Published in final edited form as:
Autism Res. 2013 April ; 6(2): 103–107. doi:10.1002/aur.1271.

Stability and Validity of an Automated Measure of Vocal Development From Day-Long Samples in Children With and Without Autism Spectrum Disorder

P. J. Yoder[1], **D. K. Oller**[2], **J. A. Richards**[3], **S. Gray**[5], and **J. Gilkerson**[3,4]
[1]Vanderbilt University

[2]University of Memphis

[3]LENA Research Foundation

[4]University of Colorado

[5]Boulder, CO

Abstract

Lay Abstract—Measuring the degree to which young children's vocalizations, many of which are non-words, have acoustic characteristics similar to speech may eventually help us match expectations and treatment methods to individual needs and abilities. To accomplish this goal, we need vocal measures that have scientific utility. The current study indicates that a single all-day recording and subsequent computer-analysis of its acoustic characteristics produces a measure of vocal development that is highly related to expressive language in children with ASD and in children who are typically developing. These findings provide the needed basis for future use of this measure for clinical and scientific purposes.

Scientific Abstract—Individual difference measures of vocal development may eventually aid our understanding of the variability in spoken language acquisition in children with autism spectrum disorder (ASD). Large samples of child vocalizations may be needed to maximize the stability of vocal development estimates. Day-long vocal samples can now be automatically analyzed based on acoustic characteristics of speech-likeness identified in theoretically-driven and empirically cross-validated quantitative models of typical vocal development. This report indicates that a single day-long recording can produce a stable estimate for a measure of vocal development that is highly related to expressive spoken language in a group of young children with ASD and in a group that is typically developing.

Rationale for Measures of Vocal Development in Children with ASD

New measures of vocal development are needed to understand why many children with autism spectrum disorder (ASD) have expressive language delays. Recently developed technology and software provide a fully-automated candidate measure. This report addresses initial reliability and validation evidence of this new measure.

Correspondence can be sent to Paul Yoder, 228 Peabody Box, Vanderbilt University, 37203. paul.yoder@vanderbilt.edu.
Paul J. Yoder, Special Education Department, Vanderbilt University; D. K. Oller, School of Communication Sciences and Disorders, University of Memphis, Memphis and Konrad Lorenz Institute for Evolution and Cognition Research Altenberg, Austria; J. A. Richards, LENA Research Foundation, Boulder, CO; S. Gray, Boulder, CO; Jill Gilkerson, University of Colorado and LENA Research Foundation, Boulder, CO.

Jeffrey A. Richards and Jill Gilkerson work for the LENA Research Foundation, the organization that owns the database of all-day recordings collected with the LENA Digital Language Processor. The content is the responsibility of all listed authors.

Association of a candidate measure with expressive language is limited by the reliability of associated measures (Ghiselli, Campbell, & Zedeck, 1981). Reducing measurement error due to human perception differences or short behavior samples should improve the scientific utility of vocal development measures. Additionally, automated measures have the potential to quantify a greater number of aspects of vocalizations that are speech-like than traditional measures requiring human perceivers to assess vocal behavior. Inclusion of aspects of vocalizations that are over- or under-represented in the non-speech vocalizations of children with ASD in an automated measure of vocal development may enable better prediction of spoken language than relying solely on a few traditional measures based on human perception (e.g., consonant inventory and canonical babbling ratios, Paul, Fuerst, Ramsey, Chawarska, & Klin, 2011; Sheinkopf, Mundy, Oller, & Steffens, 2000).

Developments that Afford a Fully-Automated Measure of Vocal Developmental Level

Recent developments provide an opportunity for improved measurement of vocal development. A device (from the LENA Research Foundation) is now available to record day-long vocal samples (Warren et al., 2010). Software that is not yet commercially available has been used to analyze and quantify speech-like characteristics of vocalizations of infants and children without time-consuming human perceptual judgments (Oller, et al. 2010). This completely automated analysis of machine-identified child vocalizations has been implemented to assess 12 acoustic aspects of vocalization that are known to be related to speech development based on over 30 years of research (Oller, 2000; Oller et al., 2010). In addition to acoustic parameters that assess formant transitions pertaining to consonants, this automated system quantifies (among other features) the degree to which syllable-like acoustic units have (a) voicing characteristics of mature syllables, (b) speech-like variation in vocal quality, (c) high and low pitch control (as evidenced in squeals and growls), (d) speech-like frequency range, and (e) speech-like duration. The Supporting Appendix of the Oller et al. report provides data suggesting acceptable agreement between these automatically-quantified acoustic parameters and human judgments of such.

At least four advantages may be afforded by fully-automated, day-long vocal development analysis. First, day-long vocal samples may afford unusually good test-retest stability because of the greater amount of vocal data collected compared to the most common approaches to assessing speech-likeness of vocalizations (i.e., observational coding and transcription of brief, small vocal samples). Second, using the software first reported in the Oller et al., 2010 paper, analysis of multiple dimensions of speech-likeness may afford unusually good construct validity, as seen by its association with expressive language assessed by other means, compared to methods only assessing one or two aspects of speech-likeness. Third, by using an automated analysis method, the resulting measure may become accessible to practitioners and scientists who are not trained to or do not have time to perceptually analyze acoustic properties of vocalizations. Fourth, because the semantic content of recordings cannot be determined by the automated system, privacy of participants can be more easily protected than methods that require video recording or human judgment of vocalizations from audio recordings. Although the LENA recorder produces WAV files, listening is unnecessary to produce the variables under study. All the automated procedures have, however, been validated with human transcription of selected samples from the thousands of hours of recording available in the database.

Current Study Goals

The current study attempted to address two goals. First, we sought to determine how many day-long sessions are required to derive a stable vocal age equivalency score in (a) children with autism spectrum disorders (ASD), and (b) children who are typically developing (TD). Second, we sought to determine the extent to which the vocal age equivalency score is associated with expressive language in children with ASD and children who are TD.

Method

Participants

Forty children with ASD and 29 children who were TD were selected from the LENA Research Foundation database based on proximity and similarity of the interval among days on which the day-long recordings occurred. No explicit matching was attempted because between-group differences were not the focus of this report. Table 1 provides participant descriptive information.

Inclusion in the ASD group was based on parental documentation of the ASD diagnosis from independent clinicians. In Table 1, the mean and SDs for the Social Communication Questionnaire, M-CHAT, and Communication and Symbolic Behavior Scales Infant Toddler Parent Questionnaire-Social domain scores are provided. As indicated in the note for Table 1, the vast majority of the ASD group scored outside the presumed typically developing range of these instruments. Other between-group differences are consistent with the ASD diagnosis.

Procedures

Expressive language measures—Within 10 days of the vocal samples, children's parents filled out the expressive language subscale of the *Child Development Inventory* (CDI) (Ireton, 1992) and the *Language Development Survey* (LDS) (Rescorla, 1989). The age equivalency score from the expressive language subscale of the CDI and the raw score from the LDS were used for this project.

Day-long audio recordings of children's vocalizations—Children wore a small digital language processor (i.e., the recording component of the LENA system) on three days within a 10-day period from first to last vocal sample. Computational analysis of vocal samples was based on the procedure described in Oller et al. (2010), which assesses the degree to which child vocalizations have 12 (i.e., parameters) of speech. Briefly, the 12 parameters quantify various acoustic products of rhythmic movements of articulators that produce syllabic sounds, voice quality characteristics found in speech, and pitch and duration characteristics that are consistent with speech.

Scores are based on presence or absence of criterion values for each of the 12 parameters in child speech-related vocal islands. Vocal islands are at least 50 ms periods of target-child produced speech-related (non-cry, non-vegetative, and non-fixed signal) sounds with high-acoustic-energy bounded by low-acoustic-energy periods. A vocal island is an acoustic concept that is highly correlated with, but not synonymous with, the linguistic concept of "syllable." A "speech-related child utterance" is speech-related vocalization with a duration of at least 50 ms that is bounded by silence or any other sounds (TV, child's non-meaningful sounds, noise, overlapping speeches) or any other voices (adults or other children) lasting for more than 300 ms.

The raw score for each parameter of speech-likeness is the proportion of speech-related child utterances with a vocal island or vocal island(s) showing the property in question. For example, the proportion of speech-related child utterances with vocal island(s) possessing the criterion level of acoustic properties associated with a canonical syllable is one such raw score.

The age-equivalency score is an aggregate of these raw scores. The weights used to aggregate the raw scores are the unstandardized regression coefficients from the multiple regression equation quantifying the relation of the 12 raw scores to chronological age in the normative sample presented in the Oller et al. (2010) paper. The Oller et al. (2010) report indicated the raw scores were highly accurate in predicting group membership (i.e., TD, ASD, language delayed). The number of day-long recordings needed to obtain a stable vocal age equivalency measure for each child has not yet been determined, nor has the relation of the vocal age equivalency scores with expressive language been reported.

Generalizability (G) Study

Generalizability (G) theory suggests using the variance estimates from the analysis of variance (ANOVA) to estimate sources of variability in reliability samples (Cronbach, 1972). For example, studies that use G theory (G studies) can be used to identify the proportion of variance on a dependent variable accounted for by (a) what we want to measure (e.g., individual differences among participants) vs. (b) how we are measuring the variable (e.g., estimates from multiple days of vocal samples) (Shavelson & Webb, 1991).

Because the current data set has three day-long samples of vocalizations for each child, we can estimate the relative portion of variance due to participants (the variance in the across-day average of vocal developmental age among participants) relative to the total variance in the sample (participant variance + error). Using the current application of G theory, error variance is that variance due to a particular day-long sample (variance in across-participant average scores among sampling days) and to the interaction between participant and sampling day. The result is a type of intra-class correlation coefficient (ICC) that can be interpreted as the proportion of total variance due to participants. Unlike inter-class correlations (e.g., Pearson’s r), G theory (a) allows us to estimate the stability across more than a pair of observations (e.g., three sessions), and (b) provides a better match with classical measurement theory (Cronbach, 1972). In an extension of the data provided by G studies (i.e., a Decision [D] study), one can estimate the ICC for a varying number of sampling days (Yoder & Symons, 2010). We sought to determine the number day-long vocal samples across which the average vocal development scores would achieve an ICC of at least .80 (our criterion for considering the estimate “stable across short-periods”).

Results

Stability of Vocal Development Age Equivalency Scores

In the TD group, vocal development age equivalency was highly stable even with only one day’s recording (.86). Although the stability for 2 (.92), and 3 (.95) days increased slightly, additional day-long samples were not be necessary for many purposes. In the ASD group, vocal development age equivalency was also highly stable with only one day’s recording (.91). Again, the stability increased only slightly for 2 (.96), and 3 (.97) days.

Concurrent Associations with Expressive Language

In the TD group, vocal development age equivalency was strongly and positively associated with number of words parents reported their children said ($r = .78$; $p < .05$) on the LDS and with the age equivalency score derived from the Child Development Inventory, Expressive Language Scale ($r = .61$; $p < .05$). In the ASD group, vocal development age equivalency was strongly and positively associated with number of words parents reported their children said ($r = .73$; $p < .05$) and with the expressive language age equivalency score ($r = .74$; $p < .05$). See Figure 1 for an illustration of one of these associations. All tests of associations passed tests for two important statistical assumptions: no undue influence of particular participants on the regression coefficients and homoscedasticity.

Discussion

Generally speaking, the maximum observable validity is limited by the reliability of the measures used to quantify the associated variables (Ghiselli et al., 1981). This study provides initial evidence of the stability of the vocal development age equivalency score from a single day's recording in young children with ASD and in TD children. The more stable the vocal development measure, the more valid it can be for applications such as documenting or predicting treatment response and predicting expressive language.

When we consider how unstable most behaviorally-measured variables are in young children with ASD, the magnitudes of the concurrent associations in the ASD sample are remarkable. As a benchmark, the concurrent correlations for frequently used vocal development measures (e.g., canonical syllable ratios and consonant inventories) with expressive language in even typically developing children or those at-risk for autism are a maximum of .65 (e.g., Paul et al., 2011).

High correlations should not be interpreted to mean that the vocal development age equivalency score *is* a measure of expressive language, per se. Instead, the high correlations with expressive language indicate that the automatically-generated age equivalency score is *relevant to* expressive language, as would be expected for a measure of vocal development. This interpretation of the results is consistent with the nomological network theory of construct validation (Cronbach & Meehl, 1955).

This is the first paper, to our knowledge, demonstrating that purely acoustic properties of the vocalizations can be used to generate a stable quantification of speech-likeness of early vocalizations with a single day-long recording. Aside from the advantage of objectivity, the lack of necessity to involve human perception in generating the vocal development age equivalency score reduces threat to the privacy of participants, and provides great potential for access for non-technical users.

Limitations

The generalizability of current findings is limited by the characteristics of the participant samples and the concurrent correlational design. Parents of children in the ASD group provided diagnostic reports from the clinicians that had established the classification. The accuracy of these reports is unknown, but the proportion of children scoring in the autism range on available measures of autism symptomology (see the "Note" in Table 1) supports the sharp differentiation of the ASD and TD groups. Furthermore, data on children in both the TD and ASD groups showed strong justification for their classifications as reported in Oller, et al. (2010, see SI Appendix Figures S2–S4 and accompanying text), where all the children from the present study were also evaluated. Additionally, the participant samples were quite varied in their expressive language levels. This variability tends to maximize the correlations between vocal development and measures of expressive language. Future longitudinal correlation studies of children beginning in a single developmental stage (e.g., preverbal children) will likely have smaller associations between variability of vocal development measure and later expressive language.

Future Research

While concurrent associations are useful first steps towards testing the validity of the vocal development age equivalency, some of the most promising uses of this new measure involve participants with confirmed ASD diagnoses in the preverbal stage of development predicting later spoken language. Linking early vocal development with the more clearly important variable of expressive language, as in the present study, provides a relatively inexpensive basis for the next step of examining whether individual differences in early vocal development have a causal influence on later expressive language. Additionally, demonstrations that vocal development can be affected by treatment would be helpful in testing models of causal influence on expressive language. Finally, even if the new vocal measure is *not* affected by treatment, vocal development age equivalency could still be a useful predictor of differential response to non-speech vs. spoken language treatments.

Conclusion

This report contributes to the evidence testing the validation and illustrating stability of a recently-developed vocal development measure that is derived from automatic analysis of day-long samples of child vocalizations. The automatic analysis protects participant privacy and will eventually afford use by clinicians and scientists who are untrained in perceptual analysis of vocalizations. If future research continues to support the construct validity of this new measure, we may have greater ability to predict which children with ASD will learn to talk, have a proximal measure of preverbal interventions that is related to later expressive language, and have improved ability to predict whether a child is ready for spoken language intervention.

Acknowledgments

This research was supported by NIDCD R01 DC011027 to D. K. Oller and NICHD Grant P30HD15052 to the Vanderbilt Kennedy Center for Research on Human Development, and the Plough Foundation. The LENA Research Foundation generously provided the data for this article. We are grateful to the families who participated in the original LENA research.

References

Cronbach, LJ. The Dependability of behavioral measurements: theory of generalizability for scores and profiles. New York: Wiley; 1972.

Cronbach LJ, Meehl PE. Construct validity in psychological tests. Psychological Bulletin. 1955; 52:281– 302. [PubMed: 13245896]

Ghiselli, EE.; Campbell, JP.; Zedeck, S. Measurement theory for the behavioral sciences. San Francisco: Freeman; 1981.

Ireton, H. Child Development Inventory, Manual. Minneapolis: Behavior Science Systems; 1992.

Oller, DK. The Emergence of the Speech Capacity. Lawrence Erlbaum and Associates, Inc; 2000.

Oller DK, Niyogi P, Gray S, Richards J, Gilkerson J, Xu D, Warren S. Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development. Proceedings of the National Academy of Sciences. 2010; 107:13354–13359.

Paul R, Fuerst Y, Ramsey G, Chawarska K, Klin A. Out of the mouths of babes: vocal production in infant siblings of children with ASD. Journal of Child Psychology & Psychiatry. 2011; 52:588–598. [PubMed: 21039489]

Rescorla LA. Language Development Survey. Journal of Speech and Hearing Disorders. 1989; 54:587– 599. [PubMed: 2811339]

Shavelson, RJ.; Webb, NM. Generalizability theory: A primer. Newbury Park, CA: Sage; 1991.

Sheinkopf SJ, Mundy P, Oller DK, Steffens M. Vocal atypicalities of preverbal autistic children. Journal of Autism and Developmental Disorders. 2000; 30:345–354. [PubMed: 11039860]

Warren SF, Gilkerson J, Richards J, Oller DK, Xu D, Yapanel U, et al. What automated vocal analysis reveals about the vocal production and language learning environment of young children with autism. Journal of Autism and Developmental Disorders. 2010; 40:555–569. [PubMed: 19936907]

Yoder, P.; Symons, F. Observational measurement of behavior. New York: Springer; 2010.

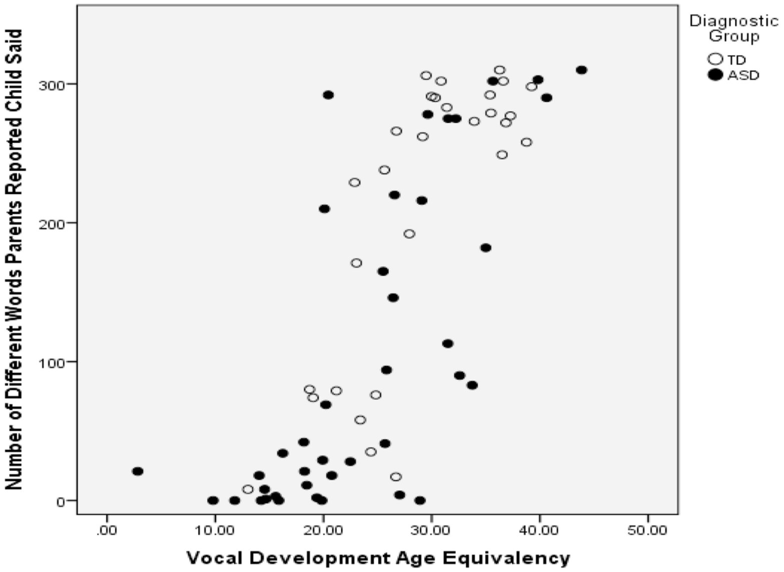

Figure 1.

Scatterplot of (a) vocal developmental age from the automated analysis of day-long samples from the 1st of the three recordings in relation to (b) parent-estimated individual differences in expressive vocabulary size from the Language Development Survey by diagnostic group.

Table 1

Characteristics of the Participants

	ASD Sample (SD)	TD Sample (SD)
% Female	18%	52% *
% Mothers college grad or above	68%	62%
Age Mean (SD)	38.6 (6.8)	28.1 (5.7) *
Predicted Vocal Age	23.7 (9.1)	29.1 (6.8) *
LDS Raw Score	105 (114)	209 (104) *
CDI-Expressive Age	21.8 (8.9)	33.7 (13.1) *
SCQ Total	18.1 (5.8)	4.8 (2.8) *
M-CHAT critical items failed	3 (2)	
CSBS-IT-Social Standard Score	4.9 (2.7)	12.7 (2.2) *

Note: BA = B; LDS = Language Development Survey; CDI = Child Development Inventory; SCQ = Social Communication Questionnaire (87.5% scored on or worse than ASD cutoff in ASD sample; high scores indicate deficits); M-CHAT number of critical items failed (73% of ASD sample scored on or worse than ASD cutoff); CSBS = Communication and Symbolic Behavior Scale Infant-Toddler Checklist (85% at or worse than risk cutoff); NA = not available.

*significant between-group difference at $p < .05$.