/10.3389/fnbeh.2015.00300

Lord, F. M. (1980). *Applications of Item Response Theory To Practical Testing Problems* (1 edition). Hillsdale, N.J: Lawrence Erlbaum Associates.

Luu, T. M., Vohr, B. R., Schneider, K. C., Katz, K. H., Tucker, R., Allan, W. C., & Ment, L. R. (2009). Trajectories of receptive language development from 3 to 12 years of age for very preterm children. *Pediatrics*, *124*(1), 333–341. https://doi.org/10.1542/peds.2008-2587

Luyster, R. J. S., AnneTalbott, Meagan R.Hele.Tager-Flusberg. (2011). Identifying Early-Risk Markers and Developmental Trajectories for Language Impairment in Neurodevelopmental Disorders. *Developmental Disabilities Research Reviews*, *17*(2), 151–159. https://doi.org/10.1002/ddrr.1109

Lyytinen, P., KennethLyytinen, Heikki. (2005). Language Development and Literacy Skills in Late-talking Toddlers with and without Familial Risk for Dyslexia. *Annals of Dyslexia*, *55*(2), 166–192.

Mei, C., Reilly, S., Reddihough, D., Mensah, F., Pennington, L., & Morgan, A. (2016). Language outcomes of children with cerebral palsy aged 5 years and 6 years: a population-based study. *Developmental Medicine & Child Neurology*, *58*(6), 605–611. https://doi.org/10.1111/dmcn.12957

Miller, J. F., & Paul, R. (1995). *The Clinical Assessment of Language Comprehension*. Baltimore: Paul H Brookes Pub Co.

Mullen, E. M. (1995). *Mullen Scales of Early Learning manual* (AGS ed). American Guidance Service.

Næss, K.-A. B., Lyster, S.-A. H., Hulme, C., & Melby-Lervåg, M. (2011). Language and verbal short-term memory skills in children with Down syndrome: A meta-analytic review. *Research in Developmental Disabilities*, *32*(6), 2225–2234. https://doi.org/10.1016/j.ridd.2011.05.014

Nasir, M. (2014, October 7). *Application of Classical Test Theory and Item Response Theory to Analyze Multiple Choice Questions* (Thesis). University of Calgary. Retrieved from http://theses.ucalgary.ca/jspui/handle/11023/1917

Paul, R. (2000). Predicting outcomes of early expressive language delay: Ethical implications. In D. V. M. Bishop, L. B. Leonard, D. V. M. (Ed) Bishop, & L. B. (Ed) Leonard (Eds.), *Speech and language impairments in children: Causes, characteristics, intervention and outcome.* (pp. 195–209). New York, NY, US: Psychology Press.

Pecyna, P. M., & Sommers, R. K. (1985). Testing the Receptive Language Skills of Severely Handicapped Preschool Children. *Language, Speech, and Hearing Services in Schools*, *16*(1), 41–52.

Pirila, S., van der Meere, J., Pentikainen, T., Ruusu-Niemi, P., Korpela, R., Kilpinen, J., & Nieminen, P. (2007). Language and motor speech skills in children with cerebral palsy. *Journal of Communication Disorders*, *40*(2), 116–128. https://doi.org/10.1016/j.jcomdis.2006.06.002

Reynell, J. (1990). *Reynell developmental language scales: Manual*. WPS, Western Psychological Services.

Romski, M. A., & Sevcik, R. A. (1993). Language comprehension: Considerations for augmentative and alternative communication. *AAC: Augmentative and Alternative Communication*, *9*(4), 281–285. https://doi.org/10.1080/07434619312331276701

Romski, MA, Sevcik, R., & Adamson, L. (1997). Framework for studying how children with developmental disabilities develop language through augmented means. *AAC: Augmentative & Alternative Communication*, *13*(3), 172–178.

Romski, MaryAnn, Sevcik, R. A., Adamson, L. B., Cheslock, M., Smith, A., Barker, R. M., & Bakeman, R. (2010). Randomized comparison of augmented and nonaugmented language interventions for toddlers with developmental delays and their parents. *Journal of Speech, Language, and Hearing Research*, *53*(2), 350–364.

Rousset, A., Dowell, R., & Leigh, J. (2016). Receptive language as a predictor of cochlear implant outcome for prelingually deaf adults. *International Journal of Audiology*, *55*(Suppl 2), S24–S30. https://doi.org/10.3109/14992027.2016.1157269

Sabbadini, M., Bonanni, R., Carlesimo, G. A., & Caltagirone, C. (2001). Neuropsychological assessment of patients with severe neuromotor and verbal disabilities. *Journal of Intellectual Disability Research: JIDR*, *45*(Pt 2), 169–179.

Semel, E., Wiig, E., & Secord, W. A. (2003). *Clinical Evaluation of Language Fundamentals, Fourth Edition*. San Antonio, TX: Pearson.

Sevcik, R. A., & Romski, M. (2005). Early Visual-Graphic Symbol Acquisition by Children With Developmental Disabilities. In L. L. Namy & L. L. Namy (Ed) (Eds.), *Symbol use and symbolic representation: Developmental and comparative perspectives.* (pp. 155–170). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Sparrow, S., Ciccheti, D., & Balla, D. (1984). *Vineland Adaptive Behavior Scales: Survey form manual* (1st ed.). American Guidance Services.

Sparrow, S., Cicchetti, D., & Balla, D. (2005). *Vineland Adaptive Behavior Scales: Survey form manual* (2nd ed.). American Guidance Services.

Spearman, C. (1904). "General Intelligence," Objectively Determined and Measured. *The American Journal of Psychology*, *15*(2), 201–292. https://doi.org/10.2307/1412107

Spoken Language Disorders: Overview. (2015). Retrieved February 6, 2016, from http://www.asha.org/Practice-Portal/Clinical-Topics/Spoken-Language-Disorders/

Suh, J., Eigsti, I.-M., Canfield, A., Irvine, C., Kelley, E., Naigles, L. R., & Fein, D. (2017). Language representation and language use in children with optimal outcomes from ASD. In L. R. Naigles & L. R. Naigles (Ed) (Eds.), *Innovative investigations of language in autism spectrum disorder.* (pp. 225–243). Washington, DC, US; Berlin, Germany: American Psychological Association.

Thal, D., & Tobias, S. (1991). Language and gesture in late talkers: A 1-year follow up. *Journal of Speech & Hearing Research*, *34*(3), 604.

Venker, C. E., Eernisse, E. R., Saffran, J. R., & Ellis Weismer, S. (2013). Individual differences in the real-time comprehension of children with ASD. *Autism Research: Official Journal Of The International Society For Autism Research*, *6*(5), 417–432. https://doi.org/10.1002/aur.1304

Whitley, B. E. (2003). *Principles of Research in Behavioral Science with Internet Guide and PowerWeb* (2 edition). New York: McGraw-Hill Humanities/Social Sciences/Languages.

Yen, W. (1993). Scaling Performance Assessments: Strategies for Managing Local Item Dependence. *Journal of Educational Measurement*, *30*(3), 187–213. https://doi.org/10.1111/j.1745-3984.1993.tb00423.x

Yoder, P., Watson, L. R., & Lambert, W. (2015). Value-Added Predictors of Expressive and Receptive Language Growth in Initially Nonverbal Preschoolers with Autism Spectrum Disorders. *Journal of Autism and Developmental Disorders*, *45*(5), 1254–1270.

Zimmerman, I. L., Steiner, V. G., & Pond, R. E. (2011). *Preschool Language Scales, Fifth Edition* (5th ed.). San Antonio, TX: Pearson.

Zwaigenbaum, L., Bryson, S., Lord, C., Rogers, S., Carter, A., Carver, L., … Yirmiya, N. (2009). Clinical Assessment and Management of Toddlers With Suspected Autism Spectrum Disorder: Insights From Studies of High-Risk Infants. *Pediatrics*, *123*(5), 1383–1391. https://doi.org/10.1542/peds.2008-1606

APPENDICES

Appendix A Item Characteristics

Appendix A.1 Item Characteristics Part 1

		Linguistic Content									Administration Format					Response Format							
	Item Description	Phrases & Routines	Nouns	Verbs	Locations & Prepositions	Adjectives & Adverbs	Comparative Concepts	Auditory Recall	Grammar	Pragmatics	Early Academics	Parent report	Examiner administered	Parent administered	Non-Specific Movement	Specific Movement/Gesture	Social Response	Compliance w/ Command	Gaze	Point	Object Manipulation	Answers Questions
MSEL																							
1	Reacts to a loud noise												✓		✓								
2	Alerts to sound												✓		✓								
3	Responds to voice and face by smiling												✓	✓			✓		✓				
4	Coordinates listening and turning												✓		✓								
5	Responds to voice and face by vocalizing												✓	✓			✓						
6	Coordinates listening and looking												✓				✓						
7	Enjoys self/mirror interaction												✓	✓			✓						
8	Attends to words and movement	✓												✓		✓							
9	Recognizes familiar names, words	✓	✓	✓									✓	✓		✓	✓		✓	✓			
10	Recognizes own name	✓											✓				✓						
11	Understands inhibitory words	✓												✓		✓		✓					

12	Understands simple verbal input	✓											✓			✓		✓				
13	Understands gesture and command	✓								✓			✓			✓		✓				
14	Identifies objects		✓										✓			✓		✓				
15	Gives toy on verbal request	✓	✓										✓					✓				
16	Comprehends questions I		✓										✓			✓			✓	✓		
17	Follows directions		✓	✓					✓				✓			✓		✓				
18	Recognizes ≥1 body part		✓										✓							✓		
	Recognizes ≥4 body parts		✓										✓							✓		
	Recognizes ≥6 body parts		✓										✓							✓		
19	Comprehends questions II		✓										✓							✓		
20	Follows related commands		✓	✓					✓				✓					✓				
21	Identifies pictures		✓										✓							✓		
22	Auditory spatial awareness, ≥1 location		✓	✓	✓				✓				✓								✓	
	Auditory spatial awareness, ≥2 locations		✓	✓	✓				✓				✓								✓	
	Auditory spatial awareness, ≥3 locations		✓	✓	✓				✓				✓								✓	
	Auditory spatial awareness, ≥4 locations		✓	✓	✓				✓				✓								✓	
23	Comprehends ≥1 action word			✓									✓							✓		
	Comprehends ≥2 action words			✓									✓							✓		
24	Identifies object function			✓									✓							✓		
25	Follows 2 unrelated commands		✓	✓				✓	✓				✓					✓			✓		
26	Size concepts					✓							✓							✓		
27	Identifies colors					✓					✓		✓							✓		
28	Length concepts					✓							✓							✓		
29	Comparative concepts, ≥3 concepts						✓						✓							✓		
	Comparative concepts, ≥4 concepts						✓						✓							✓		
	Comparative concepts, ≥5 concepts						✓						✓							✓		
	Comparative concepts, ≥6 concepts						✓						✓							✓		
30	General knowledge questions, ≥6 correct		✓								✓		✓									✓
	General knowledge questions, ≥7 correct		✓								✓		✓									✓
	General knowledge questions, ≥8 correct		✓								✓		✓									✓
	General knowledge questions, ≥9 correct		✓								✓		✓									✓

	General knowledge questions, ≥10 correct		✓							✓		✓									✓
31	Follows 3 unrelated commands		✓	✓	✓			✓	✓				✓					✓			✓	
32	Has concept of numbers, ≥1 correct						✓				✓		✓					✓			✓	
	Has concept of numbers, 2 correct						✓				✓		✓					✓			✓	
33	Identifies ≥12 letters										✓		✓							✓		
	Identifies all letters										✓		✓							✓		
VABS (II)																						
1	Turns eyes and head toward sound											✓			✓							
2	Looks toward parent when hears voice											✓					✓					
3	Responds to name											✓					✓					
4	Understands no	✓										✓				✓		✓				
5	Understands yes	✓										✓				✓	✓					
6	Listens to a story ≥5 min	✓								✓		✓					✓	✓	✓			
7	Points to ≥3 body parts		✓									✓								✓		
8	Points to objects		✓									✓								✓		
9	Listens to instructions	✓								✓		✓					✓	✓	✓			
10	Follows 1-step instructions	✓	✓	✓				✓				✓						✓			✓	
11	Points to ≥5 body parts		✓									✓								✓		
12	Follows 2-step instructions	✓	✓	✓				✓				✓						✓			✓	
13	Follows if-then instructions	✓	✓	✓				✓	✓			✓						✓				
14	Listens to a story ≥15 min	✓								✓		✓					✓	✓	✓			
15	Listens to a story ≥30 min	✓								✓		✓					✓	✓	✓			
16	Follows 3-step instructions	✓	✓	✓				✓				✓						✓			✓	
17	Follows instructions from 5 min before	✓	✓	✓				✓				✓						✓			✓	
18	Understands figures of speech									✓		✓										
19	Listens to informational talk ≥15 min									✓		✓					✓	✓	✓			
20	Listens to informational talk ≥30 min									✓		✓					✓	✓	✓			
SICD-R																						
1	Responds to sounds [a]											✓			✓		✓					
2	Responds to name	✓											✓				✓		✓			
3	Turns to localize, 90 left side	✓											✓				✓		✓			

	Turns to localize, 90 right side	✓											✓				✓		✓			
4	Turns to localize, 135 left side												✓		✓				✓			
	Turns to localize, 135 right side												✓		✓				✓			
	Turns to localize, 135 left side												✓		✓				✓			
	Turns to localize, 135 right side												✓		✓				✓			
5	Responds to name said by examiner	✓											✓				✓		✓			
	Responds to name said by parent	✓												✓			✓		✓			
6	Responds to come here with movement [a]											✓				✓						
	Responds to come her with movement												✓			✓						
7	Responds to noises around the house [a]											✓			✓		✓					
8.	Understands ≥2 words for toys [a]		✓									✓										
	Understands ≥2 names of family members [a]		✓									✓										
	Understands ≥2 words for clothing [a]		✓									✓										
	Understands ≥2 verbs [a]			✓								✓										
	Understands ≥2 names of acquaintances [a]		✓									✓										
	Understands ≥2 words for outdoor items [a]		✓									✓										
	Understands ≥2 adjectives [a]					✓						✓										
	Understands ≥2 words for household tools [a]											✓										
	Understands ≥2 pronouns [a]											✓										
	Understands ≥2 words for places [a]		✓									✓										
	Understands ≥2 words for games [a]		✓									✓										
9	Response to intonation	✓											✓			✓		✓				
10	Response to command with gesture	✓								✓			✓			✓		✓				
11	Response to specific word		✓										✓			✓				✓		
	Response to specific word [a]	✓										✓										
12	Response to situational commands, get object		✓										✓			✓						
	Response to situational commands, put down		✓		✓								✓			✓					✓	
	Response to situational commands, give to me	✓											✓			✓		✓				
13	Body part comprehension, ears		✓										✓			✓				✓		
	Body part comprehension, eyes		✓										✓			✓				✓		
	Body part comprehension, hair		✓										✓			✓				✓		

	Body part comprehension, mouth		✓									✓			✓				✓		
	Body part comprehension, nose		✓									✓			✓				✓		
	Body part comprehension, ears [a]		✓								✓				✓				✓		
	Body part comprehension, eyes [a]		✓								✓				✓				✓		
	Body part comprehension, hair [a]		✓								✓				✓				✓		
	Body part comprehension, mouth [a]		✓								✓				✓				✓		
	Body part comprehension, nose [a]		✓								✓				✓				✓		
14	Response to stand up and sit down			✓								✓					✓				
	Response to stand up and sit down [a]			✓							✓						✓				
15	Speech discrimination, socks		✓									✓			✓				✓		
	Speech discrimination, tree		✓									✓			✓				✓		
	Speech discrimination, bear		✓									✓			✓				✓		
	Speech discrimination, chair		✓									✓			✓				✓		
	Speech discrimination, key		✓									✓			✓				✓		
	Speech discrimination, box		✓									✓			✓				✓		
16	Responds to name of familiar person	✓	✓									✓						✓	✓		✓
17	Responds to prepositional commands, on		✓		✓							✓								✓	
	Responds to prepositional commands, in		✓		✓							✓								✓	
	Responds to prepositional commands, beside		✓		✓							✓								✓	
	Responds to prepositional commands, under		✓		✓							✓								✓	
18	Responds to bye-bye	✓										✓			✓						✓
19	Responds to number concepts, one						✓					✓								✓	
	Responds to number concepts, all						✓					✓								✓	
20	Responds to commands, walk to parent		✓	✓		✓						✓			✓		✓				
	Responds to commands, walk fast		✓	✓		✓						✓			✓		✓				
	Responds to commands, walk slowly		✓	✓		✓						✓			✓		✓				
21	Understands object function, stove		✓	✓								✓			✓				✓		
	Understands object function, shoe		✓	✓								✓			✓				✓		
	Understands object function, book		✓	✓								✓			✓				✓		
22	Takes turns [a]									✓		✓			✓		✓				
23	Identification of big and little		✓			✓						✓			✓				✓		

24	Discrimination of noises, rattle							✓					✓			✓				✓		✓
	Discrimination of noises, bell							✓					✓			✓				✓		✓
	Discrimination of noises, cellophane							✓					✓			✓				✓		✓
25	Identification of colors, orange					✓					✓		✓			✓				✓		
	Identification of colors, purple					✓					✓		✓			✓				✓		
	Identification of colors, red					✓					✓		✓			✓				✓		
	Identification of colors, yellow					✓					✓		✓			✓				✓		
	Identification of colors, green					✓					✓		✓			✓				✓		
	Identification of colors, blue					✓					✓		✓			✓				✓		
26	Responds to two object commands, ex. 1		✓	✓				✓					✓			✓						
	Responds to two object commands, ex. 2		✓	✓				✓					✓			✓						
	Responds to two object commands, ex. 3		✓	✓				✓					✓			✓						
27	Responds to two action commands, ex. 1		✓	✓									✓			✓					✓	
	Responds to two action commands, ex. 2		✓	✓									✓			✓					✓	
	Responds to two action commands, ex. 3		✓	✓									✓			✓					✓	
28	Understands plurals, ex. 1		✓	✓	✓				✓				✓			✓					✓	
	Understands plurals, ex. 2		✓	✓	✓		✓		✓				✓			✓					✓	
29	Sound discrimination, high												✓			✓				✓		
	Sound discrimination, low												✓			✓				✓		
	Sound discrimination, medium												✓			✓				✓		
30	Identification of hard and soft					✓							✓			✓				✓		
31	Identification of rough and smooth					✓							✓			✓				✓		
32	Identification of coins, penny		✓								✓		✓			✓				✓		
	Identification of coins, dime		✓								✓		✓			✓				✓		
	Identification of coins, nickel		✓								✓		✓			✓				✓		
33	Response to 3-step commands, ex. 1		✓	✓	✓			✓					✓								✓	
	Response to 3-step commands, ex. 2		✓	✓	✓			✓					✓								✓	
	Response to 3-step commands, ex. 3		✓	✓	✓			✓					✓								✓	
34	Understanding of numbers, ex. 1		✓								✓		✓			✓						
	Understanding of numbers, ex. 2		✓								✓		✓			✓						
	Understanding of numbers, ex. 3		✓								✓		✓			✓						

CALC																					
2.1	Familiar routines	✓								✓				✓		✓						
2.2	Joint reference activity									✓				✓		✓			✓			
2.3	Object and person names		✓										✓			✓			✓			
2.4	Action words		✓	✓									✓			✓						
2.5	Absent persons and objects		✓										✓			✓			✓			
2.6	Two-word relations		✓						✓	✓			✓			✓						
2.7	Turn-taking									✓				✓			✓					✓
3.3	Two-to-three word instructions		✓	✓					✓				✓			✓					✓	
3.7	Word order		✓	✓					✓				✓							✓		

Note. [a] Specifies an SICD item that is parent report rather than direct testing. Non-specific movement refers to movements in response to sound that suggest intact hearing but not necessarily comprehension (e.g. an infant turning toward the sound of a rattle). Social responses include smiling, laughing, or making eye-contact, particularly in response to voices or faces.

Appendix A.2 Item Characteristics Part 2

		Materials					Cues		Flexibility			Field Size & Scoring	
	Item Description	Actual Objects	Miniatures	2D Drawings in Color	2D Drawings w/ out Color	Visual	Social	Allows Repetition	Individualized Vocab	Individualized Materials	Field Size	Number of Correct Trials Required	Total Number of Trials
MSEL													
1	Reacts to a loud noise											1	1
2	Alerts to sound											1	1
3	Responds to voice and face by smiling											1	1
4	Coordinates listening and turning							✓				1	1
5	Responds to voice and face by vocalizing											1	1
6	Coordinates listening and looking											1	1
7	Enjoys self/mirror interaction	✓										1	1
8	Attends to words and movement						✓	✓				1	2
9	Recognizes familiar names, words							✓	✓			1	
10	Recognizes own name											1	1
11	Understands inhibitory words											1	1
12	Understands simple verbal input								✓			1	1
13	Understands gesture and command						✓	✓		✓		1	1
14	Identifies objects	✓	✓			✓		✓			2	1	1
15	Gives toy on verbal request									✓		1	1
16	Comprehends questions I	✓										1	2
17	Follows directions	✓	✓								2	2	3

18	Recognizes ≥1 body part				✓							1	7
	Recognizes ≥4 body parts				✓							4	7
	Recognizes ≥6 body parts				✓							6	7
19	Comprehends questions II				✓						3	1	3
20	Follows related commands	✓				✓	✓					1	2
21	Identifies pictures				✓						4	2	4
22	Auditory spatial awareness, ≥1 location	✓				✓	✓					1	5
	Auditory spatial awareness, ≥2 locations		✓									2	5
	Auditory spatial awareness, ≥3 locations		✓									3	5
	Auditory spatial awareness, ≥4 locations		✓									4	5
23	Comprehends ≥1 action word				✓						3	1	3
	Comprehends ≥2 action words				✓						3	2	3
24	Identifies object function	✓	✓								4	3	4
25	Follows 2 unrelated commands	✓	✓			✓	✓					1	2
26	Size concepts				✓						2	3	4
27	Identifies colors	✓									8	4	8
28	Length concepts	✓									2	3	4
29	Comparative concepts, ≥3 concepts				✓						3	3	10
	Comparative concepts, ≥4 concepts				✓						3	4	10
	Comparative concepts, ≥5 concepts				✓						5	5	10
	Comparative concepts, ≥6 concepts				✓						5	6	10
30	General knowledge questions, ≥6 correct											6	12
	General knowledge questions, ≥7 correct											7	12
	General knowledge questions, ≥8 correct											8	12
	General knowledge questions, ≥9 correct											9	12
	General knowledge questions, ≥10 correct											10	12
31	Follows 3 unrelated commands	✓										1	1
32	Has concept of numbers, ≥1 correct	✓									10	1	2
	Has concept of numbers, 2 correct	✓										2	2

33	Identifies ≥12 letters				✓							12	14
	Identifies all letters				✓							14	14
VABS (II)													
1	Turns eyes and head toward sound								✓				
2	Looks toward parent when hears voice								✓				
3	Responds to name												
4	Understands no												
5	Understands yes												
6	Listens to a story ≥5 min								✓	✓			
7	Points to ≥3 body parts								✓			3	
8	Points to objects								✓	✓			
9	Listens to instructions								✓				
10	Follows 1-step instructions								✓				
11	Points to ≥5 body parts								✓			5	
12	Follows 2-step instructions								✓				
13	Follows if-then instructions								✓				
14	Listens to a story ≥15 min								✓	✓			
15	Listens to a story ≥30 min								✓	✓			
16	Follows 3-step instructions								✓				
17	Follows instructions from 5 min before								✓				
18	Understands figures of speech								✓				
19	Listens to informational talk ≥15 min								✓				
20	Listens to informational talk ≥30 min								✓				
SICD-R													
1	Responds to sounds [a]											2	
2	Responds to name							✓				1	2
3	Turns to localize, 90 left side							✓				1	1
	Turns to localize, 90 right side							▪				1	1
4	Turns to localize, 135 left side							✓				1	1

	Turns to localize, 135 right side							✓				1	1
	Turns to localize, 135 left side							✓				1	1
	Turns to localize, 135 right side							✓				1	1
5	Responds to name said by examiner							✓				1	1
	Responds to name said by parent							✓				1	1
6	Responds to come here with movement [a]						✓						
	Responds to come her with movement						✓	✓				1	2
7	Responds to noises around the house [a]											2	
8.	Understands ≥2 words for toys [a]								✓			2	
	Understands ≥2 names of family members [a]								✓			2	
	Understands ≥2 words for clothing [a]								✓			2	
	Understands ≥2 verbs [a]								✓			2	
	Understands ≥2 names of acquaintances [a]								✓			2	
	Understands ≥2 words for outdoor items [a]								✓			2	
	Understands ≥2 adjectives [a]								✓			2	
	Understands ≥2 words for household tools [a]								✓			2	
	Understands ≥2 pronouns [a]								✓			2	
	Understands ≥2 words for places [a]								✓			2	
	Understands ≥2 words for games [a]								✓			2	
9	Response to intonation	✓	✓					✓				1	1
10	Response to command with gesture	✓	✓				✓	✓				1	1
11	Response to specific word	✓	✓					✓			3	1	1
	Response to specific word [a]											1	
12	Response to situational commands, get object	✓	✓			✓		✓			3	1	1
	Response to situational commands, put down	✓	✓					✓				1	1
	Response to situational commands, give to me	✓	✓					✓				1	1
13	Body part comprehension, ears							✓				1	1
	Body part comprehension, eyes							✓				1	1
	Body part comprehension, hair							✓				1	1

	Body part comprehension, mouth							✓				1	1
	Body part comprehension, nose							✓				1	1
	Body part comprehension, ears [a]											1	1
	Body part comprehension, eyes [a]											1	1
	Body part comprehension, hair [a]											1	1
	Body part comprehension, mouth [a]											1	1
	Body part comprehension, nose [a]											1	1
14	Response to stand up and sit down							✓				1	1
	Response to stand up and sit down [a]											1	1
15	Speech discrimination, socks	✓	✓					✓			6	1	1
	Speech discrimination, tree	✓	✓					✓			6	1	1
	Speech discrimination, bear	✓	✓					✓			6	1	1
	Speech discrimination, chair	✓	✓					✓			6	1	1
	Speech discrimination, key	✓	✓					✓			6	1	1
	Speech discrimination, box	✓	✓					✓			6	1	1
16	Responds to name of familiar person							✓				1	1
17	Responds to prepositional commands, on	✓						✓				1	1
	Responds to prepositional commands, in	✓						✓				1	1
	Responds to prepositional commands, beside	✓						✓				1	1
	Responds to prepositional commands, under	✓						✓				1	1
18	Responds to bye-bye							✓				1	1
19	Responds to number concepts, one	✓						✓				1	1
	Responds to number concepts, all	✓						✓				1	1
20	Responds to commands, walk to parent							✓				1	1
	Responds to commands, walk fast							✓				1	1
	Responds to commands, walk slowly							✓				1	1
21	Understands object function, stove				✓			✓			4	1	1
	Understands object function, shoe				✓			✓			4	1	1
	Understands object function, book				✓			✓			4	1	1

22	Takes turns [a]											1	1
23	Identification of big and little	✓									2	3	3
24	Discrimination of noises, rattle	✓									3	1	1
	Discrimination of noises, bell	✓									3	1	1
	Discrimination of noises, cellophane	✓									3	1	1
25	Identification of colors, orange	✓									6	1	1
	Identification of colors, purple	✓									6	1	1
	Identification of colors, red	✓									6	1	1
	Identification of colors, yellow	✓									6	1	1
	Identification of colors, green	✓									6	1	1
	Identification of colors, blue	✓									6	1	1
26	Responds to two object commands, ex. 1	✓	✓								5	1	3
	Responds to two object commands, ex. 2	✓	✓								5	2	3
	Responds to two object commands, ex. 3	✓	✓								5	3	3
27	Responds to two action commands, ex. 1	✓	✓								5	1	3
	Responds to two action commands, ex. 2	✓	✓								5	2	3
	Responds to two action commands, ex. 3	✓	✓								5	3	3
28	Understands plurals, ex. 1	✓	✓									1	1
	Understands plurals, ex. 2	✓	✓									1	1
29	Sound discrimination, high	✓									3	1	1
	Sound discrimination, low	✓									3	1	1
	Sound discrimination, medium	✓									3	1	1
30	Identification of hard and soft	✓						✓			2	3	3
31	Identification of rough and smooth	✓						✓			2	3	3
32	Identification of coins, penny	✓						✓			4	1	1
	Identification of coins, dime	✓						✓			4	1	1
	Identification of coins, nickel	✓						✓			4	1	1
33	Response to 3-step commands, ex. 1	✓	✓									1	3
	Response to 3-step commands, ex. 2	✓	✓									2	3

	Response to 3-step commands, ex. 3	✓	✓									3	3
34	Understanding of numbers, ex. 1	✓						✓				1	1
	Understanding of numbers, ex. 2	✓						✓				1	1
	Understanding of numbers, ex. 3	✓						✓				1	1
CALC													
2.1	Familiar routines							✓	✓	✓		3	4
2.2	Joint reference activity						✓	✓	✓	✓		3	4
2.3	Object and person names							✓	✓	✓		3	4
2.4	Action words							✓	✓	✓		3	4
2.5	Absent persons and objects							✓	✓	✓		3	4
2.6	Two-word relations							✓	✓	✓		3	4
2.7	Turn-taking								✓	✓			
3.3	Two-to-three word instructions							✓	✓	✓		3	4
3.7	Word order				✓						4	3	4

Note. [a] Specifies an SICD item that is parent report rather than direct testing. Visual cues refers to placing correct objects to select closer to the child or requesting perceptually intuitive responses, such as placing a small object in a container. Social cues refers to hints provided by the examiner in the form of gaze, gesture, or demonstration of a correct response.

Appendix B Item-Level Parameters and Fit

Appendix B1 Item-Level Data for MSEL

Item	Description	Item Parameters		Item Fit			Baseline Correct (Prop.)
		Item Difficulty	Difficulty S.E.	Infit MNSQ	Outfit MNSQ	Monotonicity	
1.	Reacts to a loud noise	-9.43	1.84	–	–	–	113/113 (1)
2.	Alerts to sound	-9.43	1.84	–	–	–	113/113 (1)
3.	Responds to voice and face (smiling)	-9.43	1.84	–	–	–	113/113 (1)
4.	Coordinates listening and turning	-9.43	1.84	–	–	–	113/113 (1)
5.	Responds to voice and face (vocalizing)	-7.4	0.76	0.8	0.12	No	111/113 (0.98)
6.	Coordinates listening and looking	-7.4	0.76	0.78	0.11	No	111/113 (0.98)
7.	Enjoys self/mirror interaction	-7.4	0.76	0.8	0.12	No	111/113 (0.98)
8.	Attends to words and movement	-8.17	1.04	**0.73**	**0.04**	**Yes**	112/113 (0.99)
9.	Recognizes familiar names, words	-6.28	0.51	**0.87**	**0.28**	**Yes**	108/113 (0.96)
10.	Recognizes own name	-4.88	0.36	1	0.49	Yes	100/113 (0.88)
11.	Understands inhibitory words	-6.03	0.48	0.97	0.42	No	107/113 (0.95)
12.	Understands simple verbal input	-4.75	0.35	**0.75**	**0.44**	**Yes**	99/113 (0.88)
13.	Understands gesture and command	-4.31	0.32	0.88	0.43	No	95/113 (0.84)
14.	Identifies objects	-2.82	0.28	**0.74**	**0.41**	**Yes**	78/113 (0.69)
15.	Gives toy on verbal request	-0.38	0.29	**0.66**	**0.37**	**Yes**	47/113 (0.42)
16.	Comprehends questions I	-1.5	0.28	**0.78**	**0.57**	**Yes**	61/113 (0.54)
17.	Follows directions	0.2	0.29	**0.8**	**0.45**	**Yes**	40/113 (0.35)
18.	Recognizes ≥1 body part	-0.94	0.28	**0.61**	**0.39**	**Yes**	54/113 (0.48)
	Recognizes ≥4 body parts	0.11	0.29	**0.59**	**0.34**	**Yes**	41/113 (0.36)
	Recognizes ≥6 body parts	0.81	0.3	**0.61**	**0.3**	**Yes**	33/113 (0.29)
19.	Comprehends questions II	-0.63	0.28	**0.63**	**0.34**	**Yes**	50/113 (0.44)
20.	Follows related commands	0.63	0.3	**0.82**	**0.45**	**Yes**	35/113 (0.31)
21.	Identifies pictures	0.03	0.29	**0.67**	**0.36**	**Yes**	42/113 (0.37)

22.	Auditory spatial awareness, ≥1 location	0.9	0.3	**0.86**	**0.45**	**Yes**	32/113 (0.28)
	Auditory spatial awareness, ≥2 locations	3.35	0.41	0.98	0.58	No	11/113 (0.1)
	Auditory spatial awareness, ≥3 locations	5.16	0.66	0.69	0.09	No	4/113 (0.04)
	Auditory spatial awareness, ≥4 locations	5.16	0.66	0.69	0.09	No	4/113 (0.04)
23.	Comprehends ≥1 action word	1.08	0.31	**0.78**	**0.48**	**Yes**	30/113 (0.27)
	Comprehends ≥2 action words	2	0.34	1.07	0.81	No	21/110 (0.19)
24.	Identifies object function	2.74	0.37	**0.8**	**0.31**	**Yes**	15/113 (0.13)
25.	Follows 2 unrelated commands	4.18	0.5	1.19	0.44	No	7/113 (0.06)
26.	Size concepts	5.16	0.66	0.98	0.33	No	4/113 (0.04)
27.	Identifies colors	5.16	0.66	1.29	0.38	No	4/113 (0.04)
28.	Length concepts	7.92	1.46	0.19	0.01	Yes	1/113 (0.01)
29.	Comparative concepts, ≥3 concepts	7.92	1.46	0.19	0.01	Yes	1/113 (0.01)
	Comparative concepts, ≥4 concepts	9.91	2.06	–	–	–	0/113 (0)
	Comparative concepts, ≥5 concepts	9.91	2.06	–	–	–	0/113 (0)
	Comparative concepts, ≥6 concepts	9.91	2.06	–	–	–	0/113 (0)
30.	General knowledge questions, ≥6 correct	9.91	2.06	–	–	–	0/113 (0)
	General knowledge questions, ≥7 correct	9.91	2.06	–	–	–	0/113 (0)
	General knowledge questions, ≥8 correct	9.91	2.06	–	–	–	0/113 (0)
	General knowledge questions, ≥9 correct	9.91	2.06	–	–	–	0/113 (0)
	General knowledge questions, ≥10 correct	9.91	2.06	–	–	–	0/113 (0)
31.	Follows 3 unrelated commands	9.91	2.06	–	–	–	0/113 (0)
32.	Has concept of numbers, ≥1 correct	9.91	2.06	–	–	–	0/113 (0)
	Has concept of numbers, 2 correct	9.91	2.06	–	–	–	0/113 (0)

Note. Bolded item fit statistics indicate that the item fit the IRT model adequately. Bolded difference in proportion correct indicates that > 20% of the sample improved on the item from baseline to post-intervention.

Appendix B2 Item-Level Data for VABS & VABS II

Item Description		Item Parameters		Item Fit			
		Item Difficulty	Difficulty S.E.	Infit MNSQ	Outfit MNSQ	Monotonicity	Baseline Correct (Prop.)
1.	Turns eyes and head toward sound, partial [a]	-9.43	1.84	–	–	–	113/113 (1)
	Turns eyes and head toward sound, full credit [a]	-6.92	0.63	1.4	9.9	No	110/113 (0.97)
2.	Looks toward parent when hears voice, partial	-9.08	1.86	–	–	–	51/51 (1)
	Looks toward parent when hears voice, full credit	-5.56	0.58	1.49	2.3	No	46/51 (0.9)
3.	Responds to name, partial	-6.91	0.81	1.21	1.1	No	49/51 (0.96)
	Responds to name, full credit	-3.56	0.45	1.53	7.58	No	38/51 (0.75)
4.	Understands no, partial [a]	-8.17	1.04	1.16	9.9	No	112/113 (0.99)
	Understands no, full credit [a]	-5.15	0.38	1.28	2.56	No	102/113 (0.9)
5.	Understands yes, partial [a]	-3.68	0.3	**1.23**	**1.29**	**Yes**	86/110 (0.78)
	Understands yes, full credit [a]	-2.42	0.28	**1.06**	**1.45**	**Yes**	71/110 (0.65)
6.	Listens to a story ≥5 min, partial [a]	-2.04	0.28	2.06	3.39	No	68/113 (0.6)
	Listens to a story ≥5 min, full credit [a]	-0.3	0.29	2.17	2.84	No	46/113 (0.41)
7.	Points to ≥3 body parts, partial	-1.68	0.43	**1.26**	**0.89**	**Yes**	28/51 (0.55)
	Points to ≥3 body parts, full credit	-0.94	0.43	**0.81**	**0.6**	**Yes**	24/51 (0.47)
8.	Points to objects, partial	-1.86	0.43	1.33	1.27	No	29/51 (0.57)
	Points to objects, full credit	-0.02	0.43	**1.1**	**0.89**	**Yes**	19/51 (0.37)
9.	Listens to instructions, partial [a]	-3.39	0.29	**0.98**	**1.72**	**Yes**	85/113 (0.75)
	Listens to instructions, full credit [a]	-0.54	0.28	**1.36**	**1.26**	**Yes**	49/113 (0.43)
10.	Follows 1-step instructions, partial [a]	-3.23	0.29	**1.02**	**1.6**	**Yes**	83/113 (0.73)
	Follows 1-step instructions, full credit [a]	-1.26	0.28	1.51	1.37	No	58/113 (0.51)
11.	Points to ≥5 body parts, partial	1.97	0.48	**0.83**	**0.38**	**Yes**	9/51 (0.18)
	Points to ≥5 body parts, full credit	1.97	0.48	**0.83**	**0.38**	**Yes**	9/51 (0.18)
12.	Follows 2-step instructions, partial	1.97	0.48	1.05	0.82	No	9/51 (0.18)
	Follows 2-step instructions, full credit	3.8	0.68	1.13	2.54	No	3/51 (0.06)
13.	Follows if-then instructions, partial [a]	4.77	0.59	1.93	1.26	No	5/113 (0.04)
	Follows if-then instructions, full credit [a]	7.92	1.46	3.28	7.36	No	1/113 (0.01)
14.	Listens to a story ≥15 min, partial	1.97	0.48	0.75	0.35	No	9/51 (0.18)

	Listens to a story ≥15 min, full credit	2.73	0.54	**0.57**	**0.19**	**Yes**	6/51 (0.12)
15.	Listens to a story ≥30 min, partial	3.31	0.62	0.87	0.23	No	4/50 (0.08)
	Listens to a story ≥30 min, full credit	6.46	1.86	–	–	–	0/50 (0)
16.	Follows 3-step instructions, partial	6.5	1.86	–	–	–	0/51 (0)
	Follows 3-step instructions, full credit	6.5	1.86	–	–	–	0/51 (0)
17.	Follows instructions from 5 min before, partial	6.5	1.86	–	–	–	0/51 (0)
	Follows instructions from 5 min before, full credit	6.5	1.86	–	–	–	0/51 (0)
18.	Understands figures of speech, partial	6.5	1.86	–	–	–	0/51 (0)
	Understands figures of speech, full credit	6.5	1.86	–	–	–	0/51 (0)
19.	Listens to informational talk ≥15 min, partial	6.5	1.86	–	–	–	0/51 (0)
	Listens to informational talk ≥15 min, full credit	6.5	1.86	–	–	–	0/51 (0)
20.	Listens to informational talk ≥30 min, partial	6.5	1.86	–	–	–	0/51 (0)
	Listens to informational talk ≥30 min, full credit	6.5	1.86	–	–	–	0/51 (0)

Note. Bolded item fit statistics indicate that the item fit the IRT model adequately. Bolded difference in proportion correct indicates that > 20% of the sample improved on the item from baseline to post-intervention. [a] Item is on both VABS and VABS II, all other items are only on VABS II.

Appendix B3 Item-Level Data for SICD-R

Item	Description	Item Parameters		Item Fit			Change Over Time		
		Item Difficulty	Difficulty S.E.	Infit MNSQ	Outfit MNSQ	Mono-tonicity	Baseline Correct (Prop.)	Post Correct (Prop.)	Difference in Prop.
1.	Responds to sounds [a]	-9.43	1.84	–	–	–	113/113 (1)	111/112 (0.99)	-0.01
2.	Responds to name	-5.79	0.45	**0.81**	**0.34**	**Yes**	105/112 (0.94)	108/112 (0.96)	0.03
3.	Turns to localize, 90 left side	-6.9	0.64	**0.71**	**0.11**	**Yes**	109/112 (0.97)	109/112 (0.97)	0.00
	Turns to localize, 90 right side	-6.9	0.64	**0.71**	**0.11**	**Yes**	109/112 (0.97)	109/112 (0.97)	0.00
4.	Turns to localize, 135 left side	-8.17	1.04	0.93	0.08	No	112/113 (0.99)	111/112 (0.99)	0.00
	Turns to localize, 135 right side	-9.43	1.84	–	–	–	113/113 (1)	111/112 (0.99)	-0.01
	Turns to localize, 135 left side	-8.17	1.04	0.93	0.08	No	112/113 (0.99)	111/112 (0.99)	0.00
	Turns to localize, 135 right side	-9.43	1.84	–	–	–	113/113 (1)	111/112 (0.99)	-0.01
5.	Responds to name said by ex.	-5.15	0.38	**0.9**	**0.38**	**Yes**	102/113 (0.9)	107/112 (0.96)	0.05
	Responds to name said by parent	-6.57	0.56	1.14	0.53	No	109/113 (0.96)	109/112 (0.97)	0.01
6.	Responds to come with movement [a]	-8.17	1.04	1.13	0.61	No	112/113 (0.99)	108/112 (0.96)	-0.03
	Responds to come with movement	-5.63	0.43	1.5	1.03	No	105/113 (0.93)	105/112 (0.94)	0.01
7.	Responds to noises [a]	-5.82	0.45	**0.97**	**0.39**	**Yes**	106/113 (0.94)	108/112 (0.96)	0.03
8.	Understands ≥2 words for toys [a]	-4.64	0.34	1.04	0.63	No	98/113 (0.87)	111/112 (0.99)	0.12
	Understands ≥2 names of family members [a]	-5.15	0.38	0.92	0.36	Yes	102/113 (0.9)	104/112 (0.93)	0.03
	Understands ≥2 words for clothing [a]	-3.31	0.29	**1.08**	**1.17**	**Yes**	84/113 (0.74)	93/112 (0.83)	0.09
	Understands ≥2 verbs [a]	-4.31	0.32	**0.99**	**0.55**	**Yes**	95/113 (0.84)	104/112 (0.93)	0.09
	Understands ≥2 names of acquaintances [a]	-2.35	0.28	**1.19**	**1.39**	**Yes**	72/113 (0.64)	88/112 (0.79)	0.15
	Understands ≥2 words for outdoor items [a]	-2.43	0.28	**0.78**	**0.46**	**Yes**	73/113 (0.65)	88/112 (0.79)	0.14
	Understands ≥2 adjectives [a]	-1.73	0.28	**0.86**	**0.89**	**Yes**	64/113 (0.57)	83/112 (0.74)	0.17
	Understands ≥2 words for household tools [a]	-1.5	0.28	**0.87**	**0.81**	**Yes**	61/113 (0.54)	77/112 (0.69)	0.15
	Understands ≥2 pronouns [a]	-1.34	0.28	**0.75**	**0.68**	**Yes**	59/113 (0.52)	75/112 (0.67)	0.15
	Understands ≥2 words for places [a]	-1.5	0.28	**0.79**	**0.74**	**Yes**	61/113 (0.54)	76/112 (0.68)	0.14
	Understands ≥2 words for games [a]	-1.89	0.28	**0.69**	**0.41**	**Yes**	66/113 (0.58)	77/112 (0.69)	0.10
9.	Response to intonation	-4.52	0.33	1.53	6.78	No	97/113 (0.86)	102/112 (0.91)	0.05
10.	Response to command with gesture	-3.74	0.3	1.22	0.8	No	89/113 (0.79)	98/112 (0.88)	0.09

11.	Response to specific word	-1.89	0.28	**0.67**	**0.39**	**Yes**	66/113 (0.58)	85/112 (0.76)	0.17
	Response to specific word [a]	-4.31	0.32	**0.76**	**0.32**	**Yes**	95/113 (0.84)	103/112 (0.92)	0.08
12.	Response to situational commands, get object	-2.59	0.28	**0.97**	**1.03**	**Yes**	75/113 (0.66)	85/112 (0.76)	0.10
	Response to situational commands, put down	0.37	0.29	**0.9**	**0.58**	**Yes**	38/113 (0.34)	56/112 (0.5)	0.16
	Response to situational commands, give	-0.14	0.29	**1.12**	**1.17**	**Yes**	44/113 (0.39)	59/112 (0.53)	0.14
13.	Body part comprehension, ears	-0.22	0.29	**0.55**	**0.3**	**Yes**	45/113 (0.4)	58/112 (0.52)	0.12
	Body part comprehension, eyes	-0.38	0.29	**0.71**	**0.47**	**Yes**	47/113 (0.42)	62/112 (0.55)	0.14
	Body part comprehension, hair	0.03	0.29	**0.64**	**0.48**	**Yes**	42/113 (0.37)	60/112 (0.54)	0.16
	Body part comprehension, mouth	-0.63	0.28	**0.73**	**0.41**	**Yes**	50/113 (0.44)	60/112 (0.54)	0.09
	Body part comprehension, nose	-0.87	0.28	**0.72**	**0.56**	**Yes**	53/113 (0.47)	65/112 (0.58)	0.11
	Body part comprehension, ears [a]	-1.1	0.28	**1.02**	**0.7**	**Yes**	56/113 (0.5)	71/112 (0.63)	0.14
	Body part comprehension, eyes [a]	-1.42	0.28	**1.17**	**0.8**	**Yes**	60/113 (0.53)	75/112 (0.67)	0.14
	Body part comprehension, hair [a]	-0.71	0.28	**0.95**	**0.71**	**Yes**	51/113 (0.45)	68/112 (0.61)	0.16
	Body part comprehension, mouth [a]	-1.34	0.28	**0.96**	**0.63**	**Yes**	59/113 (0.52)	75/112 (0.67)	0.15
	Body part comprehension, nose [a]	-2.12	0.28	**1.06**	**0.68**	**Yes**	69/113 (0.61)	79/112 (0.71)	0.09
14.	Response to stand up and sit down	-0.71	0.28	**1.02**	**0.79**	**Yes**	51/113 (0.45)	66/112 (0.59)	0.14
	Response to stand up and sit down [a]	-1.73	0.28	**1.06**	**1.21**	**Yes**	64/113 (0.57)	82/112 (0.73)	0.17
15.	Speech discrimination, socks	-0.79	0.28	**0.92**	**0.6**	**Yes**	52/113 (0.46)	64/112 (0.57)	0.11
	Speech discrimination, tree	-0.05	0.29	**1.11**	**0.71**	**Yes**	43/113 (0.38)	52/112 (0.46)	0.08
	Speech discrimination, bear	-0.38	0.29	**0.95**	**0.68**	**Yes**	47/113 (0.42)	56/112 (0.5)	0.08
	Speech discrimination, chair	0.2	0.29	**0.88**	**0.48**	**Yes**	40/113 (0.35)	52/112 (0.46)	0.11
	Speech discrimination, key	0.2	0.29	**0.94**	**1.07**	**Yes**	40/113 (0.35)	56/112 (0.5)	0.15
	Speech discrimination, box	0.2	0.29	**0.84**	**0.51**	**Yes**	40/113 (0.35)	54/112 (0.48)	0.13
16.	Responds to name of familiar person	-1.26	0.28	**0.71**	**0.44**	**Yes**	58/113 (0.51)	67/112 (0.6)	0.08
17.	Responds to prepositional commands, on	2.74	0.37	1.38	1.42	No	15/113 (0.13)	28/111 (0.25)	0.12
	Responds to prepositional commands, in	-0.54	0.28	1.02	0.8	No	49/113 (0.43)	54/111 (0.49)	0.05
	Responds to prepositional commands, beside	6.44	0.99	0.51	0.03	Yes	2/113 (0.02)	14/111 (0.13)	0.11
	Responds to prepositional commands, under	3.94	0.47	1.05	0.81	No	8/113 (0.07)	25/111 (0.23)	0.15
18.	Responds to bye-bye	-0.46	0.28	1.05	1.3	No	47/112 (0.42)	66/112 (0.59)	0.17
19.	Responds to number concepts, one	4.53	0.59	0.96	0.57	No	5/111 (0.05)	26/111 (0.23)	0.19
	Responds to number concepts, all	1.39	0.32	1.43	1.82	No	25/111 (0.23)	42/111 (0.38)	0.15
20.	Responds to commands, walk	1.18	0.31	**1.06**	**0.64**	**Yes**	29/113 (0.26)	48/112 (0.43)	0.17

	Responds to commands, walk fast	3.73	0.45	0.9	0.31	No	9/113 (0.08)	22/112 (0.2)	0.12
	Responds to commands, walk slowly	3.94	0.47	1.14	0.4	No	8/113 (0.07)	22/111 (0.2)	0.13
21.	Understands object function, stove	2.88	0.38	0.92	0.36	No	14/113 (0.12)	25/111 (0.23)	0.10
	Understands object function, shoe	2.35	0.35	0.99	0.41	No	18/113 (0.16)	37/111 (0.33)	0.17
	Understands object function, book	2.74	0.37	1.13	0.53	No	15/113 (0.13)	34/111 (0.31)	0.17
22.	Takes turns [a]	2.61	0.36	**1.01**	**1.33**	**Yes**	16/113 (0.14)	39/112 (0.35)	**0.21**
23.	Identification of big and little	4.77	0.59	1.33	0.42	No	5/113 (0.04)	19/112 (0.17)	0.13
24.	Discrimination of noises, rattle	3.35	0.41	1.2	0.8	No	11/113 (0.1)	30/112 (0.27)	0.17
	Discrimination of noises, bell	3.03	0.39	0.96	0.35	No	13/113 (0.12)	29/112 (0.26)	0.14
	Discrimination of noises, cellophane	3.94	0.47	1.14	0.44	No	8/113 (0.07)	25/112 (0.22)	0.15
25.	Identification of colors, orange	3.94	0.47	0.89	0.31	No	8/113 (0.07)	25/112 (0.22)	0.15
	Identification of colors, purple	3.53	0.43	**0.96**	**0.3**	**Yes**	10/113 (0.09)	24/112 (0.21)	0.13
	Identification of colors, red	4.18	0.5	1.03	0.37	No	7/113 (0.06)	23/112 (0.21)	0.14
	Identification of colors, yellow	3.53	0.43	0.99	0.31	No	10/113 (0.09)	25/112 (0.22)	0.13
	Identification of colors, green	3.73	0.45	**0.9**	**0.29**	**Yes**	9/113 (0.08)	23/112 (0.21)	0.13
	Identification of colors, blue	3.53	0.43	**0.96**	**0.3**	**Yes**	10/113 (0.09)	21/112 (0.19)	0.10
26.	Responds to two object commands, ex. 1	3.73	0.45	**0.9**	**0.34**	**Yes**	9/113 (0.08)	18/112 (0.16)	0.08
	Responds to two object commands, ex. 2	3.94	0.47	0.73	0.19	No	8/113 (0.07)	18/112 (0.16)	0.09
	Responds to two object commands, ex. 3	3.94	0.47	**0.92**	**0.24**	**Yes**	8/113 (0.07)	21/112 (0.19)	0.12
27.	Responds to two action commands, ex. 1	7.92	1.46	0.19	0.01	Yes	1/113 (0.01)	4/112 (0.04)	0.03
	Responds to two action commands, ex. 2	6.44	0.99	0.51	0.03	Yes	2/113 (0.02)	5/112 (0.04)	0.03
	Responds to two action commands, ex. 3	7.92	1.46	0.19	0.01	Yes	1/113 (0.01)	4/112 (0.04)	0.03
28.	Understands plurals, example 1	7.92	1.46	0.19	0.01	Yes	1/113 (0.01)	3/112 (0.03)	0.02
	Understands plurals, example 2	7.92	1.46	0.19	0.01	Yes	1/113 (0.01)	1/112 (0.01)	0.00
29.	Sound discrimination, high	7.92	1.46	0.19	0.01	Yes	1/113 (0.01)	5/112 (0.04)	0.04
	Sound discrimination, low	7.92	1.46	3.17	0.35	No	1/113 (0.01)	5/112 (0.04)	0.04
	Sound discrimination, medium	9.91	2.06	–	–	–	0/113 (0)	3/112 (0.03)	0.03
30.	Identification of hard and soft	7.92	1.46	0.19	0.01	Yes	1/113 (0.01)	6/112 (0.05)	0.04
31.	Identification of rough and smooth	7.92	1.46	0.19	0.01	Yes	1/113 (0.01)	1/112 (0.01)	0.00
32.	Identification of coins, penny	9.91	2.06	–	–	–	0/113 (0)	2/112 (0.02)	0.02
	Identification of coins, dime	9.91	2.06	–	–	–	0/113 (0)	2/112 (0.02)	0.02
	Identification of coins, nickel	9.91	2.06	–	–	–	0/113 (0)	1/112 (0.01)	0.01

33.	Response to 3-step commands, ex. 1	9.91	2.06	–	–	–	0/113 (0)	1/112 (0.01)	0.01
	Response to 3-step commands, ex. 2	9.91	2.06	–	–	–	0/113 (0)	1/112 (0.01)	0.01
	Response to 3-step commands, ex. 3	9.91	2.06	–	–	–	0/113 (0)	1/112 (0.01)	0.01
34.	Understanding of numbers, ex. 1	7.92	1.46	0.19	0.01	Yes	1/113 (0.01)	1/112 (0.01)	0.00
	Understanding of numbers, ex. 2	7.92	1.46	0.19	0.01	Yes	1/113 (0.01)	1/112 (0.01)	0.00
	Understanding of numbers, ex. 3	9.91	2.06	–	–	–	0/113 (0)	1/112 (0.01)	0.01

Note. Bolded item fit statistics indicate that the item fit the IRT model adequately. Bolded difference in proportion correct indicates that > 20% of the sample improved on the item from baseline to post-intervention. [a] Item is parent report only, all other items are direct assessment

Appendix B4 Item-Level Data for CALC

Item	Description	Item Parameters: Item Difficulty	Item Parameters: Difficulty S.E.	Item Fit: Infit MNSQ	Item Fit: Outfit MNSQ	Item Fit: Mono-tonicity	Change Over Time: Baseline Correct (Prop.)	Change Over Time: Post Correct (Prop.)	Change Over Time: Difference in Prop.
2.1	Familiar routines, partial	-6.57	0.56	0.81	6.95	Yes	109/113 (0.96)	110/111 (0.99)	0.03
	Familiar routines, full credit	-5.01	0.37	1.07	9.9	No	101/113 (0.89)	107/111 (0.96)	0.07
2.2	Joint reference activity, partial	-6.92	0.63	1.02	9.9	No	110/113 (0.97)	109/111 (0.98)	0.01
	Joint reference activity, full credit	-4.21	0.32	1.19	6.42	No	94/113 (0.83)	96/111 (0.86)	0.03
2.3	Object/person names, partial	-3.83	0.3	**0.89**	**0.83**	**Yes**	90/113 (0.8)	98/111 (0.88)	0.09
	Object/person names, full credit	-1.26	0.28	1.04	1.61	No	58/113 (0.51)	72/111 (0.65)	0.14
2.4	Action words, partial	-1.65	0.28	**0.75**	**0.97**	**Yes**	63/113 (0.56)	80/111 (0.72)	0.16
	Action words, full credit	0.9	0.3	1.24	0.97	No	32/113 (0.28)	52/111 (0.47)	0.19
2.5	Absent persons/objects, partial	-1.1	0.28	**1.13**	**1.35**	**Yes**	56/113 (0.5)	75/111 (0.68)	0.18
	Absent persons/objects, full credit	1.47	0.32	1.35	1.46	No	26/113 (0.23)	49/111 (0.44)	**0.21**
2.6	Two-word relations, partial	-0.38	0.29	**0.89**	**0.63**	**Yes**	47/113 (0.42)	64/111 (0.58)	0.16
	Two-word relations, full credit	2.48	0.36	**0.92**	**0.39**	**Yes**	17/113 (0.15)	41/111 (0.37)	**0.22**
2.7	Turn-taking, partial	1.78	0.33	1.44	1.53	No	23/113 (0.2)	52/111 (0.47)	**0.26**
	Turn-taking, full credit	3.35	0.41	1.13	0.52	No	11/113 (0.1)	27/111 (0.24)	0.15
3.3	Multi-word instructions, partial	3.03	0.39	**0.85**	**0.34**	**Yes**	13/112 (0.12)	23/111 (0.21)	0.09
	Multi-word instructions, full credit	4.18	0.5	**0.68**	**0.16**	**Yes**	7/112 (0.06)	9/111 (0.08)	0.02
3.7	Word order, partial	6.44	0.99	2.74	2.73	No	2/112 (0.02)	5/110 (0.05)	0.03
	Word order, full credit	9.91	2.06	–	–	–	0/112 (0)	2/111 (0.02)	0.02

Note. Bolded item fit statistics indicate that the item fit the IRT model adequately. Bolded difference in proportion correct indicates that > 20% of the sample improved on the item from baseline to post-intervention.

Appendix B5 ***Item-Level Data for MacArthur CDI***

	Item Parameters		Item Fit			Change Over Time		
Item Description	Item Difficulty	Difficulty S.E.	Infit MNSQ	Outfit MNSQ	Mono-tonicity	Baseline Correct (Prop.)	Post Correct (Prop.)	Difference in Prop.
A1 First Signs of Understanding								
Responds to his or her name	-6.09	0.58	1.21	4.98	No	104/108 (0.96)	98/101 (0.97)	0.01
Responds to no	-6.09	0.58	1.63	2.02	No	104/108 (0.96)	99/101 (0.98)	0.02
Responds to mommy/daddy	-4.21	0.35	**0.86**	**0.76**	**Yes**	94/108 (0.87)	94/101 (0.93)	0.06
Early Phrases								
Are you hungry?	-1.79	0.24	**1.23**	**1.17**	**Yes**	66/113 (0.58)	77/102 (0.75)	0.17
Are you sleepy?	-0.85	0.24	**1.06**	**0.91**	**Yes**	50/113 (0.44)	69/102 (0.68)	**0.23**
Be careful	-0.43	0.25	1.46	1.62	No	43/113 (0.38)	59/102 (0.58)	0.20
Be quiet	-0.04	0.26	1.08	1.50	No	37/113 (0.33)	54/102 (0.53)	0.20
Clap your hands	-3.16	0.28	1.20	1.06	No	87/113 (0.77)	90/102 (0.88)	0.11
Change diaper	-2.15	0.25	1.04	9.90	Yes	72/113 (0.64)	83/102 (0.81)	0.18
Come here	-4.05	0.33	1.61	1.54	No	97/113 (0.86)	95/102 (0.93)	0.07
Mommy/Daddy's home	-3.01	0.27	**0.84**	**0.59**	**Yes**	85/113 (0.75)	86/102 (0.84)	0.09
Do you want more?	-3.48	0.29	**1.18**	**1.15**	**Yes**	91/113 (0.81)	93/102 (0.91)	0.11
Don't do that	-2.21	0.25	1.38	1.55	No	73/113 (0.65)	82/102 (0.8)	0.16
Don't touch	-1.91	0.24	1.26	1.38	No	68/113 (0.6)	76/102 (0.75)	0.14
Get up	-2.09	0.25	1.39	2.19	No	71/113 (0.63)	81/102 (0.79)	0.17
Give to mommy	-2.79	0.26	**0.98**	**0.78**	**Yes**	82/113 (0.73)	87/102 (0.85)	0.13
Give a hug	-2.73	0.26	1.19	5.27	No	81/113 (0.72)	86/102 (0.84)	0.13
Give a kiss	-3.48	0.29	1.51	9.90	No	91/113 (0.81)	94/102 (0.92)	0.12
Go get _	-1.97	0.24	**1.00**	**0.84**	**Yes**	69/113 (0.61)	72/102 (0.71)	0.10
Good boy/girl	-2.46	0.25	1.14	2.58	No	77/113 (0.68)	80/102 (0.78)	0.10
Hold still	0.87	0.29	1.35	1.15	No	25/113 (0.22)	40/102 (0.39)	0.17
Go bye-bye	-3.32	0.28	1.32	1.36	No	89/113 (0.79)	86/102 (0.84)	0.06
Look at this	-1.91	0.24	0.93	2.62	No	68/113 (0.6)	80/102 (0.78)	0.18

Open mouth	-0.85	0.24	1.15	7.02	No	50/113 (0.44)	71/102 (0.7)	**0.25**
Sit down	-3.01	0.27	1.49	2.04	No	85/113 (0.75)	88/102 (0.86)	0.11
Spit out	-0.24	0.25	1.49	5.41	No	40/113 (0.35)	52/102 (0.51)	0.16
Stop it	-2.59	0.26	**1.22**	**1.15**	**Yes**	79/113 (0.7)	80/102 (0.78)	0.09
Time to go night-night	-2.73	0.26	**1.29**	**1.31**	**Yes**	81/113 (0.72)	85/102 (0.83)	0.12
Throw ball	-2.94	0.27	**1.01**	**0.81**	**Yes**	84/113 (0.74)	85/102 (0.83)	0.09
This little piggy	0.46	0.28	1.31	2.98	No	30/113 (0.27)	46/102 (0.45)	0.19
Want to go for a ride?	-0.36	0.25	**1.16**	**1.15**	**Yes**	42/113 (0.37)	61/102 (0.6)	**0.23**
D1 Sound Effects								
baa	-1.09	0.24	1.32	1.50	No	54/113 (0.48)	63/102 (0.62)	0.14
choo	-1.09	0.24	**1.22**	**1.19**	**Yes**	54/113 (0.48)	64/102 (0.63)	0.15
cockadoodle	0.78	0.29	1.26	1.11	No	26/113 (0.23)	42/102 (0.41)	0.18
grr	-0.36	0.25	1.41	5.82	No	42/113 (0.37)	57/102 (0.56)	0.19
meow	-1.55	0.24	**1.03**	**1.08**	**Yes**	62/113 (0.55)	74/102 (0.73)	0.18
moo	-1.55	0.24	**0.87**	**0.65**	**Yes**	62/113 (0.55)	73/102 (0.72)	0.17
ouch	-0.67	0.25	**1.19**	**1.19**	**Yes**	47/113 (0.42)	59/102 (0.58)	0.16
quack	-1.09	0.24	**0.92**	**0.85**	**Yes**	54/113 (0.48)	66/102 (0.65)	0.17
uh-oh	-2.27	0.25	1.34	4.75	No	74/113 (0.65)	73/102 (0.72)	0.06
vroom	-1.21	0.24	1.12	1.98	No	56/113 (0.5)	60/102 (0.59)	0.09
woof	-1.38	0.24	1.20	2.31	No	59/113 (0.52)	71/102 (0.7)	0.17
yum	-0.30	0.25	1.28	1.60	No	41/113 (0.36)	64/102 (0.63)	**0.26**
D2 Animal Names								
animal	1.05	0.3	**0.95**	**0.69**	**Yes**	23/113 (0.2)	41/102 (0.4)	0.20
bear	-0.55	0.25	**0.81**	**0.60**	**Yes**	45/113 (0.4)	64/102 (0.63)	**0.23**
bee	1.05	0.3	**1.02**	**0.83**	**Yes**	23/113 (0.2)	36/102 (0.35)	0.15
bird	-0.91	0.24	**0.90**	**0.73**	**Yes**	51/113 (0.45)	73/102 (0.72)	**0.26**
bug	0.96	0.3	0.79	0.44	No	24/113 (0.21)	42/102 (0.41)	0.20
bunny	0.10	0.26	**0.88**	**0.67**	**Yes**	35/113 (0.31)	48/102 (0.47)	0.16
butterfly	0.62	0.28	**0.84**	**0.50**	**Yes**	28/113 (0.25)	44/102 (0.43)	0.18
cat	-1.55	0.24	**0.85**	**0.62**	**Yes**	62/113 (0.55)	75/102 (0.74)	0.19
chicken	0.31	0.27	**1.06**	**0.73**	**Yes**	32/113 (0.28)	52/102 (0.51)	**0.23**
cow	-0.91	0.24	**0.79**	**0.59**	**Yes**	51/113 (0.45)	71/102 (0.7)	**0.24**

deer	2.90	0.46	**0.71**	**0.39**	**Yes**	9/113 (0.08)	27/102 (0.26)	0.19
dog	-2.40	0.25	**0.84**	**0.65**	**Yes**	76/113 (0.67)	90/102 (0.88)	**0.21**
donkey	2.51	0.42	**0.73**	**0.30**	**Yes**	11/113 (0.1)	23/102 (0.23)	0.13
duck	-1.55	0.24	**0.78**	**0.57**	**Yes**	62/113 (0.55)	72/102 (0.71)	0.16
elephant	-0.10	0.26	**1.02**	**0.80**	**Yes**	38/113 (0.34)	59/102 (0.58)	**0.24**
fish	-0.97	0.24	**0.85**	**0.71**	**Yes**	52/113 (0.46)	71/102 (0.7)	**0.24**
frog	0.24	0.27	**0.93**	**0.83**	**Yes**	33/113 (0.29)	50/102 (0.49)	0.20
giraffe	1.34	0.32	1.07	0.72	No	20/113 (0.18)	42/102 (0.41)	**0.23**
goose	2.19	0.39	1.00	0.54	No	13/113 (0.12)	22/102 (0.22)	0.10
horse	-0.43	0.25	**0.94**	**0.72**	**Yes**	43/113 (0.38)	65/102 (0.64)	**0.26**
kitty	-0.04	0.26	**1.07**	**0.89**	**Yes**	37/113 (0.33)	51/102 (0.5)	0.17
lamb	1.67	0.34	**0.93**	**0.69**	**Yes**	17/113 (0.15)	31/102 (0.3)	0.15
lion	-0.04	0.26	**1.16**	**1.05**	**Yes**	37/113 (0.33)	53/102 (0.52)	0.19
monkey	0.10	0.26	**0.99**	**0.74**	**Yes**	35/113 (0.31)	60/102 (0.59)	**0.28**
mouse	1.34	0.32	**1.01**	**1.00**	**Yes**	20/113 (0.18)	38/102 (0.37)	0.20
owl	1.67	0.34	**0.68**	**0.33**	**Yes**	17/113 (0.15)	32/102 (0.31)	0.16
penguin	2.70	0.44	**0.69**	**0.21**	**Yes**	10/113 (0.09)	22/102 (0.22)	0.13
pig	-0.61	0.25	**0.86**	**0.64**	**Yes**	46/113 (0.41)	64/102 (0.63)	**0.22**
pony	2.35	0.4	0.85	0.55	No	12/113 (0.11)	22/102 (0.22)	0.11
puppy	0.03	0.26	0.96	0.68	No	36/113 (0.32)	50/102 (0.49)	0.17
sheep	0.54	0.28	**1.02**	**0.80**	**Yes**	29/113 (0.26)	47/102 (0.46)	0.20
squirrel	1.79	0.35	**0.83**	**0.41**	**Yes**	16/113 (0.14)	31/102 (0.3)	0.16
teddy	-0.04	0.26	**0.81**	**0.57**	**Yes**	37/113 (0.33)	56/102 (0.55)	**0.22**
tiger	0.70	0.29	**1.11**	**0.83**	**Yes**	27/113 (0.24)	51/102 (0.5)	**0.26**
turkey	2.90	0.46	**0.54**	**0.16**	**Yes**	9/113 (0.08)	25/102 (0.25)	0.17
turtle	1.05	0.3	**0.87**	**0.75**	**Yes**	23/113 (0.2)	45/102 (0.44)	**0.24**
D3 Vehicles								
airplane	-0.79	0.24	**0.83**	**0.68**	**Yes**	49/113 (0.43)	66/102 (0.65)	**0.21**
bike	0.03	0.26	**0.93**	**0.75**	**Yes**	36/113 (0.32)	54/102 (0.53)	**0.21**
bus	-0.79	0.24	**1.05**	**1.02**	**Yes**	49/113 (0.43)	71/102 (0.7)	**0.26**
car	-3.24	0.28	0.81	7.76	Yes	88/113 (0.78)	93/102 (0.91)	0.13
firetruck	0.03	0.26	0.98	1.10	No	36/113 (0.32)	52/102 (0.51)	0.19

motorcycle	1.55	0.33	1.12	1.04	No	18/113 (0.16)	41/102 (0.4)	**0.24**
stroller	-1.91	0.24	**1.11**	**0.87**	**Yes**	68/113 (0.6)	66/102 (0.65)	0.05
train	-1.03	0.24	**0.88**	**0.74**	**Yes**	53/113 (0.47)	72/102 (0.71)	**0.24**
truck	-0.91	0.24	**0.93**	**1.10**	**Yes**	51/113 (0.45)	68/102 (0.67)	**0.22**
D4 Toys								
ball	-5.35	0.46	1.11	0.47	No	106/113 (0.94)	98/102 (0.96)	0.02
balloon	-1.67	0.24	**0.80**	**0.61**	**Yes**	64/113 (0.57)	77/102 (0.75)	0.19
block	-1.73	0.24	**0.97**	**0.92**	**Yes**	65/113 (0.58)	71/102 (0.7)	0.12
book	-3.66	0.3	**0.83**	**0.58**	**Yes**	93/113 (0.82)	94/102 (0.92)	0.10
bubbles	-2.73	0.26	**0.93**	**0.66**	**Yes**	81/113 (0.72)	93/102 (0.91)	0.19
doll	-0.36	0.25	**1.00**	**1.02**	**Yes**	42/113 (0.37)	53/102 (0.52)	0.15
pen	0.54	0.28	**0.91**	**1.21**	**Yes**	29/113 (0.26)	41/102 (0.4)	0.15
toy	-0.91	0.24	**0.85**	**0.72**	**Yes**	51/113 (0.45)	60/102 (0.59)	0.14
D5 Food & Drink								
apple	-1.21	0.24	**0.81**	**0.68**	**Yes**	56/113 (0.5)	69/102 (0.68)	0.18
banana	-1.61	0.24	**1.06**	**0.91**	**Yes**	63/113 (0.56)	75/102 (0.74)	0.18
bread	-0.24	0.25	**0.94**	**0.69**	**Yes**	40/113 (0.35)	58/102 (0.57)	**0.21**
butter	2.05	0.37	0.90	0.47	No	14/113 (0.12)	24/102 (0.24)	0.11
cake	0.54	0.28	1.05	1.01	No	29/113 (0.26)	47/102 (0.46)	0.20
candy	0.17	0.27	1.30	1.46	No	34/113 (0.3)	49/102 (0.48)	0.18
carrot	0.78	0.29	**0.81**	**0.45**	**Yes**	26/113 (0.23)	42/102 (0.41)	0.18
cereal	-0.36	0.25	**1.00**	**0.94**	**Yes**	42/113 (0.37)	56/102 (0.55)	0.18
cheerios	-0.73	0.25	1.29	1.74	No	48/113 (0.42)	44/102 (0.43)	0.01
cheese	0.03	0.26	1.08	1.46	No	36/113 (0.32)	52/102 (0.51)	0.19
chicken	0.39	0.27	**1.05**	**0.76**	**Yes**	31/113 (0.27)	55/102 (0.54)	**0.26**
coffee	2.90	0.46	1.14	1.06	No	9/113 (0.08)	12/102 (0.12)	0.04
cookie	-2.09	0.25	1.31	1.61	No	71/113 (0.63)	79/102 (0.77)	0.15
cracker	-0.91	0.24	1.05	1.56	No	51/113 (0.45)	64/102 (0.63)	0.18
drink	-1.73	0.24	**1.20**	**1.08**	**Yes**	65/113 (0.58)	81/102 (0.79)	**0.22**
egg	0.96	0.3	**0.86**	**0.53**	**Yes**	24/113 (0.21)	39/102 (0.38)	0.17
fish	1.05	0.3	**1.09**	**0.74**	**Yes**	23/113 (0.2)	47/102 (0.46)	**0.26**
food	-0.36	0.25	**1.05**	**1.20**	**Yes**	42/113 (0.37)	54/102 (0.53)	0.16

ice cream	-0.30	0.25	**0.98**	**0.80**	**Yes**	41/113 (0.36)	56/102 (0.55)	0.19
juice	-1.85	0.24	**1.18**	**1.05**	**Yes**	67/113 (0.59)	77/102 (0.75)	0.16
meat	2.05	0.37	1.02	1.41	No	14/113 (0.12)	29/102 (0.28)	0.16
milk	-1.73	0.24	**1.26**	**1.18**	**Yes**	65/113 (0.58)	82/102 (0.8)	**0.23**
noodles	1.14	0.31	0.84	0.72	No	22/113 (0.19)	28/102 (0.27)	0.08
orange	0.54	0.28	**1.13**	**1.18**	**Yes**	29/113 (0.26)	45/102 (0.44)	0.18
peas	1.44	0.33	**0.92**	**0.50**	**Yes**	19/113 (0.17)	27/102 (0.26)	0.10
pizza	-0.17	0.26	**0.86**	**0.65**	**Yes**	39/113 (0.35)	54/102 (0.53)	0.18
raisin	1.67	0.34	1.23	0.94	No	17/113 (0.15)	37/102 (0.36)	**0.21**
spaghetti	0.78	0.29	0.90	0.64	No	26/113 (0.23)	37/102 (0.36)	0.13
toast	1.55	0.33	1.09	1.18	No	18/113 (0.16)	22/102 (0.22)	0.06
water	-1.50	0.24	**0.99**	**0.92**	**Yes**	61/113 (0.54)	72/102 (0.71)	0.17
D6 Clothing								
beads	2.05	0.37	**0.71**	**0.28**	**Yes**	14/113 (0.12)	21/102 (0.21)	0.08
bib	-0.10	0.26	**0.98**	**0.77**	**Yes**	38/113 (0.34)	41/102 (0.4)	0.07
boots	1.44	0.33	**0.77**	**0.37**	**Yes**	19/113 (0.17)	33/102 (0.32)	0.16
button	0.78	0.29	**0.73**	**0.42**	**Yes**	26/113 (0.23)	40/102 (0.39)	0.16
coat	-0.04	0.26	**0.78**	**0.59**	**Yes**	37/113 (0.33)	51/102 (0.5)	0.17
diaper	-2.40	0.25	**0.89**	**0.65**	**Yes**	76/113 (0.67)	82/102 (0.8)	0.13
dress	1.14	0.31	0.93	0.70	No	22/113 (0.19)	36/102 (0.35)	0.16
hat	-1.67	0.24	**0.76**	**0.54**	**Yes**	64/113 (0.57)	69/102 (0.68)	0.11
jacket	-0.24	0.25	**0.87**	**0.77**	**Yes**	40/113 (0.35)	54/102 (0.53)	0.18
jeans	1.91	0.36	**0.66**	**0.34**	**Yes**	15/113 (0.13)	23/102 (0.23)	0.09
necklace	1.91	0.36	**0.70**	**0.34**	**Yes**	15/113 (0.13)	27/102 (0.26)	0.13
pajamas	-0.10	0.26	**0.72**	**0.50**	**Yes**	38/113 (0.34)	56/102 (0.55)	**0.21**
pants	-1.85	0.24	**0.72**	**0.52**	**Yes**	67/113 (0.59)	74/102 (0.73)	0.13
shirt	-1.61	0.24	**0.64**	**0.47**	**Yes**	63/113 (0.56)	74/102 (0.73)	0.17
shoe	-3.32	0.28	**0.68**	**0.40**	**Yes**	89/113 (0.79)	92/102 (0.9)	0.11
shorts	0.54	0.28	**0.90**	**0.52**	**Yes**	29/113 (0.26)	40/102 (0.39)	0.14
socks	-2.15	0.25	**0.85**	**0.61**	**Yes**	72/113 (0.64)	83/102 (0.81)	0.18
sweater	1.67	0.34	**0.74**	**0.38**	**Yes**	17/113 (0.15)	27/102 (0.26)	0.11
zipper	0.46	0.28	**0.90**	**0.61**	**Yes**	30/113 (0.27)	46/102 (0.45)	0.19

D7 Body Parts								
arm	-0.24	0.25	**0.67**	**0.56**	**Yes**	40/113 (0.35)	57/102 (0.56)	0.20
belly button	0.24	0.27	**0.79**	**0.49**	**Yes**	33/113 (0.29)	46/102 (0.45)	0.16
cheek	0.96	0.3	**0.85**	**0.59**	**Yes**	24/113 (0.21)	39/102 (0.38)	0.17
ear	-1.91	0.24	**0.91**	**0.68**	**Yes**	68/113 (0.6)	81/102 (0.79)	0.19
eye	-2.73	0.26	**0.98**	**0.73**	**Yes**	81/113 (0.72)	91/102 (0.89)	0.18
face	-0.17	0.26	**0.98**	**0.98**	**Yes**	39/113 (0.35)	54/102 (0.53)	0.18
foot	-1.38	0.24	0.76	1.13	No	59/113 (0.52)	69/102 (0.68)	0.15
finger	-0.36	0.25	**0.84**	**0.79**	**Yes**	42/113 (0.37)	59/102 (0.58)	**0.21**
hair	-1.73	0.24	**0.89**	**1.47**	**Yes**	65/113 (0.58)	70/102 (0.69)	0.11
hand	-1.38	0.24	**0.79**	**1.15**	**Yes**	59/113 (0.52)	69/102 (0.68)	0.15
head	-2.03	0.25	1.04	1.91	Yes	70/113 (0.62)	74/102 (0.73)	0.11
knee	0.54	0.28	**0.89**	**0.83**	**Yes**	29/113 (0.26)	48/102 (0.47)	**0.21**
leg	-0.17	0.26	0.90	1.08	No	39/113 (0.35)	57/102 (0.56)	**0.21**
mouth	-2.27	0.25	1.09	2.50	No	74/113 (0.65)	80/102 (0.78)	0.13
nose	-3.24	0.28	1.02	4.57	No	88/113 (0.78)	91/102 (0.89)	0.11
booboo	0.62	0.28	**1.07**	**0.79**	**Yes**	28/113 (0.25)	40/102 (0.39)	0.14
tooth	-0.61	0.25	1.08	1.19	No	46/113 (0.41)	57/102 (0.56)	0.15
toe	-0.55	0.25	**1.06**	**1.27**	**Yes**	45/113 (0.4)	60/102 (0.59)	0.19
tongue	-0.24	0.25	**0.92**	**0.87**	**Yes**	40/113 (0.35)	56/102 (0.55)	0.20
tummy	-1.15	0.24	**0.81**	**1.05**	**Yes**	55/113 (0.49)	64/102 (0.63)	0.14
D8 Furniture & Rooms								
bathroom	-0.61	0.25	**0.99**	**0.95**	**Yes**	46/113 (0.41)	66/102 (0.65)	**0.24**
tub	-1.73	0.24	**0.92**	**0.84**	**Yes**	65/113 (0.58)	71/102 (0.7)	0.12
bed	-1.79	0.24	**0.75**	**0.53**	**Yes**	66/113 (0.58)	81/102 (0.79)	**0.21**
bedroom	-0.30	0.25	**0.90**	**0.86**	**Yes**	41/113 (0.36)	50/102 (0.49)	0.13
chair	-1.97	0.24	**0.88**	**0.97**	**Yes**	69/113 (0.61)	79/102 (0.77)	0.16
couch	0.10	0.26	**0.99**	**0.80**	**Yes**	35/113 (0.31)	54/102 (0.53)	**0.22**
crib	0.17	0.27	**0.96**	**0.78**	**Yes**	34/113 (0.3)	36/102 (0.35)	0.05
door	-2.40	0.25	**0.78**	**0.69**	**Yes**	76/113 (0.67)	82/102 (0.8)	0.13
drawer	1.55	0.33	**0.79**	**0.57**	**Yes**	18/113 (0.16)	38/102 (0.37)	**0.21**
garage	0.96	0.3	**0.77**	**0.50**	**Yes**	24/113 (0.21)	33/102 (0.32)	0.11

highchair	-0.24	0.25	**1.16**	**1.10**	**Yes**	40/113 (0.35)	47/102 (0.46)	0.11
kitchen	-0.55	0.25	**0.87**	**0.73**	**Yes**	45/113 (0.4)	60/102 (0.59)	0.19
living room	1.24	0.31	**0.89**	**0.68**	**Yes**	21/113 (0.19)	32/102 (0.31)	0.13
oven	1.34	0.32	0.97	0.83	No	20/113 (0.18)	36/102 (0.35)	0.18
playpen	3.37	0.52	0.84	0.18	No	7/113 (0.06)	9/102 (0.09)	0.03
potty	-1.03	0.24	**1.13**	**1.01**	**Yes**	53/113 (0.47)	69/102 (0.68)	**0.21**
fridge	-0.73	0.25	**0.85**	**0.68**	**Yes**	48/113 (0.42)	62/102 (0.61)	0.18
rocker	1.34	0.32	**0.88**	**0.60**	**Yes**	20/113 (0.18)	27/102 (0.26)	0.09
sink	0.70	0.29	**0.75**	**0.42**	**Yes**	27/113 (0.24)	47/102 (0.46)	**0.22**
stairs	-0.79	0.24	**0.85**	**0.71**	**Yes**	49/113 (0.43)	62/102 (0.61)	0.17
stove	1.34	0.32	**0.82**	**0.39**	**Yes**	20/113 (0.18)	34/102 (0.33)	0.16
table	-1.15	0.24	**1.07**	**1.17**	**Yes**	55/113 (0.49)	66/102 (0.65)	0.16
TV	-2.86	0.27	**0.90**	**0.76**	**Yes**	83/113 (0.73)	85/102 (0.83)	0.10
window	-0.55	0.25	**0.87**	**0.71**	**Yes**	45/113 (0.4)	55/102 (0.54)	0.14
D9 Small Household Items								
blanket	-1.09	0.24	**0.99**	**0.85**	**Yes**	54/113 (0.48)	66/102 (0.65)	0.17
bottle	-1.21	0.24	1.46	1.52	No	56/113 (0.5)	58/102 (0.57)	0.07
bowl	-0.30	0.25	0.66	0.59	No	41/113 (0.36)	62/102 (0.61)	**0.25**
box	0.70	0.29	**0.69**	**0.38**	**Yes**	27/113 (0.24)	45/102 (0.44)	0.20
broom	0.24	0.27	**0.83**	**0.68**	**Yes**	33/113 (0.29)	48/102 (0.47)	0.18
brush	-1.09	0.24	**0.82**	**0.67**	**Yes**	54/113 (0.48)	69/102 (0.68)	0.20
clock	1.24	0.31	**0.75**	**0.39**	**Yes**	21/113 (0.19)	34/102 (0.33)	0.15
comb	0.10	0.26	0.99	1.00	No	35/113 (0.31)	47/102 (0.46)	0.15
cup	-2.86	0.27	**1.00**	**0.80**	**Yes**	83/113 (0.73)	90/102 (0.88)	0.15
dish	1.55	0.33	**1.02**	**1.34**	**Yes**	18/113 (0.16)	33/102 (0.32)	0.16
fork	-0.55	0.25	**0.82**	**0.63**	**Yes**	45/113 (0.4)	63/102 (0.62)	**0.22**
glass	1.14	0.31	**1.11**	**0.74**	**Yes**	22/113 (0.19)	36/102 (0.35)	0.16
glasses	-0.17	0.26	**0.72**	**0.52**	**Yes**	39/113 (0.35)	45/102 (0.44)	0.10
hammer	1.79	0.35	**0.84**	**0.41**	**Yes**	16/113 (0.14)	27/102 (0.26)	0.12
keys	-1.21	0.24	**0.94**	**0.78**	**Yes**	56/113 (0.5)	67/102 (0.66)	0.16
lamp	1.34	0.32	**0.86**	**1.23**	**Yes**	20/113 (0.18)	32/102 (0.31)	0.14
light	-1.03	0.24	**0.99**	**0.90**	**Yes**	53/113 (0.47)	62/102 (0.61)	0.14

medicine	0.39	0.27	**0.90**	**0.67**	**Yes**	31/113 (0.27)	45/102 (0.44)	0.17
money	1.67	0.34	**0.95**	**0.61**	**Yes**	17/113 (0.15)	24/102 (0.24)	0.08
paper	0.24	0.27	**0.82**	**0.53**	**Yes**	33/113 (0.29)	50/102 (0.49)	0.20
penny	3.37	0.52	0.96	0.75	No	7/113 (0.06)	15/102 (0.15)	0.09
picture	0.46	0.28	**0.61**	**0.35**	**Yes**	30/113 (0.27)	48/102 (0.47)	**0.21**
pillow	-0.36	0.25	**1.07**	**1.02**	**Yes**	42/113 (0.37)	58/102 (0.57)	0.20
plant	1.24	0.31	**0.98**	**0.91**	**Yes**	21/113 (0.19)	27/102 (0.26)	0.08
plate	-0.04	0.26	**0.79**	**0.54**	**Yes**	37/113 (0.33)	57/102 (0.56)	**0.23**
purse	0.78	0.29	**1.07**	**1.07**	**Yes**	26/113 (0.23)	33/102 (0.32)	0.09
radio	0.78	0.29	**1.32**	**1.37**	**Yes**	26/113 (0.23)	35/102 (0.34)	0.11
scissors	1.24	0.31	0.90	0.51	No	21/113 (0.19)	34/102 (0.33)	0.15
soap	0.46	0.28	**0.74**	**0.49**	**Yes**	30/113 (0.27)	51/102 (0.5)	**0.23**
spoon	-1.91	0.24	**0.88**	**0.65**	**Yes**	68/113 (0.6)	80/102 (0.78)	0.18
telephone	-1.79	0.24	**0.91**	**0.68**	**Yes**	66/113 (0.58)	75/102 (0.74)	0.15
toothbrush	-1.91	0.24	**0.97**	**0.72**	**Yes**	68/113 (0.6)	74/102 (0.73)	0.12
towel	-0.17	0.26	**0.63**	**0.43**	**Yes**	39/113 (0.35)	55/102 (0.54)	0.19
trash	-0.55	0.25	0.90	2.48	No	45/113 (0.4)	56/102 (0.55)	0.15
vacuum	-0.10	0.26	**0.92**	**0.76**	**Yes**	38/113 (0.34)	49/102 (0.48)	0.14
watch	1.34	0.32	**0.68**	**0.36**	**Yes**	20/113 (0.18)	34/102 (0.33)	0.16
D10 Outside & Places to Go								
backyard	0.24	0.27	**1.06**	**1.18**	**Yes**	33/113 (0.29)	44/102 (0.43)	0.14
beach	2.51	0.42	0.68	0.26	No	11/113 (0.1)	22/102 (0.22)	0.12
church	1.05	0.3	1.17	1.27	No	23/113 (0.2)	38/102 (0.37)	0.17
flower	-0.55	0.25	**0.81**	**0.65**	**Yes**	45/113 (0.4)	53/102 (0.52)	0.12
garden	2.90	0.46	**0.66**	**0.17**	**Yes**	9/113 (0.08)	14/102 (0.14)	0.06
home	-1.15	0.24	**1.13**	**0.94**	**Yes**	55/113 (0.49)	59/102 (0.58)	0.09
house	-0.24	0.25	**0.80**	**0.56**	**Yes**	40/113 (0.35)	58/102 (0.57)	**0.21**
moon	0.46	0.28	**0.96**	**0.75**	**Yes**	30/113 (0.27)	44/102 (0.43)	0.17
outside	-2.33	0.25	1.10	0.88	No	75/113 (0.66)	88/102 (0.86)	0.20
park	0.62	0.28	**1.22**	**1.17**	**Yes**	28/113 (0.25)	43/102 (0.42)	0.17
party	2.90	0.46	**0.94**	**0.43**	**Yes**	9/113 (0.08)	19/102 (0.19)	0.11
pool	0.24	0.27	**1.02**	**0.79**	**Yes**	33/113 (0.29)	51/102 (0.5)	**0.21**

rain	0.10	0.26	**1.02**	**0.84**	**Yes**	35/113 (0.31)	48/102 (0.47)	0.16
rock	0.96	0.3	**0.98**	**0.85**	**Yes**	24/113 (0.21)	30/102 (0.29)	0.08
school	-0.04	0.26	**1.05**	**0.87**	**Yes**	37/113 (0.33)	61/102 (0.6)	**0.27**
shovel	1.91	0.36	**0.84**	**0.98**	**Yes**	15/113 (0.13)	18/102 (0.18)	0.04
sky	1.14	0.31	**0.85**	**0.53**	**Yes**	22/113 (0.19)	38/102 (0.37)	0.18
slide	-1.03	0.24	**0.92**	**0.75**	**Yes**	53/113 (0.47)	65/102 (0.64)	0.17
snow	2.19	0.39	**0.82**	**0.68**	**Yes**	13/113 (0.12)	24/102 (0.24)	0.12
star	0.78	0.29	**0.89**	**0.64**	**Yes**	26/113 (0.23)	41/102 (0.4)	0.17
store	1.14	0.31	**0.88**	**0.64**	**Yes**	22/113 (0.19)	43/102 (0.42)	**0.23**
sun	1.05	0.3	**0.81**	**0.40**	**Yes**	23/113 (0.2)	42/102 (0.41)	**0.21**
swing	-1.15	0.24	**1.02**	**0.81**	**Yes**	55/113 (0.49)	70/102 (0.69)	0.20
tree	-0.30	0.25	**0.88**	**0.77**	**Yes**	41/113 (0.36)	61/102 (0.6)	**0.24**
water	-1.55	0.24	**0.89**	**0.70**	**Yes**	62/113 (0.55)	74/102 (0.73)	0.18
work	1.91	0.36	**0.76**	**0.38**	**Yes**	15/113 (0.13)	31/102 (0.3)	0.17
zoo	2.90	0.46	**0.65**	**0.30**	**Yes**	9/113 (0.08)	20/102 (0.2)	0.12
D11 People								
aunt	1.67	0.34	1.49	1.51	No	17/113 (0.15)	30/102 (0.29)	0.14
baby	-1.55	0.24	**0.88**	**0.64**	**Yes**	62/113 (0.55)	73/102 (0.72)	0.17
babysitter	2.70	0.44	1.18	4.85	No	10/113 (0.09)	13/102 (0.13)	0.04
babysitter's name	2.19	0.39	1.43	2.29	No	13/113 (0.12)	23/102 (0.23)	0.11
boy	0.78	0.29	**0.88**	**0.60**	**Yes**	26/113 (0.23)	38/102 (0.37)	0.14
brother	0.24	0.27	1.82	3.92	No	33/113 (0.29)	30/102 (0.29)	0.00
child	2.90	0.46	**0.55**	**0.13**	**Yes**	9/113 (0.08)	14/102 (0.14)	0.06
daddy	-3.66	0.3	1.30	7.07	No	93/113 (0.82)	84/102 (0.82)	0.00
girl	0.87	0.29	1.11	0.91	No	25/113 (0.22)	35/102 (0.34)	0.12
grandma	-1.67	0.24	**1.19**	**1.06**	**Yes**	64/113 (0.57)	65/102 (0.64)	0.07
grandpa	-0.79	0.24	1.22	3.31	No	49/113 (0.43)	52/102 (0.51)	0.08
lady	2.90	0.46	**0.94**	**0.37**	**Yes**	9/113 (0.08)	21/102 (0.21)	0.13
man	1.79	0.35	**0.72**	**0.27**	**Yes**	16/113 (0.14)	23/102 (0.23)	0.08
mommy	-4.53	0.37	**0.91**	**1.01**	**Yes**	101/113 (0.89)	94/102 (0.92)	0.03
child's own name	-2.94	0.27	1.32	9.90	Yes	84/113 (0.74)	81/102 (0.79)	0.05
people	2.19	0.39	**0.73**	**0.39**	**Yes**	13/113 (0.12)	23/102 (0.23)	0.11

person	2.90	0.46	1.09	1.56	No	9/113 (0.08)	14/102 (0.14)	0.06
sister	0.03	0.26	1.56	2.64	No	36/113 (0.32)	36/102 (0.35)	0.03
teacher	1.34	0.32	**1.13**	**0.78**	**Yes**	20/113 (0.18)	45/102 (0.44)	**0.26**
uncle	2.05	0.37	1.47	1.52	No	14/113 (0.12)	22/102 (0.22)	0.09
D12 Games & Routines								
bath	-3.66	0.3	**0.84**	**0.65**	**Yes**	93/113 (0.82)	91/102 (0.89)	0.07
breakfast	-0.30	0.25	**0.70**	**0.51**	**Yes**	41/113 (0.36)	52/102 (0.51)	0.15
bye	-3.57	0.3	1.14	0.71	No	92/113 (0.81)	93/102 (0.91)	0.10
dinner	0.24	0.27	**0.79**	**0.55**	**Yes**	33/113 (0.29)	45/102 (0.44)	0.15
don't	-1.03	0.24	**1.25**	**1.50**	**Yes**	53/113 (0.47)	68/102 (0.67)	0.20
hello	-0.85	0.24	**0.91**	**0.90**	**Yes**	50/113 (0.44)	66/102 (0.65)	0.20
hi	-2.53	0.26	**1.03**	**1.15**	**Yes**	78/113 (0.69)	80/102 (0.78)	0.09
lunch	0.24	0.27	**0.85**	**0.62**	**Yes**	33/113 (0.29)	51/102 (0.5)	**0.21**
nap	-1.50	0.24	1.06	0.98	No	61/113 (0.54)	68/102 (0.67)	0.13
night	-3.01	0.27	1.17	1.96	No	85/113 (0.75)	86/102 (0.84)	0.09
no	-3.66	0.3	1.28	0.84	No	93/113 (0.82)	92/102 (0.9)	0.08
pattycake	-0.79	0.24	1.33	1.24	No	49/113 (0.43)	52/102 (0.51)	0.08
peekaboo	-2.53	0.26	1.14	1.04	No	78/113 (0.69)	81/102 (0.79)	0.10
please	-1.15	0.24	**1.16**	**0.95**	**Yes**	55/113 (0.49)	61/102 (0.6)	0.11
hush	-0.91	0.24	**0.89**	**1.03**	**Yes**	51/113 (0.45)	56/102 (0.55)	0.10
thank you	-1.79	0.24	1.07	1.09	No	66/113 (0.58)	75/102 (0.74)	0.15
wait	-0.73	0.25	**1.07**	**1.01**	**Yes**	48/113 (0.42)	56/102 (0.55)	0.12
want	0.62	0.28	**0.76**	**0.57**	**Yes**	28/113 (0.25)	53/102 (0.52)	**0.27**
yes	-1.73	0.24	**0.85**	**0.69**	**Yes**	65/113 (0.58)	71/102 (0.7)	0.12
D13 Verbs								
bite	-0.91	0.24	**1.01**	**0.86**	**Yes**	51/113 (0.45)	65/102 (0.64)	0.19
blow	-0.85	0.24	**0.79**	**0.56**	**Yes**	50/113 (0.44)	65/102 (0.64)	0.19
break	1.55	0.33	**0.96**	**0.55**	**Yes**	18/113 (0.16)	30/102 (0.29)	0.13
bring	-0.67	0.25	**0.92**	**0.72**	**Yes**	47/113 (0.42)	52/102 (0.51)	0.09
bump	2.19	0.39	**0.95**	**0.47**	**Yes**	13/113 (0.12)	28/102 (0.27)	0.16
clean	-0.43	0.25	**0.92**	**0.72**	**Yes**	43/113 (0.38)	64/102 (0.63)	**0.25**
close	-1.38	0.24	**0.96**	**0.75**	**Yes**	59/113 (0.52)	65/102 (0.64)	0.12

cry	-0.30	0.25	**0.70**	**0.59**	**Yes**	41/113 (0.36)	56/102 (0.55)	0.19
dance	-1.32	0.24	**1.01**	**0.80**	**Yes**	58/113 (0.51)	70/102 (0.69)	0.17
draw	-0.04	0.26	**0.81**	**0.60**	**Yes**	37/113 (0.33)	53/102 (0.52)	0.19
drink	-1.73	0.24	**0.85**	**0.72**	**Yes**	65/113 (0.58)	83/102 (0.81)	**0.24**
drive	0.96	0.3	**0.84**	**0.61**	**Yes**	24/113 (0.21)	41/102 (0.4)	0.19
eat	-3.16	0.28	**0.87**	**0.79**	**Yes**	87/113 (0.77)	85/102 (0.83)	0.06
fall	0.54	0.28	**0.75**	**0.50**	**Yes**	29/113 (0.26)	41/102 (0.4)	0.15
feed	0.62	0.28	**0.92**	**1.15**	**Yes**	28/113 (0.25)	42/102 (0.41)	0.16
finish	-0.43	0.25	1.44	1.49	No	43/113 (0.38)	59/102 (0.58)	0.20
get	-0.79	0.24	**0.79**	**0.57**	**Yes**	49/113 (0.43)	59/102 (0.58)	0.14
give	-1.38	0.24	**0.83**	**0.65**	**Yes**	59/113 (0.52)	66/102 (0.65)	0.12
go	-1.73	0.24	**0.91**	**0.89**	**Yes**	65/113 (0.58)	80/102 (0.78)	**0.21**
help	-0.49	0.25	**1.01**	**0.82**	**Yes**	44/113 (0.39)	69/102 (0.68)	**0.29**
hit	0.62	0.28	**0.77**	**0.66**	**Yes**	28/113 (0.25)	50/102 (0.49)	**0.24**
hug	-2.21	0.25	**0.84**	**0.62**	**Yes**	73/113 (0.65)	80/102 (0.78)	0.14
hurry	2.05	0.37	0.84	0.43	No	14/113 (0.12)	26/102 (0.25)	0.13
jump	-0.91	0.24	**1.12**	**0.87**	**Yes**	51/113 (0.45)	73/102 (0.72)	**0.26**
kick	-0.97	0.24	**1.04**	**1.18**	**Yes**	52/113 (0.46)	61/102 (0.6)	0.14
kiss	-3.66	0.3	**1.17**	**0.83**	**Yes**	93/113 (0.82)	91/102 (0.89)	0.07
look	-0.97	0.24	**0.80**	**0.76**	**Yes**	52/113 (0.46)	75/102 (0.74)	**0.28**
love	0.31	0.27	**1.18**	**1.47**	**Yes**	32/113 (0.28)	56/102 (0.55)	**0.27**
open	-1.73	0.24	**0.80**	**0.61**	**Yes**	65/113 (0.58)	75/102 (0.74)	0.16
play	-0.67	0.25	**0.84**	**0.64**	**Yes**	47/113 (0.42)	69/102 (0.68)	**0.26**
pull	0.39	0.27	**1.06**	**0.92**	**Yes**	31/113 (0.27)	42/102 (0.41)	0.14
push	-0.49	0.25	1.26	1.26	No	44/113 (0.39)	57/102 (0.56)	0.17
put	-0.10	0.26	**1.17**	**1.07**	**Yes**	38/113 (0.34)	42/102 (0.41)	0.08
read	-0.61	0.25	**0.64**	**0.45**	**Yes**	46/113 (0.41)	73/102 (0.72)	**0.31**
ride	0.31	0.27	**0.76**	**0.47**	**Yes**	32/113 (0.28)	51/102 (0.5)	**0.22**
run	-0.24	0.25	**0.86**	**0.58**	**Yes**	40/113 (0.35)	56/102 (0.55)	0.20
say	0.31	0.27	**0.82**	**0.80**	**Yes**	32/113 (0.28)	46/102 (0.45)	0.17
see	0.03	0.26	**0.89**	**0.62**	**Yes**	36/113 (0.32)	47/102 (0.46)	0.14
show	0.70	0.29	**0.90**	**0.62**	**Yes**	27/113 (0.24)	40/102 (0.39)	0.15

sing	-0.10	0.26	**1.07**	**1.05**	**Yes**	38/113 (0.34)	56/102 (0.55)	**0.21**
sleep	-1.26	0.24	**0.84**	**0.75**	**Yes**	57/113 (0.5)	66/102 (0.65)	0.14
smile	-0.17	0.26	**1.13**	**1.18**	**Yes**	39/113 (0.35)	49/102 (0.48)	0.14
splash	0.17	0.27	**0.87**	**0.88**	**Yes**	34/113 (0.3)	51/102 (0.5)	0.20
stop	-1.21	0.24	**1.06**	**1.37**	**Yes**	56/113 (0.5)	72/102 (0.71)	**0.21**
swim	0.54	0.28	**1.03**	**0.85**	**Yes**	29/113 (0.26)	42/102 (0.41)	0.16
swing	-1.15	0.24	**0.93**	**0.76**	**Yes**	55/113 (0.49)	62/102 (0.61)	0.12
take	0.78	0.29	0.85	0.56	No	26/113 (0.23)	37/102 (0.36)	0.13
throw	-1.61	0.24	**0.84**	**0.62**	**Yes**	63/113 (0.56)	70/102 (0.69)	0.13
tickle	-0.61	0.25	**0.87**	**0.69**	**Yes**	46/113 (0.41)	62/102 (0.61)	0.20
touch	0.31	0.27	**0.85**	**1.01**	**Yes**	32/113 (0.28)	51/102 (0.5)	**0.22**
watch	1.34	0.32	**0.54**	**0.24**	**Yes**	20/113 (0.18)	38/102 (0.37)	0.20
walk	-1.15	0.24	**0.81**	**0.60**	**Yes**	55/113 (0.49)	67/102 (0.66)	0.17
wash	-0.43	0.25	**0.76**	**0.58**	**Yes**	43/113 (0.38)	62/102 (0.61)	**0.23**
wipe	-0.24	0.25	**0.82**	**0.70**	**Yes**	40/113 (0.35)	52/102 (0.51)	0.16
write	1.34	0.32	1.17	1.03	No	20/113 (0.18)	36/102 (0.35)	0.18
D14 Time Words								
day	2.35	0.4	0.78	0.35	No	12/113 (0.11)	23/102 (0.23)	0.12
later	2.90	0.46	**1.04**	**1.18**	**Yes**	9/113 (0.08)	17/102 (0.17)	0.09
morning	3.12	0.49	**0.99**	**1.57**	**Yes**	8/113 (0.07)	26/102 (0.25)	0.18
night	1.67	0.34	1.40	1.50	No	17/113 (0.15)	37/102 (0.36)	**0.21**
now	2.05	0.37	1.04	0.70	No	14/113 (0.12)	27/102 (0.26)	0.14
today	4.42	0.68	**0.67**	**3.25**	**Yes**	4/113 (0.04)	13/102 (0.13)	0.09
tomorrow	4.01	0.61	**0.67**	**1.22**	**Yes**	5/113 (0.04)	8/102 (0.08)	0.03
tonight	4.42	0.68	**0.42**	**0.05**	**Yes**	4/113 (0.04)	8/102 (0.08)	0.04
D15 Adjectives								
all gone	-3.24	0.28	**0.96**	**0.68**	**Yes**	88/113 (0.78)	92/102 (0.9)	0.12
asleep	0.78	0.29	**1.08**	**0.92**	**Yes**	26/113 (0.23)	44/102 (0.43)	0.20
bad	1.24	0.31	1.15	1.09	No	21/113 (0.19)	39/102 (0.38)	0.20
big	0.96	0.3	1.15	1.54	No	24/113 (0.21)	41/102 (0.4)	0.19
blue	1.05	0.3	**0.98**	**0.90**	**Yes**	23/113 (0.2)	40/102 (0.39)	0.19
broken	1.55	0.33	**0.95**	**0.69**	**Yes**	18/113 (0.16)	35/102 (0.34)	0.18

careful	0.31	0.27	1.43	1.36	No	32/113 (0.28)	46/102 (0.45)	0.17
clean	0.46	0.28	**0.82**	**0.75**	**Yes**	30/113 (0.27)	56/102 (0.55)	**0.28**
cold	0.39	0.27	**0.96**	**0.63**	**Yes**	31/113 (0.27)	53/102 (0.52)	**0.25**
cute	2.35	0.4	**0.90**	**0.63**	**Yes**	12/113 (0.11)	23/102 (0.23)	0.12
dark	2.19	0.39	1.27	2.47	No	13/113 (0.12)	25/102 (0.25)	0.13
dirty	0.70	0.29	**0.91**	**0.89**	**Yes**	27/113 (0.24)	44/102 (0.43)	0.19
dry	1.79	0.35	**0.89**	**0.57**	**Yes**	16/113 (0.14)	30/102 (0.29)	0.15
empty	0.62	0.28	1.19	1.03	No	28/113 (0.25)	47/102 (0.46)	**0.21**
fast	2.51	0.42	1.00	0.76	No	11/113 (0.1)	28/102 (0.27)	0.18
fine	4.42	0.68	**0.47**	**0.06**	**Yes**	4/113 (0.04)	8/102 (0.08)	0.04
gentle	0.39	0.27	1.20	1.20	No	31/113 (0.27)	38/102 (0.37)	0.10
good	-0.04	0.26	1.14	1.08	No	37/113 (0.33)	57/102 (0.56)	**0.23**
happy	1.14	0.31	1.04	0.97	No	22/113 (0.19)	43/102 (0.42)	**0.23**
hard	4.42	0.68	**0.67**	**1.45**	**Yes**	4/113 (0.04)	17/102 (0.17)	0.13
hot	-0.97	0.24	1.06	1.78	No	52/113 (0.46)	63/102 (0.62)	0.16
hungry	-0.36	0.25	1.12	1.65	No	42/113 (0.37)	60/102 (0.59)	**0.22**
hurt	0.70	0.29	1.37	1.58	No	27/113 (0.24)	40/102 (0.39)	0.15
little	2.70	0.44	0.99	0.46	No	10/113 (0.09)	29/102 (0.28)	0.20
naughty	4.42	0.68	0.63	0.19	Yes	4/113 (0.04)	10/102 (0.1)	0.06
nice	1.34	0.32	1.18	0.87	No	20/113 (0.18)	34/102 (0.33)	0.16
old	4.42	0.68	**0.42**	**0.05**	**Yes**	4/113 (0.04)	10/102 (0.1)	0.06
pretty	2.05	0.37	1.56	2.53	No	14/113 (0.12)	31/102 (0.3)	0.18
red	1.24	0.31	**1.02**	**1.01**	**Yes**	21/113 (0.19)	38/102 (0.37)	0.19
scared	2.51	0.42	0.70	0.31	No	11/113 (0.1)	16/102 (0.16)	0.06
sick	2.90	0.46	**0.63**	**0.55**	**Yes**	9/113 (0.08)	18/102 (0.18)	0.10
sleepy	0.03	0.26	**1.13**	**0.91**	**Yes**	36/113 (0.32)	53/102 (0.52)	0.20
soft	1.91	0.36	1.29	3.12	No	15/113 (0.13)	28/102 (0.27)	0.14
thirsty	0.17	0.27	**1.01**	**0.90**	**Yes**	34/113 (0.3)	50/102 (0.49)	0.19
tired	1.55	0.33	**0.94**	**0.74**	**Yes**	18/113 (0.16)	41/102 (0.4)	**0.24**
wet	0.62	0.28	1.00	1.00	No	28/113 (0.25)	44/102 (0.43)	0.18
yucky	1.34	0.32	**1.09**	**0.90**	**Yes**	20/113 (0.18)	42/102 (0.41)	**0.23**
D16 Pronouns								

her	3.66	0.56	0.76	0.19	No	6/113 (0.05)	8/102 (0.08)	0.03
his	3.37	0.52	0.92	1.06	No	7/113 (0.06)	13/102 (0.13)	0.07
I	1.34	0.32	1.24	2.36	No	20/113 (0.18)	30/102 (0.29)	0.12
it	2.35	0.4	**0.97**	**0.65**	**Yes**	12/113 (0.11)	13/102 (0.13)	0.02
me	0.31	0.27	1.38	1.76	No	32/113 (0.28)	43/102 (0.42)	0.14
mine	0.70	0.29	1.25	1.15	No	27/113 (0.24)	49/102 (0.48)	**0.24**
my	1.67	0.34	1.02	0.50	No	17/113 (0.15)	36/102 (0.35)	0.20
that	2.35	0.4	1.19	0.76	No	12/113 (0.11)	23/102 (0.23)	0.12
this	2.19	0.39	1.08	1.04	No	13/113 (0.12)	28/102 (0.27)	0.16
you	0.62	0.28	**1.13**	**0.97**	**Yes**	28/113 (0.25)	36/102 (0.35)	0.11
your	2.19	0.39	1.24	1.02	No	13/113 (0.12)	26/102 (0.25)	0.14
D17 Question Words								
how	3.12	0.49	0.99	2.17	No	8/113 (0.07)	10/102 (0.1)	0.03
what	0.96	0.3	**1.12**	**0.88**	**Yes**	24/113 (0.21)	42/102 (0.41)	0.20
when	4.01	0.61	**0.62**	**0.15**	**Yes**	5/113 (0.04)	8/102 (0.08)	0.03
where	0.62	0.28	**1.06**	**0.93**	**Yes**	28/113 (0.25)	43/102 (0.42)	0.17
who	1.24	0.31	**1.06**	**0.83**	**Yes**	21/113 (0.19)	33/102 (0.32)	0.14
why	3.66	0.56	**0.81**	**0.45**	**Yes**	6/113 (0.05)	11/102 (0.11)	0.05
D18 Prepositions & Locations								
away	2.05	0.37	1.53	1.83	No	14/113 (0.12)	20/102 (0.2)	0.07
back	1.24	0.31	1.30	0.96	No	21/113 (0.19)	21/102 (0.21)	0.02
down	-1.44	0.24	**1.27**	**1.32**	**Yes**	60/113 (0.53)	71/102 (0.7)	0.17
in	-1.03	0.24	1.40	2.06	No	53/113 (0.47)	67/102 (0.66)	0.19
inside	1.14	0.31	1.24	0.80	No	22/113 (0.19)	43/102 (0.42)	**0.23**
off	-0.97	0.24	**1.10**	**1.41**	**Yes**	52/113 (0.46)	59/102 (0.58)	0.12
on	-0.85	0.24	**1.14**	**1.59**	**Yes**	50/113 (0.44)	64/102 (0.63)	0.18
out	-0.67	0.25	1.49	2.21	No	47/113 (0.42)	68/102 (0.67)	**0.25**
there	1.55	0.33	1.36	1.34	No	18/113 (0.16)	22/102 (0.22)	0.06
under	1.14	0.31	**1.15**	**0.88**	**Yes**	22/113 (0.19)	43/102 (0.42)	**0.23**
up	-1.61	0.24	**1.27**	**1.40**	**Yes**	63/113 (0.56)	71/102 (0.7)	0.14
D19 Quantifiers								
all	1.79	0.35	**1.11**	**0.88**	**Yes**	16/113 (0.14)	26/102 (0.25)	0.11

another	2.70	0.44	1.08	1.10	No	10/113 (0.09)	16/102 (0.16)	0.07
more	-1.91	0.24	1.35	1.61	No	68/113 (0.6)	86/102 (0.84)	**0.24**
none	2.90	0.46	**0.96**	**0.91**	**Yes**	9/113 (0.08)	18/102 (0.18)	0.10
not	3.37	0.52	0.88	0.65	No	7/113 (0.06)	17/102 (0.17)	0.10
other	3.12	0.49	0.98	1.34	No	8/113 (0.07)	19/102 (0.19)	0.12
same	2.35	0.4	**0.99**	**0.96**	**Yes**	12/113 (0.11)	18/102 (0.18)	0.07
some	2.19	0.39	1.48	1.52	No	13/113 (0.12)	22/102 (0.22)	0.10

Note: Bolded item fit statistics indicate that the item fit the IRT model adequately. Bolded difference in proportion correct indicates that > 20% of the sample improved on the item from baseline to post-intervention.