The Stability and Validity of Automated Vocal Analysis in Preverbal Preschoolers With Autism Spectrum Disorder

Tiffany Woynaroski, D. Kimbrough Oller, Bahar Keceli-Kaysili, Dongxin Xu, Jeffrey A. Richards, Jill Gilkerson, Sharmistha Gray, and Paul Yoder

Theory and research suggest that vocal development predicts "useful speech" in preschoolers with autism spectrum disorder (ASD), but conventional methods for measurement of vocal development are costly and time consuming. This longitudinal correlational study examines the reliability and validity of several automated indices of vocalization development relative to an index derived from human coded, conventional communication samples in a sample of preverbal preschoolers with ASD. Automated indices of vocal development were derived using software that is presently "in development" and/or only available for research purposes and using commercially available Language ENvironment Analysis (LENA) software. Indices of vocal development that could be derived using the software available for research purposes: (a) were highly stable with a single day-long audio recording, (b) predicted future spoken vocabulary to a degree that was nonsignificantly different from the index derived from conventional communication samples, and (c) continued to predict future spoken vocabulary even after controlling for concurrent vocabulary in our sample. The score derived from standard LENA software was similarly stable, but was not significantly correlated with future spoken vocabulary. Findings suggest that automated vocal analysis is a valid and reliable alternative to time intensive and expensive conventional communication samples for measurement of vocal development of preverbal preschoolers with ASD in research and clinical practice. ***Autism Res*** *2016, 00: 000–000.* © 2016 International Society for Autism Research, Wiley Periodicals, Inc.

Keywords: useful speech; language; vocalizations; automated vocal analysis; LENA; preschool; preverbal; autism

Introduction

Children with autism spectrum disorder (ASD) show a wide range of individual differences in their ability to use spoken words to communicate [Tager-Flusberg, Paul, & Lord, 2005]. Explaining individual differences in, or "predicting," spoken word use of preverbal preschool children with ASD is especially important because learning to use words to communicate, or acquiring "useful speech," during the preschool years has been linked repeatedly with long-term outcomes in ASD [e.g., Billstedt, Carina Gillberg, & Gillberg, 2007; Eisenberg, 1956; Gillberg & Steffenburg, 1987].

Theoretical and Empirical Support for Measuring Child Vocalizations to Predict Spoken Word Use

Infraphonological theory and related research on typically developing infants and children suggest that measuring the vocalizations of preschoolers with ASD can help us predict their spoken word use (refer to Oller [2000] for a comprehensive overview of vocal development theory and research). Children follow a predictable path in vocal development en route to spoken word use (Fig. 1). Their earliest vocalizations include sounds, such as quasi-vowels, squeals, and growls, which are not very "speech-like" relative to adult productions. These sounds do, however, reveal the emergence of foundational (or "infraphonological") capabilities for speech, such as the ability to phonate voluntarily, and they manifest the child's inclination to explore such abilities. Over the course of development, these infraphonological capacities grow, and children begin to produce canonical syllables (i.e., consonant and vowel combinations produced with adult-like speech timing) and, ultimately, an increasing number of spoken words. Thus, measuring a child's vocalizations across the early stages of language development should provide insight into the child's status on the path to spoken word use.

From the Department of Hearing and Speech Sciences, Vanderbilt University Medical Center, Nashville, Tennessee (T.W.); School of Communication Sciences and Disorders, Institute for Intelligent Systems; University of Memphis; Konrad Lorenz Institute for Evolution and Cognition Research, Austria, Memphis, Tennessee, USA (D.K.O.); Department of Special Education, Ankara University, Ankara, Turkey (B.K.-K.); LENA Research Foundation, Boulder, Colorado, USA (D.X., J.A.R., J.G.); Nuance Communications, Burlington, MA, USA (S.G.); Special Education Department, Vanderbilt University, Nashville, Tennessee (P.Y.)

Received October 02, 2015; accepted for publication June 13, 2016

Address for correspondence and reprints: Tiffany Woynaroski, Department of Hearing and Speech Sciences, Vanderbilt University, Nashville, TN. E-mail: tiffany.g.woynaroski@vanderbilt.edu

Published online 00 Month 2016 in Wiley Online Library (wileyonlinelibrary.com)
DOI: 10.1002/aur.1667

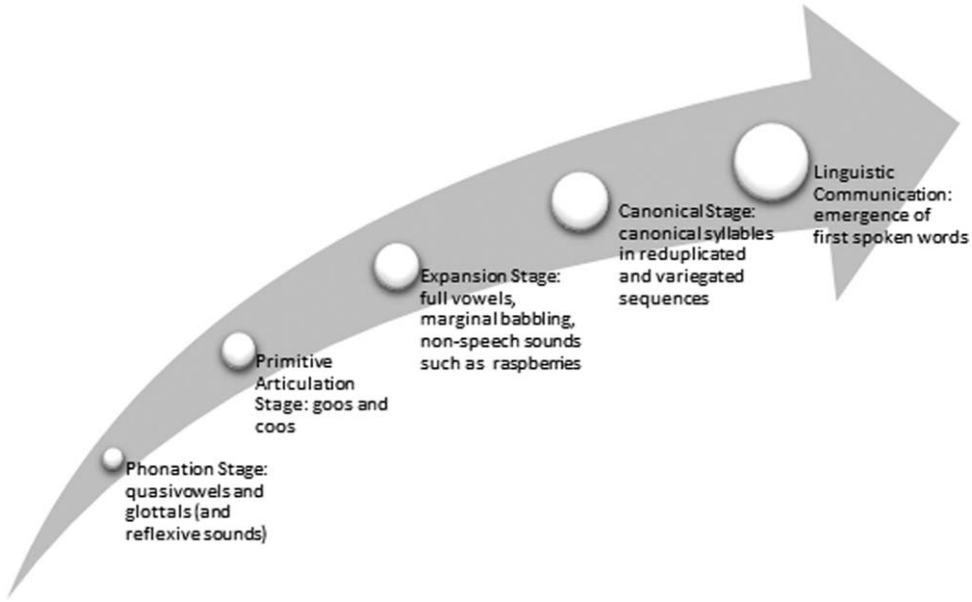

Figure 1. The stages of vocal development recognized by international consensus (Oller, 2000).

Empirical Support for Using Conventional Communication Samples to Measure Child Vocalizations

Historically, we have measured children's vocal development by collecting relatively brief (i.e., 10–15 min) samples of a child's communication in the laboratory or other settings. Child vocalizations within the samples are then identified and coded for select features of vocal development, such as canonical syllable use or the number of different consonants that are included in the child's productions. Previous research has demonstrated that such indices of vocal development can explain individual differences in concurrent spoken language use and predict future spoken language use of preschoolers with ASD [Plumb & Wetherby, 2013; Sheinkopf, Mundy, Oller, & Steffens, 2000; Wetherby, Watt, Morgan, & Shumway, 2007], even after we control for factors such as level of cognitive impairment, severity of autism, and other zero-order predictors of later spoken language growth [Yoder, Watson, & Lambert, 2015]. Thus, we know that indices of vocal development derived via conventional communication samples are valid or "useful" for predicting spoken word use of preschoolers with ASD.

Importantly, Plumb and Wetherby [2013] reported that the proportion of syllabic vocalizations *used communicatively* improved predictions of later spoken language above and beyond syllabic vocalizations produced for noncommunicative purposes in a sample of young children with ASD. Previous studies have found that preschoolers with ASD differ from their typically developing peers in their communicative use of vocalizations in addition to the complexity of their vocalizations. For example, toddlers who are later diagnosed with ASD are less likely to vocalize when communicating and more likely to vocalize for a noncommunicative purpose in comparison to typically developing peers of the same chronological age [Plumb & Wetherby, 2013; Shumway & Wetherby, 2009]. We suspected that the ability to hone in on the development of vocalizations that are specifically produced for a communicative purpose would be one advantage of conventional communication sampling over any approach to vocal analysis that does not make use of human judgments.

The use of conventional communication sampling is limited though, primarily by the large amount of time, and thus the high cost, associated with the collection and coding of communication samples. Even a brief sample may take several hours to collect and to code for features like consonant production and canonical syllable use. For example, collection and coding of a single communication sample for aspects of vocal development, such as canonical syllable and consonant use within communication acts, takes trained coders within our laboratory approximately 2–3 hr. (Excluding discrepancy discussions necessary to address coder disagreements). This collection and coding comes at a massive expense in large-scale research studies and makes it difficult, if not impossible, for clinicians to use conventional communication samples in everyday practice.

Automated Vocal Analysis as a Potential Alternative for Measurement of Vocal Development

Automated vocal analysis may provide an alternative for measurement of vocal development in preschoolers

with ASD. Automated vocal analysis using the recently available Language ENvironment Analysis (LENA, LENA Research Foundation, 2016) system allows for the identification and evaluation of child vocalizations from day-long (i.e., up to 16 hr) audio-recorded samples collected in naturalistic settings. LENA software programs that are available commercially and/or for research purposes can be used to derive several different indices that are purported to reflect children's vocal development [Oller et al., 2010; Richards, Gilkerson, Paul, & Xu, 2009; Xu, Richards, & Gilkerson, 2014].

Aside from an initial investment in hardware and software and an optional fee for maintenance and support of the LENA software (https://www.lenafoundation.org/), the automated method requires less time and thus cost relative to conventional communication sampling techniques. For example, personnel time required to prepare a LENA data collection packet, provide parental instructions on LENA use, collect a data collection packet, and upload LENA data has generally been less than 30 min in our research to date. Perhaps even more important in terms of advantages over conventional communication sampling, the automated approach then yields indices of vocal development without any further human involvement (and therefore without any further cost in listening and coding time). Thus, the automated approach could be readily applied by researchers interested in measuring children's vocal development and ultimately by clinicians hoping to incorporate measurement of vocal development into their everyday practice.

A Need to Establish the Validity and Reliability of Automated Indices of Vocal Development

However, we must first confirm that indices of vocal development derived via automated vocal analysis are valid for predicting future word use in preverbal preschoolers with ASD. The validity of any measure is limited by its reliability [Crocker & Algina, 1986], particularly the type of reliability relating to the stability of estimates of a construct across observations or contexts [McCrae, Kurtz, Yamagata, & Terracciano, 2011]. Therefore, we must ascertain both the validity and stability of automated indices of vocal development.

A recent study demonstrated that one automated index of vocal development, which we refer to as the infraphonological vocal development (IVD) score, was highly stable and valid for predicting spoken word use in a previous sample of preschoolers with ASD [Yoder, Oller, Richards, Gray, & Gilkerson, 2013]. However, the stability and validity of this score was not compared to any index derived via conventional communication samples. Additionally, other indices that can now be derived via automated vocal analysis [Richards et al., 2009; Xu et al., 2014] were not examined. Finally, and perhaps most importantly, most of the children with ASD who participated in Yoder et al. [2013] were already using many words to communicate, and the measure of language was concurrent with vocal sample collection. No studies to date have examined how stable or valid the various indices of vocal development that can be derived via automated analysis are *relative to an index from conventional communication samples* for predicting *future spoken word use* in preschoolers with ASD *who are preverbal or not yet using many words to communicate*.

Research Questions

Thus, our specific research questions were:

a. How stable are the various indices of vocal development that can presently be derived via automated vocal analysis relative to an index of vocal development derived via conventional communication samples in preverbal children with ASD?
b. How valid are indices of vocal development that can presently be derived via automated vocal analysis relative to an index of vocal development derived via conventional communication samples for predicting future spoken word use of preverbal children with ASD?

Method

Overview of Study Design

To answer these research questions, we carried out a longitudinal correlational investigation that examined the extent to which indices of vocal development derived via automated vocal analysis versus conventional communication samples predicted future spoken vocabulary in a sample of preverbal preschoolers with ASD. At Time 1, children's vocal development was measured in: (a) Two day-long audio recordings that were collected in children's natural settings and analyzed via automated vocal analysis and (b) two 15 min conventional communication samples that were collected in the laboratory and coded by the research team. Children's spoken vocabulary was measured via parent report at Time 1 and 4 months later at Time 2 for the present study. Collection of multiple audio recorded and conventional communication samples at Time 1 permitted tests of stability for each index of interest.

Participants

Participants included 20 preschool children with ASD (17 male; 3 female) who met the following inclusion criteria: (a) chronological age between 24 and 48

months; (b) diagnosis of ASD; (c) no severe sensory or motor impairments; (d) no identified metabolic, genetic, or progressive neurological disorders; (e) reported spoken vocabulary of less than or equal to 20 words on the MacArthur Bates Communicative Development Inventory: Words and Gestures (MB-CDI) [Fenson et al., 2003] vocabulary checklist; and (f) primarily English-speaking household.

Diagnoses of ASD were based on the Autism Diagnostic Observation Schedule Module 1 (ADOS) [Lord et al., 2000] and judgment that children met criteria for Autistic Disorder or Pervasive Developmental Disorder - Not Otherwise Specified according to criteria from the Diagnostic and Statistical Manual of Mental Disorders-Fourth Edition [American Psychiatric Association, 2000] by a licensed clinician on the research team who was independently research reliable on the ADOS and experienced with evaluating young children with ASD. Revised ADOS algorithms were used for improved diagnostic validity [Gotham, Risi, Pickles, & Lord, 2006]. The Mullen Scales of Early Learning [Mullen, 1995] was also administered at entry to the study to further characterize the sample. Sample characteristics are summarized in Table 1. The present sample is entirely nonoverlapping with the sample from Yoder et al. [2013].

Child Vocal Development Derived Via Automated Vocal Analysis

Children's vocal development was assessed in Two day-long audio recordings collected on consecutive days with LENA recorders [LENA Research Foundation, 2014] in children's natural settings. Parents were instructed to turn on the recorder when their child woke up in the morning, to place the recorder on their child in the chest pocket of specially designed clothing provided by the research team, and to allow the recorder to run continuously for a full 16-hr day (i.e., max LENA recording time) on any two consecutive days of the week. All parents involved in the study were able to comply with these instructions regarding data collection. When recorders were returned, research staff transferred audio files from the recorders directly to a computer for analysis.

Recordings were first processed using standard LENA software. This software uses modified speech recognition algorithms: (a) to segment the audio stream, (b) to identify or "label" segments as likely being produced by the target child or other predefined speaker/sound sources (including other child, adult male, adult female, overlapping speech/sound, electronic media, noise, silence, or unclear), and (c) to parse out target child-produced segments that are speech-related utterances (i.e., speech-related vocalizations of at least 50 ms duration bounded by sounds of other source types or silence for more than 300 ms) versus fixed signals (cries) or vegetative sounds (such as burps). After standard utterance labeling by LENA software, target child-produced, speech-related utterances were analyzed using three different approaches to automated vocal analysis to derive three different automated indices of vocal development [Oller et al., 2010; Richards et al., 2009; Xu et al., 2014].

We used software developed by Oller et al. [2010] to derive the IVD score. The IVD score was motivated by, and developed as an implementation of, infraphonological theory [Oller, 2000]. Within target child-produced, speech-related utterances, this software identifies "vocal islands," which are intended to correspond to syllable-like units. These vocal islands are assessed across 12 theoretically based infraphonological features in four perceptual categories that tap rhythm and syllabicity, squeal quality, growl quality, and vocal island duration (see Table 2, adapted from Oller et al. [2010]). Each speech-related vocal island is scored for the presence or absence of each parameter based on criterion values, and each of the 12 parameters then receives a raw score for speech-likeness based on the proportion of speech-related child utterances with at least one vocal island demonstrating the criterion level for the acoustic property of interest. To derive the IVD score, the 12-parameter raw scores are weighted by the unstandardized regression coefficients from a multiple regression equation that predicted chronological age in the normative sample of Oller et al. [2010]. The IVD score is not yet available in the standard LENA software package.

Open-source Sphinx recognition software was used to derive indices called Average Count Per Utterance (ACPU) [Xu et al., 2014] scores. These scores are also not presently available in the standard LENA software package. Briefly, the Sphinx software is used to estimate the average count of 39 different phones (acoustic matches to 24 consonants like "p" and 15 vowels like "a"), as well as periods of silence and nonspeech elements, such as coughing or lip smacking, produced in utterances identified as being produced by the target child. We emphasize that this software can provide only an "estimate" of phone production in our sample because Sphinx software was modeled entirely on adult speech. Thus, it is not yet clear that the program actually functions as a phonemic monitor in child speech. This report will focus on the ACPU of the aforementioned Sphinx-identified elements that one would theoretically expect to predict later spoken word use of preverbal preschoolers with ASD—the estimated phone counts. We derived both ACPU-Consonants and ACPU-Vowels scores. These scores were aggregated into an ACPU-Consonants + Vowels

Table 1. Sample Characteristics at Time 1

Characteristic	Mean	SD	Minimum	Maximum	Mode
Chronological age in months	37.90	6.16	25	46	42
Mullen composite age equivalency in months	9.07	3.32	3.75	16.5	8.5
Mullen expressive age equivalency in months	7.35	3.65	2	16	5.0
Mullen receptive age equivalency in months	4.05	5.34	1	22	1.0
Number of words spoken on MB-CDI	5.40	5.23	0	20	0
ADOS algorithm total score	24.15	3.42	16	28	28

Note. $N = 20$. Mullen = Mullen Scales of Early Learning (Mullen, 1995); MB-CDI = MacArthur Bates Communicative Development Inventory: Words and Gestures [*Fenson et al.*, 2003]; ADOS = Autism Diagnostic Observation Schedule [*Lord et al.*, 2000].

(ACPU-C+V) score because they were conceptually similar and empirically related. Evidence of the empirical relation will be presented in the Preliminary Results section.

The third index of vocal development that we derived via automated vocal analysis is the Automated Vocalization Analysis Developmental Age score (AVA DA) [Richards et al., 2009]. This score quantifies vocal development using phone-level information, specifically the distribution of 2,000 + uniphone pairs (i.e., biphones, acoustic matches to productions such as "pa," reflecting all possible Sphinx category combinations) detected in LENA-identified, speech-related target child utterances. The biphone distribution is reduced to 50 principal components and weighted based on multiple linear regression modeling that predicted expressive language in a normative sample from the LENA Natural Language Study [Gilkerson & Richards, 2008] to generate an age-standardized estimate of child vocal development. This estimate is itself weighted by an age-dependent variance measure (from the same normative sample) and applied as an adjustment to chronological age to approximate a child's vocal developmental age. This score may be obtained using the standard LENA software package presently available for commercial use.

Child Vocal Development Derived From Conventional Communication Samples

Children's vocal development was additionally measured via two conventional communication samples with an unfamiliar examiner: (a) the Communication and Symbolic Behavior Scales-Developmental Profile Behavior Sample (CSBS-DP) [Wetherby & Prizant, 2002], and (b) a semistructured communication sample around a standard set of toys (SSCS) [Yoder & Stone, 2006]. The SSCS is a timed, 15 min procedure. The CSBS-DP is approximately 15 min in duration, dependent upon the time required to complete all standard components of the sample.

Table 2. The 12 Parameters Comprising the IVD Score

Rhythm/syllabicity		
1	Voiced: Pitch detectable for 50% SVI	Positive classification on these parameters suggests that vocalizations tended to show voicing features, canonical formant transitions, and spectral entropy variations consistent with speech-like rhythm and syllabicity.
2	Canonical Syllable: Formant transitions < 120 ms	
3	Spectral entropy typical of speech	
Squeal quality (low spectral tilt and high pitch control)		
4	Mean pitch high (squeal): > 600 Hz	Positive classification on these parameters suggests more active expression in the high spectral frequency range (i.e., a squeal quality to vocal productions).
5	Low spectral tilt	
6	High-frequency energy concentration	
Growl quality (wide format bandwidth and low pitch control)		
7	Mean pitch low (growl): < 250 Hz	Positive classification on these parameters suggests more active expression in the low spectral frequency range (i.e., a growl quality to vocal productions).
8	Wide bandwidth (first two formants)	
Duration of SVIs within utterances		
9	Short (110–250 ms)	Positive classification on parameters nine and ten suggests speech-like rhythmic organization because duration values are typical of syllables in adult speech. Positive classification on parameters 11 and 12 suggest the opposite because the corresponding durations are beyond the range of typical syllables.
10	Medium (250–600 ms)	
11	Long (600–900 ms)	
12	Extra long (900–3000 ms)	

Note. SVI = speech-related vocal island. Table adapted from *Oller et al.* [2010]. Copyright [2010] held by *Oller et al.* Adapted with permission.

The two conventional communication samples were coded first for intentional child communication acts using a 5 sec partial interval coding system. Intentional child communication acts were defined as: (a) vocal or gestural acts combined with coordinated attention to object and person; (b) conventional gestures (e.g., showing and pointing) with attention to an adult; and (c) symbolic forms (i.e., words and sign language). Intervals coded for child communication acts were subsequently coded for the production of canonical syllables and the number of different consonants used within communication acts. Canonical syllables were defined as vocalizations in which a rapid transition occurred between vowel-like and consonant-like speech sounds, as judged by a human observer. The number of different consonants used communicatively was coded according to Wetherby's CSBS-DP True Consonant Inventory List, which includes supraglottal consonants that emerge earliest or are produced relatively frequently by young children and that are easy to code [Wetherby & Prizant, 2002].

Two component variables of child vocal development were derived from our conventional communication samples: (a) the proportion of communication acts including canonical syllables, and (b) the number of different consonants from Wetherby's True Consonant Inventory List [Wetherby & Prizant, 2002] used communicatively. We aggregated these two component variables because they were conceptually linked and empirically related. Evidence of the empirical relation amongst component variables will be presented in the Preliminary Analyses section of the Results.

Spoken Vocabulary

At Time 1 and Time 2, children's spoken vocabulary was measured via the MB-CDI: Words and Gestures vocabulary checklist [Fenson et al., 2003]. Parents were asked to check items on the MB-CDI to indicate whether their child either "understands" or "understands and says" early lexical items in categories such as actions, household items, and animals. Spoken vocabulary was the raw number of words the child was reported to say.

Preparation of Data for Analysis

The analysis method that we planned to use in evaluating the validity of our measures of child vocal development assumes multivariate normality, and multivariate normality is more likely when univariate distributions do not grossly depart from the normal distribution [Tabachnick & Fidell, 2001]. Thus, all variables were evaluated for normality. Variables showing univariate skewness $> |1.0|$ or kurtosis $> |3.0|$ were transformed prior to imputation and analysis.

Missing data points (ranging from 0% to 5% across variables) were then multiply imputed [Enders, 2011b]. Briefly, multiple imputation involved: (a) generation of 40 data sets with plausible values for missing data points, (b) analysis of each filled-in data set, and (c) pooling of the information from the multiple data sets into a single result. Plausible values are generated according to the association of variables with missing data to other variables with observed scores. This method is preferable to traditional methods for dealing with missing data (e.g., listwise deletion, single imputation, and last observation carried forward) in longitudinal data sets because it prevents loss of information related to missing data, reduces bias, improves parameter estimates, and preserves statistical power to detect effects of interest [Enders, 2011a].

Conceptually similar and empirically related component variables of vocal development that were derived via the same method (i.e., via conventional communication samples or the same approach to automated vocal analysis) were aggregated. Aggregation across component variables not only reduces the number of variables tested, but also increases the stability of scores [Rushton, Brainerd, & Pressley, 1983; Sandbank & Yoder, 2014]. Our criterion level for evidence of an empirical relation amongst component variables was a minimum covariation of 0.40 prior to aggregation [Cohen & Cohen, 1984].

Testing the Stability of Indices of Vocal Development

We carried out Generalizability and Decision (G & D) studies to test the stability of each index of vocal development (see Yoder and Symons [2010] for a comprehensive discussion of this method). The G studies examined the extent to which each index led to children being "ranked" similarly in terms of vocal development across repeated observations (i.e., audio recorded or conventional communication samples). The extent to which each index yielded similar rankings for participants across the two relevant samples was quantified by a type of intra-class correlation coefficient (ICC) called a g coefficient. The a priori criterion that we selected as indicative of "acceptable stability" for each variable was a g of 0.80.

Sometimes the stability (as quantified by g) for an index as derived from a single sample is unacceptably low. When this occurs, one can boost stability for an index by averaging estimates across several samples (e.g., audio recordings or conventional communication samples). D studies draw on information from G studies and apply a logic similar to the Spearman prophesy formula to project g coefficients beyond the number of sessions observed. The projected g coefficients can be used to "decide" how many samples across which one

needs to average to yield acceptably stable estimates of child vocal development (i.e., to achieve a criterion level for the *g* coefficient). When results of a G & D study suggested that a single sample did not yield acceptable stability for an index, we aggregated that index across the two audio recordings or conventional communication samples that we had collected in an attempt to boost the score's stability prior to testing its validity [Rushton et al., 1983; Sandbank & Yoder, 2014].

Evaluating the Validity of Indices of Vocal Development

We used linear regression to examine the predictive and incremental validity of each index of vocal development. First, we obtained zero-order correlations to examine the extent to which each index positively correlated with (i.e., showed predictive validity in explaining) future spoken vocabulary in our sample without controlling for any other factors. We subsequently used Steiger's Z-test to compare the magnitude of the zero-order correlation for each automated index to the magnitude of the zero-order correlation for the index from conventional communication samples [Lee & Preacher, 2013; Steiger, 1980]. Finally, we obtained part correlations to examine whether indices of vocal development with significant zero-order correlations continued to predict (i.e., showed incremental validity in predicting) future spoken vocabulary after controlling for concurrent spoken vocabulary in our sample. The *a priori* alpha level established for statistical significance was $P < 0.05$. Tests were one-tailed because theory and extant data suggested that all of these correlations should be positive. Throughout regression analyses, we used Cook's D to determine whether any individual data points were unduly influencing regression coefficients. In all reported associations, there was no evidence that individual participants had undue influence on estimated associations.

Results

Preliminary Analyses

Interobserver reliability. Interobserver reliability of the vocal measures from the communication samples was estimated using absolute agreement ICCs. The mean ICC for canonical syllabic communication across CSBS and SSCS samples, calculated in a way that included both unitizing errors (i.e., errors in identifying the presence of a communication act) and classifying errors (i.e., errors in identifying whether the act had a canonical syllable), was 0.96 (range 0.94–0.97). The mean ICC for consonant inventory as aggregated across CSBS-DP and SSCS samples was 0.96 (range 0.93–0.98).

Aggregation of component variables. The intercorrelation among the proportion of communication acts including canonical syllables and the number of different consonants from Wetherby's True Consonant Inventory List [Wetherby & Prizant, 2002] used communicatively across the CSBS-DP and SSCS samples was 0.88, $P < 0.001$. The intercorrelation among the ACPU-Consonants and ACPU-Vowels scores was 0.85, $P < 0.001$.

Transformation of variables. MB-CDI scores were positively skewed at both Time 1 and Time 2, but were corrected with log10 transformation.

Primary Analyses

Relative stability of indices of vocal development. Figure 2 displays the relative stability of the indices of vocal development. The aggregated index of vocal development as measured in conventional communication samples (i.e., the proportion of communication acts including canonical syllables and the number of different early emerging consonants used communicatively) approached, but did not reach, the 0.8 threshold for acceptable stability with a single sample ($g = 0.74$). However, an estimate of vocal development that met our standard for acceptable stability could be obtained by aggregating across two conventional communication samples ($g = 0.85$). Thus, we concatenated the component variables for vocal development from conventional communication samples across the CSBS-DP and SSCS prior to testing the validity of this variable.

Two of the three automated indices of vocal development, the IVD score and the AVA DA score, surpassed the 0.8 threshold for acceptable stability with 1-day-long audio recording (g values = 0.86 and 0.85, respectively). However, the third index of vocal development, the ACPU-C+V score, was not as stable for a single audio recording ($g = 0.57$). The D study for this variable indicated that 3-day-long audio recordings would be necessary to achieve acceptable stability (projected g for three samples = 0.80). By averaging the ACPU-C+V score across the 2-day-long audio recordings that we had available, we achieved a g of 0.73 for this index. A summary of the final scores used in the subsequent tests of relative validity is provided in Table 3.

Relative validity of indices of vocal development. Table 4 summarizes the results for the validity of the indices of vocal development. The index derived from conventional communication samples was significantly and strongly correlated with future spoken vocabulary in our sample. Similarly, two automated scores, the IVD score from a single audio recording and

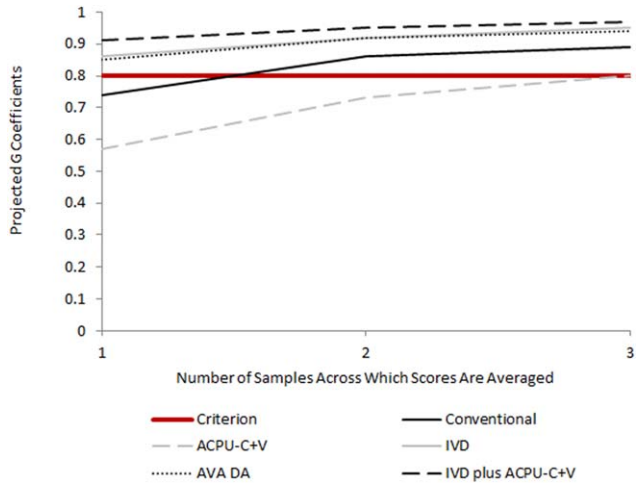

Figure 2. Relative stability of indices of vocal development derived via automated vocal analysis and conventional communication samples. IVD = Infraphonological Vocal Development. ACPU-C+V = Average Count Per Utterance of Consonants Plus Vowels. AVA DA = Automated Vocal Analysis Developmental Age. IVD plus ACPU-C+V = Aggregate of IVD and ACPU-C+V scores. Generalizability coefficients are observed for two audio recordings and projected beyond two audio recordings.

the ACPU-C+V score as averaged across two audio recordings, were also significantly and strongly associated with future spoken vocabulary. The IVD and ACPU-C+V scores were nonsignificantly different from the index from conventional communication samples in their prediction of later spoken vocabulary (*P* values for Steiger's $Zs > 0.05$). The IVD and ACPU-C+V scores, as well as the index derived from conventional communication samples, predicted future spoken vocabulary even after controlling for concurrent spoken vocabulary as reported by parents on the MB-CDI at the time children entered the study (all *P* values for part *r* values < 0.05).

In contrast, the AVA DA score did not significantly correlate with future spoken vocabulary in our sample. The association observed between the AVA DA and future spoken vocabulary was small and not in the theoretically anticipated direction. This score performed significantly worse than the index from conventional communication samples in its prediction of later spoken word use in this sample, $Z = 2.743$, $P = 0.003$.

Post Hoc Analyses

Although the two automated scores that predicted later spoken word use in our sample were nonsignificantly different from the index derived from conventional communication samples, they did account for less variance in later spoken word use (i.e., $R^2 = 0.26$ and 0.30 for IVD and ACPU-C+V scores, respectively) relative to the index derived from conventional communication samples ($R^2 = 0.41$). One of our objectives is to identify an automated index of vocal development that is "as good as" conventional communication samples in explaining individual differences in later word use in preverbal preschoolers with ASD. Thus, in post hoc analyses we explored whether a new score, which was an aggregate of IVD and ACPU-C+V scores, might account for more variance in later spoken word use than either the IVD or ACPU-C+V score alone. This score was created by averaging the z-scores for the IVD and ACPU-C+V scores. The IVD and ACPU-C+V scores met our criterion for aggregation with an intercorrelation of 0.45, $P = 0.023$.

Table 3. Summary of Constructs, Measurement Procedures, and Variables Used in Tests of Validity

Constructs	Measures	Variables used in tests of validity	Role
Future spoken vocabulary	MB-CDI	Number of words child is reported to say on MB-CDI at Time 2[a]	Outcome
Concurrent spoken vocabulary	MB-CDI	Number of words child is reported to say on MB-CDI at Time 1[a]	Covariate
Indices of vocal development derived via automated vocal analysis	Two day-long LENA-DLP recorded samples	IVD score from one audio recording ACPU-C+V from two audio recordings AVA DA score from one audio recording Combined IVD and ACPU-C+V score from one audio recording	Predictors
Index of vocal development from conventional communication samples	CSBS-DP SSCS	Average of z-scores for: (a) proportion of communication acts including canonical syllables and (b) number of different consonants used communicatively across two samples	Predictor

Note. CSBS-DP = Communication and Symbolic Behavior Scales-Developmental Profile Behavior Sample (CSBS-DP) [Wetherby & Prizant, 2002]; LENA-DLP = Language Environment Analysis—Digital Language Processor; MB-CDI = MacArthur Bates Communicative Development Inventory: Words and Gestures vocabulary checklist (*Fenson et al.*, 2003); SSCS = semistructured communication sample with the examiner. IVD = Infraphonological Vocal Development. ACPU-C+V = Average Count Per Utterance of Consonants Plus Vowels. AVA DA = Automated Vocal Analysis Developmental Age. Time 2 was 4 months after indices of vocal development were collected.

[a] MB-CDI scores were log10 transformed to correct severe positive skew.

Table 4. Validity of Indices of Child Vocalization Development

Variable	Zero-order *r* for predictive validity	Steiger's *Z* for relative validity[a]
Index derived from conventional communication samples	0.64***	NA^{ns}
Infraphonological Vocal Development (IVD)	0.51**	0.64^{ns}
Average Count Per Utterance-Consonants + Vowels (ACPU-C+V)	0.55**	0.44^{ns}
Automated Vocal Analysis Developmental Age (AVA DA)	−0.22^{ns}	2.74***
Combined IVD and ACPU-C+V	0.61***	0.16^{ns}

Note. *$P < 0.05$. **$P < 0.01$. ***$P < 0.005$. All *P* values are one tailed. ns = nonsignificant result.

[a] Indicates the Steiger's *Z* value for difference of correlation magnitude for automated index relative to conventional communication samples. Significant Steiger's *Z* indicates that this score significantly differs from index derived from conventional communication samples in prediction of future spoken vocabulary.

This combined score was highly stable with a single day-long audio recording ($g = 0.91$). The combined IVD and ACPU-C+V score from a single day-long audio recording significantly predicted future spoken vocabulary in our sample ($r(19) = 0.61$, $P = 0.002$), accounting for 37% of the variance in future word use in our sample. This score did not significantly differ from conventional communication samples in its prediction of later spoken word use, $Z = 0.16$, $P = 0.44$ (Fig. 3). It did continue to significantly predict future spoken vocabulary even after controlling for concurrent spoken vocabulary (part $r(18) = 0.25$, $P < 0.05$). This score in combination with concurrent vocabulary accounted for over three fourths of the variance in future spoken word use in our sample ($R^2 = 0.751$).

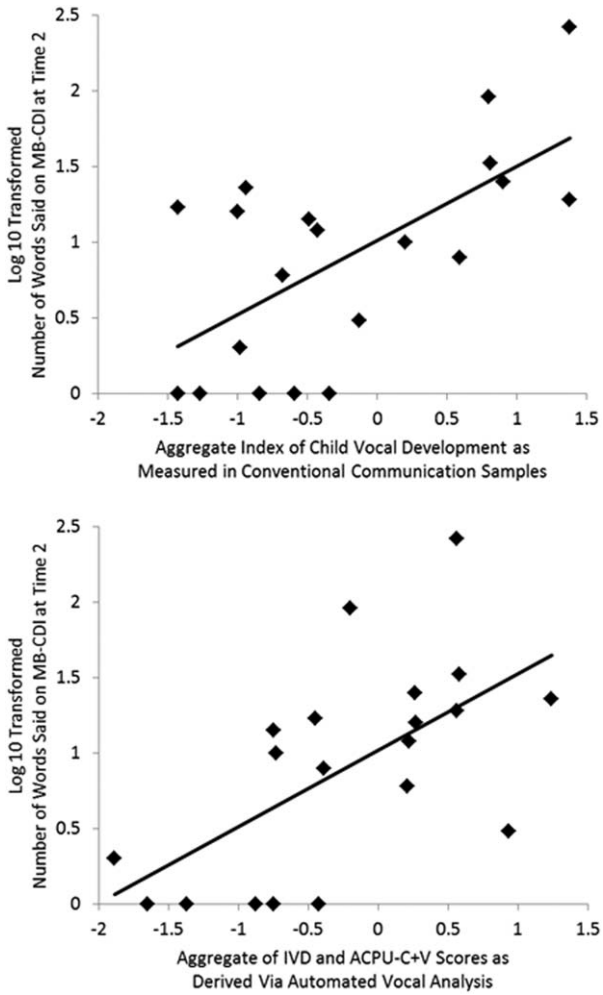

Figure 3. Relative predictive validity of the combined IVD and ACPU-C+V score and the index of vocal development derived from conventional communication samples. MB-CDI = MacArthur Bates Communicative Development Inventory: Words and Gestures vocabulary checklist [*Fenson et al.*, 2003]. Range of back-transformed values observed for number of words said at Time 2 is 0–262.

Discussion

This study examined the stability and validity of several indices that can be derived via automated vocal analysis and that are purported to measure child vocal development, a previously identified value-added predictor of "useful speech" in preverbal preschoolers with ASD [Yoder et al., 2015]. The present work extends the previous conclusions of Yoder et al. [2013], who found that one index of vocal development that can be derived via automated vocal analysis, the IVD score, correlated with concurrent spoken language in a sample of preschoolers with ASD who were largely already using words to communicate. Our results indicate that automated vocal analysis may be used to derive *several* indices of vocal development that are: (a) similarly, if not more, stable relative to indices of vocal development derived via the more conventional approach to measurement, and (b) useful as predictors of *future* spoken vocabulary for preschoolers with ASD who are *preverbal* or just beginning to use words to communicate.

Importantly, a number of automated indices were nonsignificantly different from the aggregate metric of child vocal development as measured across conventional communication samples in their prediction of future spoken vocabulary in this population. It is noteworthy that these automated indices may or may not reflect the development of vocalizations *used communicatively*. A previous study found that vocalizations used

communicatively improved predictions of spoken language in preschoolers with ASD above and beyond vocalizations that were used noncommunicatively [Plumb & Wetherby, 2013]. We found that effect sizes for the association with future spoken vocabulary were large for indices of child vocal development (with the exception of the AVA DA score), regardless of whether they were derived via a conventional approach that considered communicativeness or via an automated procedure that offered no information about the extent to which vocalizations were produced for communicative purposes.

Results suggest that an automated score that combines the infraphonological measure based on infant vocal development theory (i.e., IVD) and the measure modeled on adult phonemics (i.e., ACPU-C+V) may be most useful for predicting future word use in children with ASD. The combined IVD and ACPU-C+V score, an aggregate of z-transformed IVD and ACPU-C+V scores, performed better than either index alone and was approximately "as good as" the index that we derived from conventional communication samples in explaining individual differences in future spoken vocabulary in our sample of preverbal children with ASD. The improved validity of this combined score may, at least in part, be due to its very high stability. Future work may further evaluate whether this score may be derived in such a way (i.e., whether the many parameters that comprise the ACPU and IVD scores may be weighted such) that it performs even better in predicting word use of preschoolers with ASD.

Limitations

Three limitations are immediately apparent. First, the analyses for the combined IVD plus ACPU-C+V score were exploratory in nature. Thus, a replication of the present result is in order before we can possibly consider this particular index to be superior to the other automated scores that we evaluated. We consider the present result to support the validity of the IVD, ACPU-C+V, and the combined IVD plus ACPU-C+V score for predicting future word use in preverbal children with ASD. We additionally acknowledge that 4 months is a relatively short time frame for prediction of "future" spoken vocabulary. Subsequent studies should attempt to systematically replicate the present results over longer intervals of time. Finally, the longitudinal correlational design that we utilized does not allow us to infer a causal relation between vocal development and spoken vocabulary in this population. Although the design does demonstrate an association between vocal development and spoken vocabulary and temporal precedence for the association between early (Time 1) vocal development and later (Time 2) spoken vocabulary, it does not allow us to rule out all alternative explanations for this association. Thus, we can only conclude at present that early vocal development is associated with, but does not necessarily cause, later spoken vocabulary in preverbal preschoolers with ASD.

Future Research and Development

Future studies should directly target vocal development and test whether early effects of treatment on this construct translate to later effects of treatment on spoken vocabulary or broader spoken language skill in preverbal preschoolers with ASD. Confirmation of this type of indirect effect in a well-designed treatment study would allow us to more confidently conclude that child vocal development is causally related to later spoken word use. Incorporating automated vocal analysis into such clinical trials would allow us to additionally test whether indices of vocal development derived via the more novel, automated approach are sensitive to early effects of intervention. If we could demonstrate that effects of intervention on later word use are preceded and mediated by effects on vocal development as indexed by the automated scores that we have validated, then these automated indices may eventually be useful as an early indicator of whether a child with ASD is responding to treatment targeting the development of useful speech.

At present, the software needed to compute the automated scores that we found to be valid are not part of the standard LENA software package that is available for commercial use. Unfortunately, the one index of vocal development that can currently be derived via commercially available LENA software (the AVA DA score), though highly stable, did not prove to be useful for predicting future word use in our sample. The AVA DA score underwent extensive research and development and was previously found to be valid for predicting broader expressive language in young, typically developing children [Richards et al., 2009]. However, to our knowledge the validity of this score for predicting spoken language, or any specific aspect of spoken language (e.g., vocabulary), has not previously been tested on any clinical populations. Thus, its applicability to children with developmental disabilities, and to preverbal preschoolers with ASD in particular, is not known. As such, further research along these lines is needed.

The scores that appear to be promising (the IVD, ACPU-C+V, and/or the combined IVD and ACPU-C+V score) are presently available from LENA for research. This will allow for much-needed replication of the present results, as well as extension of this research into the stability and validity of these scores for other purposes (e.g., sensitivity to change over time and/or detection of early effects of treatment) and/or other clinical populations. Such replications and extensions

supporting the utility of the automated scores would justify inclusion of these new tools into updated versions of the LENA software made available to the general public.

Conclusion

Our findings indicate that automated vocal analysis is a valid and reliable alternative to conventional communication sampling for measuring vocal development in preverbal children with ASD. The use of automated vocal analysis to index children's vocal development may significantly reduce the time and cost necessary to conduct research in this population. This more cost-effective and time-efficient method may also eventually make it possible for clinicians to measure children's vocal development in everyday clinical practice, to make prognoses about the extent to which children with ASD are likely to use words to communicate (at least in the short term) and perhaps eventually to track whether children are progressing along the path to spoken word use in treatment, once the scores that we have tested here are further validated and made publicly available.

Acknowledgments

This work was supported by the National Institute for Deafness and other Communication Disorders (NIDCD R01 DC006893; NIDCD DC011027), the EKC of NIH under award U54HD08321 to the Vanderbilt Kennedy Center, a Preparation of Leadership Personnel grant (H325K080075; PI: Schuele), US Department of Education, CTSA award No. KL2TR000446 from the National Center for Advancing Translational Sciences, and by the Plough Foundation. The content is solely the responsibility of the author/s and does not necessarily represent the official views of the National Institutes of Health or the US Department of Education. Dongxin Xu, Jeffrey Richards, and Jill Gilkerson are employees of LENA Research Foundation. Kim Oller serves on the advisory board for, but does not derive any financial benefit from, LENA Research Foundation.

References

American Psychiatric Association. (2000). Diagnostic and statistical manual of mental disorders-IV-TR. Washington, DC: APA.

Billstedt, E., Carina Gillberg, I., & Gillberg, C. (2007). Autism in adults: Symptom patterns and early childhood predictors. Use of the DISCO in a community sample followed from childhood. Journal of Child Psychology and Psychiatry, 48, 1102-1110. doi: 10.1111/j.1469-7610.2007.01774.x.

Cohen, J., & Cohen, P. (1984). Applied multiple regression. Mahwah, NJ: Erlbaum.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. Fort Worth, TX: Harcourt.

Eisenberg, L. (1956). The autistic child in adolescence. The American Journal of Psychiatry, 112, 607-612.

Enders, C. (2011a). Analyzing longitudinal data with missing values. Rehabilitation Psychology, 56, 267-288. doi: 10.1037/a0025579.

Enders, C. (2011b). Analyzing longitudinal data with missing values. Rehabilitation Psychology, 56, 267-288. doi: 10.1037/a00255792011-22474-001.

Fenson, L., Dale, P., Reznick, J., Thal, D., Bates, E., Hartung, J., . . ., Reilly, J. (2003). MacArthur communicative development inventories: User's guide and technical manual. Baltimore, MD: Paul H. Brookes.

Gilkerson, J., & Richards, J. (2008). The LENA natural language study (LENA Technical Report 02-2) (pp. 1-26). Boulder, CO: LENA Foundation.

Gillberg, C., & Steffenburg, S. (1987). Outcome and prognostic factors in infantile autism and similar conditions: A population-based study of 46 cases followed through puberty. Journal of Autism and Developmental Disorders, 17, 273-287.

Gotham, K., Risi, S., Pickles, A., & Lord, C. (2006). The autism diagnostic observation schedule: Revised algorithms for improved diagnostic validity. Journal of Autism and Developmental Disorders, 37(4), 613-627. doi: 10.1007/s10803-006-0280-1.

Lee, I., & Preacher, K. (2013). Calculation for the test of the difference between two dependent correlations with one variable in common [Computer Software]. Retrieved from http://quantpsy.org.

LENA Research Foundation. (2014). LENA: Every word counts. Retrieved from http://www.lenababy.com/LenaHome/why-use-lena-home.aspx?.

Lord, C., Risi, S., Lambrecht, L., Cook, E.H., Jr., Leventhal, B.L., DiLavore, P.C., . . ., Rutter, M. (2000). The autism diagnostic observation schedule-generic: A standard measure of social and communication deficits associated with the spectrum of autism. Journal of Autism and Developmental Disorders, 30(3), 205-223.

McCrae, R., Kurtz, J.E., Yamagata, S., & Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scale validity. Personality and Social Psychology Review, 15(1), 28-50. doi: 10.1177/1088868310366253310888683103662531088868310366253.

Mullen, E. (1995). Mullen scales of early learning. Circle Pines, MN: American Guidance Service.

Oller, D.K. (2000). The emergence of the speech capacity. Mahwah, NJ: Lawrence Erlbaum Associates.

Oller, D.K., Niyogi, P., Gray, S., Richards, J.A., Gilkerson, J., Xu, D., . . ., Warren, S.F. (2010). Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development. Proceedings of the National Academy of Sciences of the United States, 107, 13354-13359. doi: 10.1073/pnas.1003882107.

Plumb, A., & Wetherby, A. (2013). Vocalization development in toddlers with autism spectrum disorder. Journal of Speech Language and Hearing Research, 56, 721-734. doi: 10.1044/1092-4388(2012/11-0104)1092-4388_2012_11-0104.

Richards, J., Gilkerson, J., Paul, T., & Xu, D. (2009). The LENA automatic vocalization assessment (LENA Technical Report 08-1) (pp. 1-20). Boulder, CO: LENA Foundation.

Rushton, J., Brainerd, C., & Pressley, M. (1983). Behavioral development and construct validity: The principle of aggregation. Psychological Bulletin, 94, 18-38. doi: 10.1037/0033-2909.94.1.18.

Sandbank, M., & Yoder, P.J. (2014). Measuring representative communication in 3-year-olds with intellectual disabilities. Topics in Early Childhood Special Education. 34, 133-141. doi: 10.1177/0271121414528052.

Sheinkopf, S., Mundy, P., Oller, D.K., & Steffens, M. (2000). Vocal atypicalities of preverbal autistic children. Journal of Autism and Developmental Disorders, 30, 345-354. doi: 10.1023/a:1005531501155.

Shumway, S., & Wetherby, A. (2009). Communicative acts of children with autism spectrum disorders in the second year of life. Journal of Speech Language and Hearing Research, 52, 1139-1156. doi: 10.1044/1092-4388(2009/07-0280)1092-4388_2009_07-0280.

Steiger, J. (1980). Tests for comparing elements of a correlation matrix. Psychological Bulletin, 87, 245-251.

Tabachnick, B., & Fidell, L. (2001). Using multivariate statistics (4th ed.). Boston, MA: Allyn and Bacon.

Tager-Flusberg, H., Paul, R., & Lord, C. (2005). Language and communication in autism. In F. R. Volkmar, R. Paul, A. Klin, & D. Cohen (Eds.), Handbook of autism and pervasive developmental disorders, Vol. 1: Diagnosis, development, neurobiology, and behavior (3rd ed., pp. 335-364). Hoboken, NJ: John Wiley & Sons Inc.

Wetherby, A., & Prizant, B.M. (2002). Communication and symbolic behavior scales developmental profile-first normed edition. Baltimore: Paul H. Brookes.

Wetherby, A., Watt, N., Morgan, L., & Shumway, S. (2007). Social communication profiles of children with autism spectrum disorders late in the second year of life. Journal of Autism and Developmental Disorders, 37, 960-975. doi: 10.1007/s10803-006-0237-4.

Xu, D., Richards, J.A., & Gilkerson, J. (2014). Automated analysis of child phonetic production using naturalistic recordings. Journal of Speech, Language, and Hearing Research, 57, 1638-1650. doi: 10.1044/2014_JSLHR-S-13-0037.

Yoder, P.J., Oller, D.K., Richards, J., Gray, S., & Gilkerson, J. (2013). Stability and validity of an automated measure of vocal development from day-long samples in children with and without autism spectrum disorder. Autism Research, 6, 103-107. doi: 10.1002/aur.1271.

Yoder, P.J., & Stone, W.L. (2006). A randomized comparison of the effect of two prelinguistic communication interventions on the acquisition of spoken communication in preschoolers with ASD. Journal of Speech Language and Hearing Research, 49, 698-711. doi: 10.1044/1092-4388(2006/051).

Yoder, P.J., & Symons, F. (2010). Observational measurement of behavior. New York, NY: Springer Publishing Company.

Yoder, P.J., Watson, L., & Lambert, W.E. (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, 1254-1270.