

Development and Performance of the LENA Automatic Autism Screen

Jeffrey A. Richards, Dongxin Xu & Jill Gilkerson
LENA Foundation, Boulder, CO

LTR-10-1

August 2010

Hardware Model: LR-0121
Software Version: V3.1.0

ABSTRACT

Autism Spectrum Disorders (ASD) are characterized by qualitative impairments in social interaction and communication as well as restricted and repetitive behaviors. The presence or absence of these behavioral signs underlies the criteria that clinicians use to assess children for ASD. However, the extensive training and expertise required to diagnose ASD and the sometimes incomplete nature of available information can limit the efficiency and reliability of screening using traditional indicators, especially at younger ages. Going beyond established diagnostic criteria, researchers have reported atypicality in the vocal production of children with ASD for features such as duration, pitch, and rhythm. Such anomalies potentially carry important diagnostic information, but on a practical level the means to explore this possibility in greater detail have been lacking. In particular, the identification of vocal features characteristic of ASD has been limited by the need to rely on resource-intensive expert judgment and the difficulty of obtaining, processing, and interpreting representative audio samples of sufficient quality and quantity. Here we report on the development and performance of a fully automatic and objective method that utilizes recent advances in technology to collect child vocalizations in large volume and evaluate discriminative vocal characteristics that could be used to help identify children at risk for ASD.

1.0 INTRODUCTION

Official diagnoses of Autism Spectrum Disorders (ASD) as specified in the Diagnostic and Statistical Manual of Mental Disorders, 4th Edition Text Revision (DSM-IV-TR; American Psychiatric Association, 2000) require that a child demonstrate a range of qualitative impairments in social interaction and communication as well as restricted and/or repetitive behaviors with onset in at least one area prior to age three. When evaluating children for ASD, clinicians rely primarily on assessments that indicate the presence or absence of these behavioral signs. Children with ASD also commonly present with delays in expressive language development, but although impaired speech development and conversation initiation are included among official diagnostic criteria, specific vocal irregularities are not. Atypical vocalizations have been described in children with ASD (Oller et al., 2010), but to date there have been few systematic attempts to collect and characterize a clinically useful acoustic feature set. As is true for child language research in general, adequately

representative language samples can be difficult to obtain. Furthermore, there is a lack of established standards and training regarding the systematic identification of ASD-specific acoustic features that may prove discriminative of children with ASD despite their significant individual variation.

In their review of research on the development of acoustic characteristics of infant vocalizations (both prespeech and speech) Oller et al. (2010) suggest that *infrastructural* expressive language properties (e.g., phonation, syllabicity, syllabic duration) may reveal developmental differences between children with ASD and other children.¹ They note that small sample studies of children with ASD have found evidence of atypical features of prosody in their vocalizations and indications that pitch and other vocal qualities in these children may differ somewhat from those of typically developing children. In addition to the small sample sizes common to research in this area, the relatively modest magnitudes of acoustic differences complicate the generalizability of results. Utilizing higher volume data sampling techniques, Oller et al. demonstrated that significant discrimination could be achieved between children with ASD, children with language delays, and typically developing children.²

Early diagnosis and intervention have been stressed as key factors in mitigating some of the longer-term impact of ASD (American Academy of Pediatrics, AAP, 2006), yet resources to enable large-scale screening of young children remain limited. Clinics specializing in the diagnosis and treatment of ASD typically have months-long waiting lists, and in less well-served geographic regions the opportunity for parents to seek out trained experts is especially limited. The AAP recommends screening for autism in children as young as 24 months of age. But, in addition to a lack of resources necessary to implement such recommendations, clinicians can face other challenges common to the evaluation of very young children, such as a paucity of appropriate assessments and typically lower validity and reliability for existing screening tools at those ages.

The reliable performance of the LENA (Language ENvironment Analysis) System demonstrates that it is possible to audio record children in their natural language environments in a relatively simple and unobtrusive manner and to obtain child vocalization

1 Infrastructural expressive language features are language independent and correspond to vocal control of the musculature of speech.

2 The technology and dataset used in Oller et al. (2010) were the basis for the research described in this report. Differences in methodology and acoustic feature sets are outlined in section 2.5.

samples that are both high quality and of sufficient quantity to permit meaningful analysis of acoustic features. Moreover, the fully automated statistical algorithms that underlie the system have been shown to be both valid and reliable with respect to distinguishing child vocalizations from adult speech and ambient environment sound (Xu, Yapanel & Gray, 2008).³ (See Xu, Richards, Gilkerson, Yapanel, Gray & Hansen [2009] for a detailed discussion of the feasibility, reliability and validity of the current system and the automated processing of productive language in children.)

This report describes the development and testing of a novel screening tool to identify children at risk for ASD based on automated modeling of acoustic features in their vocalizations. This modeling attempts to characterize the vocalizations of children with ASD in order to distinguish them from children with other language disorders and from typically developing children. This report details the statistical approach, evaluates its performance with respect to accurate classification, and compares the results with those of standard assessment tools. Research and clinical implications of this technology and directions for additional investigation are discussed.

3 The term *child vocalization* used with respect to the LENA System refers to *all* sounds originating from a child's vocal tract, not only those that are or may be considered to be related to speech. Any given child vocalization may include speech-related utterances and/or nonspeech sounds, such as breathing and vegetative or fixed-signal sounds.

2.0 DEVELOPMENT OF THE LENA AUTOMATIC AUTISM SCREEN (AAS)

2.1 The LENA System

Recording data used in the development of the AAS were collected using the LENA System, a language monitoring and feedback tool designed to facilitate data collection in the natural language environment and provide information about the development of infants and toddlers ages 2 months through 48 months. The hardware component of the system includes a lightweight digital audio recorder that is placed in the chest pocket of custom-made clothing worn by the child of interest (referred to in the system as the *Key Child*). The unit records *Key Child* vocalizations, adult speech, and any other sounds in the child's immediate environment (within at least a 6- to 10-foot radius) for up to 16 hours (Ford, Baer, Xu, Yapanel & Gray, 2008).

Audio recording data are uploaded to a Microsoft Windows-based computer for detailed analyses by the system's software component. The core process utilizes a Gaussian mixture model approach incorporating modified speech recognition algorithms to differentiate speech and speech-related sounds from environmental background noise. Additional analyses provide overview statistics, including count estimates for child vocalizations, adult speech, and other features of the natural language environment (Xu, Yapanel & Gray, 2008).

Though mathematically complex, this computational process may be summarized at a conceptual level. During processing, the continuous, daylong audio stream is parsed into variable length sound segments and reduced to component acoustic features. Each segment is identified or labeled as representing one of eight unique types based on its statistical similarity to corresponding sound source models: *Key Child*, *Adult Male*, *Adult Female*, *Other Child*, *Overlap*, *Noise*, *Electronic Media* (primarily TV and radio), and *Silence*. *Key Child* and *Adult* segments are further processed to estimate the child's vocalization count and adult word count. Importantly, although speech recognition algorithms are utilized in this process, neither vocalizations nor word estimates derive from the identification of individual words or sounds; instead, they are generated using the underlying acoustic properties of the labeled segments and previously defined regression models.

2.2 Validity of Key Child Vocalization Identification

Because AAS scores are generated exclusively from automatically identified Key Child vocalization segments, it is important to validate the accuracy of the segmentation system with respect to Key Child. As detailed in Xu, Yapanel & Gray (2008), the minimum duration of a child vocalization segment is 600 ms, but the segmentation identification and labeling process is performed by the software at a *frame* resolution, for which each frame is 10 ms in duration. Thus, there are a minimum of 60 frames per child segment. The system accuracy reported here is based on a frame-level comparison of 70 hours of human transcriber-labeled data with results obtained from the segmentation and identification algorithms. Table 1 presents frame totals (in 100K) for Key Child vs. Other segments identified by human transcribers compared to those identified by the LENA System. Standard classification performance statistics also are provided. Overall sensitivity (i.e., Key Child segment detection) is good at 76%, and specificity is excellent at 96%; Cohen’s kappa, which adjusts for differences in the classification distributions, is good at 69%. Therefore, we conclude that the automated system can identify Key Child vocalizations accurately and in sufficient quantity to permit their use in subsequent analyses.

Table 1: Human and LENA Algorithmic-Based Detection and Classification of Sound as Either Key Child Vocalizations or Other (Total Frames in Units of 100K)

		LENA System		Total
		Key Child Vocalizations	Other	
Human Transcribers	Key Child Vocalizations	14.7	4.6	19.3
	Other	6.8	166.4	173.3
Total		21.5	171.0	192.6

Sensitivity: 76%	(+) Predictive Power: 68%	Overall Accuracy: 94%
Specificity: 96%	(-) Predictive Power: 97%	Cohen’s Kappa: 69%

2.3 Data Collection and Training Samples

Audio recording data used to develop, train, and test AAS algorithms included 1,486 recordings contributed by children ranging in age from 10 months to 48 months from 232 families comprising three distinct diagnostic samples: Typically Developing (TD), Language Delay (LD), and Autism Spectrum Disorder (ASD). Each diagnostic sample includes two subsamples collected at different time points over a three-year period following approximately similar protocols; procedural differences specific to each sample are noted below. A summary of selection criteria is provided in Appendix A. Recordings routinely covered the period from the child’s being dressed in the morning until bedtime. No families in any of the samples used in the present study received feedback during their recording periods. All participating families provided informed consent, and protocols were approved by the Essex Institutional Review Board.

For a more complete description of sample recruitment and procedures see Oller et al. (2010). Recording session characteristics are summarized in Table 2, and additional demographic information for each sample can be found in Appendix B.

Table 2: Recording Session Characteristics for TD, LD and ASD Sample Participants

		TD		LD		ASD	
		Sample 1	Sample 2	Sample 1	Sample 2	Sample 1	Sample 2
Number of Participants		76	30	28	21	34	43
Number of Recordings		712	90	270	63	225	126
Recording Frequency		1x/month	3x/week	1-3x/week	3x/week	1x/week	3x/week
Weeks of Recording	Mean (SD)	37.9 (14.4)	0.5 (0.2)	21.7 (2.8)	0.7 (0.6)	5.7 (2.4)	0.8 (1.0)
	Range	0-53	0-1	9-25	0-2	0-9	0-7
Recordings Per Child	Mean (SD)	9.4 (3.4)	3.0 (0.0)	9.6 (1.2)	3.0 (0.0)	6.6 (2.0)	2.9 (0.3)
	Range	1-13	3	5-10	3	2-8	1-3
Child Age Month	Mean (SD)	28.5 (10.4)	27.3 (5.7)	26.7 (7.5)	32.0 (6.4)	33.6 (7.9)	37.8 (7.0)
	Range	10-48	18-37	10-40	22-44	16-48	24-48

Typically Developing Samples

Sample 1 includes 76 children from the Denver metropolitan area who participated in a longitudinal effort to establish normative values for measures produced by the LENA System (Gilkerson & Richards, 2008). Recordings were conducted in the home (with audio recorders and paperwork delivered and returned via courier service) and were contributed at monthly intervals. Children in Sample 1 were evaluated on site by a certified speech-language pathologist (SLP). Sample 2 includes an additional 30 children recruited nationally whose parents each recorded three times over the course of approximately 10 days. Parents in both samples completed questionnaires that assessed their child's language development and other development. Appendix C contains a summary of assessments collected across samples.

Language Delay Samples

Sample 1 includes 28 children recruited from the Denver metropolitan area who had been diagnosed by a pediatrician or certified SLP with some form of language delay (without ASD). Families participated over a 6-month period, recording from 1-3 times per month. Children also were assessed on site during the study by a certified SLP. Sample 2 includes an additional 21 children, on average slightly older and with more severe language delays than Sample 1 participants. This sample was recruited nationally, and participating families recorded three times over an approximate 10-day period. Parents in both samples completed questionnaires describing their child's language development and other development (see Appendix C), and parents in Sample 2 provided documentation of their child's language delay diagnosis.

ASD Samples

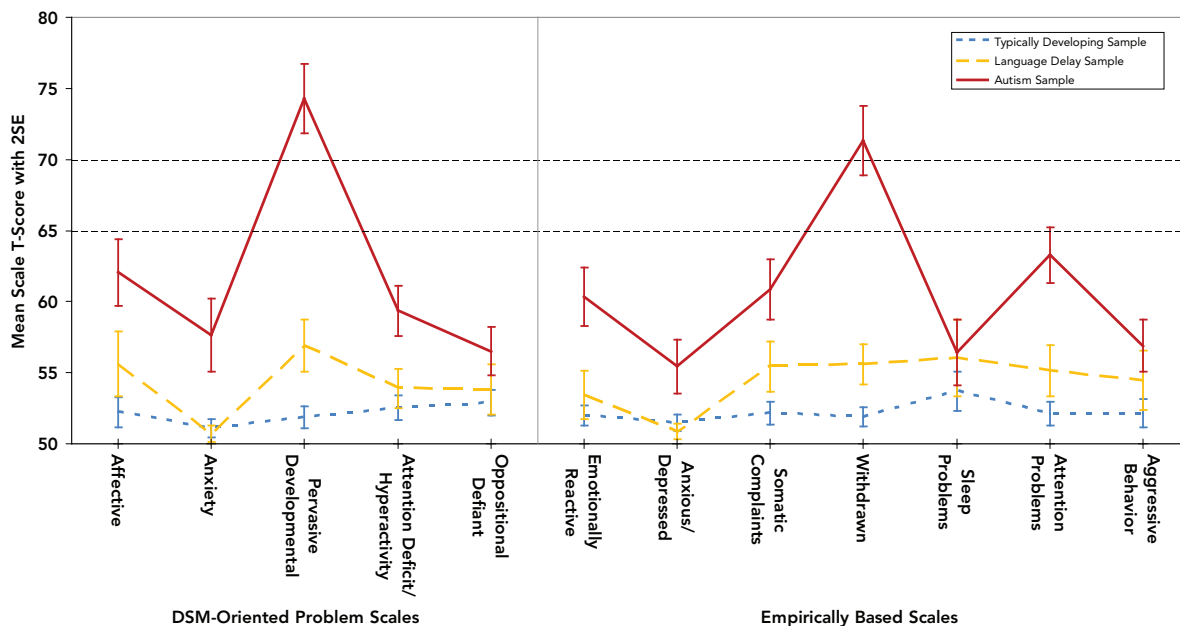
Both Samples 1 and 2 were recruited nationally, and all parents were required to provide written documentation of their child's diagnosis of ASD by a qualified professional or team of professionals. Children diagnosed with Asperger syndrome were excluded. Sample 1 includes 34 families who recorded once per week over seven weeks (with a second recording the first week), and Sample 2 includes an additional 43 families who recorded three times over the course of approximately 10 days. Parents in both samples completed questionnaires regarding their child's ASD symptomatology, language and other development (see Appendix C).

Combined Samples

For the development and testing of the AAS, recording data for Samples 1 and 2 within each diagnostic group were combined. Although not all children in each sample were assessed locally during the recording period, self-report data provided by parents demonstrate expected differences between diagnostic groups.⁴ See Appendix D for a comparison of the full ASD sample with previously published, well-documented samples on standard clinical measures.

Figure 1 shows average Child Behavior Checklist (CBCL; Achenbach & Rescorla, 2000) clinical scale scores for Sample 2 of the TD and LD groups compared with the combined ASD sample. Consistent with their diagnosis, the ASD samples on average displayed clinical range elevations on the Pervasive Developmental Problems DSM-oriented scale and the empirically based Withdrawn scale. Figure 2 plots Communication and Symbolic Behavior Scales (CSBS; Wetherby & Prizant, 2002) subscale scores for these same samples. Once more, the ASD samples on average scored in the range of clinical concern, particularly on the Social composite subscale.

Figure 1: CBCL subscale profiles across diagnostic samples



⁴ Assessment data presented here were collected from ASD Samples 1 and 2 and from TD Sample 2 and LD Sample 2.

2.4 Statistical Implementation

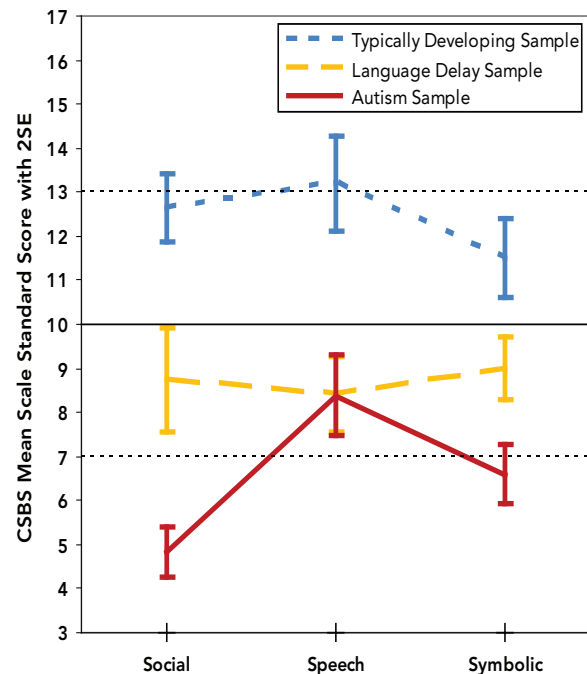
The currently implemented analytical approach (referred to here as the *Detailed Spectrum [DS]* approach) combines two complementary methods for first detecting and characterizing unique discriminative patterns in the vocalizations produced by children with ASD and then deriving classification probabilities from them. Both methods incorporate a statistical decomposition of acoustic features extracted from child vocalization segments, but they differ to some degree in the specific nature of the decomposition of the acoustic feature set. In this report we refer to the two methods respectively as *phone-based* and *cluster-based*.

Each method starts with the segmentation and identification of vocalization data from the child of interest, and each includes an analysis of acoustic characteristics in these vocalizations to identify patterns that are consistent with those produced by the ASD sample. The validity of the LENA System regarding child vocalization identification was described previously. Following the segmentation and identification process, acoustic feature sets are extracted from the child vocalizations as mel-frequency cepstral coefficients (MFCC). MFCC data are then analyzed using the phone-based and cluster-based methods.

Phone-Based Method

The phone-based method defines a unique acoustic feature set using a quantitative approach that incorporates modified components of automatic speech recognition software as described in more detail in Xu et al. (2009). Because the goal is not to recognize or translate speech, it is not necessary that extracted acoustic features correspond to specific, identifiable phones but only that the processing provide consistent, reliable results. In brief, child vocalizations are parsed into 46 pre-defined categories that correspond roughly to 39 *uniphone-like* and 7 *filler* types. Contiguous uniphone and filler pairs are subsequently joined to produce up to 46^2 (2,116) *biphone-like* combinations that encompass longer

Figure 2: CSBS subscale profiles across diagnostic samples



vocalization units. Biphone frequencies within each recording are standardized by the total count for that recording. Finally, the resulting biphone frequency distribution is subjected to a principal component analysis (PCA) to reduce the feature set to the first 50 components. For a more detailed description of this phone-based methodology, see Richards, Gilkerson, Paul & Xu (2008).

Cluster-Based Method

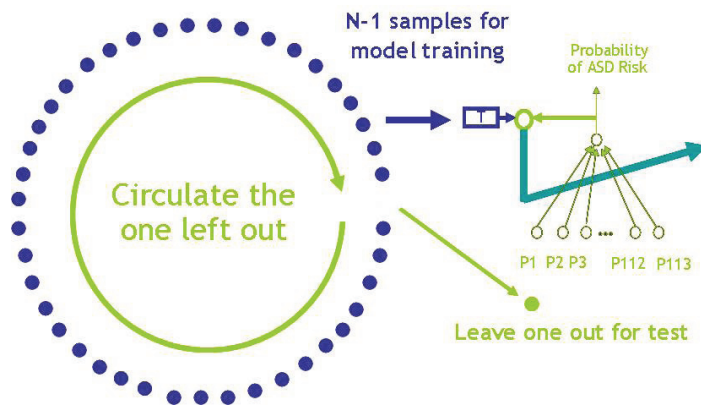
In contrast to the pre-defined categories of the phone-based method, the cluster-based method utilizes an unsupervised (i.e., data-driven) k-means clustering routine applied directly to child vocalization segments. There are no pre-defined phone-like or other a priori categories. Instead, a mathematical approach (N-dimensional analysis) is employed that searches for spectral features optimized to differentiate groups maximally. The application of this self-organized approach to vocalization MFCC data from our sample generated 63 independent phone-like clusters (or features). In short, whatever unique information could be detected that reliably distinguished vocalizations of the children with ASD from other children was utilized statistically in the most efficient manner. Once more, it is unnecessary to characterize the content of these features in order to use them for classification purposes, though with further study it should be possible to do so.

Combined-Methods Modeling

Ultimately, the 50 phone-based features and 63 cluster-based features are combined into one 113-element set of acoustic features. A linear discriminant analysis (LDA) function is computed by utilizing this feature set to generate for each recording the probability of classification to the target group of interest, in most cases the ASD group. This modeling process is conducted at the *recording-level*; that is, data from each recording are included independently, and potential non-independence effects arising from within-family similarities across multiple recording sessions are ignored.

Leave-One-Out Cross-Validation

To maximize data usage for LDA modeling and to enhance the overall generalizability of performance results to new samples, we incorporated a statistical technique called leave-one-out cross-validation (LOOCV). Figure 3 provides a graphical illustration of the process. In this case, the “one” that is left out refers to the selection of a specific child out of the overall sample of N participants. All data contributed by that child are temporarily removed from the training dataset and modeling proceeds based on the remaining N-1 participants. The resulting model is then applied to the left-out child’s data, and results for that child only are retained.

Figure 3: Illustrating Leave-One-Out Cross-Validation

Finally, the left-out child's data are reinstated, and data from a different child is removed. This process repeats until results for all children in the sample have been obtained. In this way, results for each child are derived from training models that excluded that child's data. A potential drawback of the LOOCV method is that, in general, the model applied to any

given child will differ slightly from that applied to any other child in the sample. However, given a sufficiently large sample size such differences can be expected to be negligible.

2.5 Comparison with the 12-Parameters Approach

The statistical modeling-based DS approach described in this report can be considered to be an extension of that described in Oller et al. (2010), here referred to as the 12-Parameters (12P) approach. The recording dataset from which the 12P approach was developed is also the basis for the current DS approach. However, important differences exist in the implemented methodologies.

First, the DS approach utilizes entire child vocalization segments, whereas the 12P approach pre-filters to some degree within segments to retain only potential speech-related segment portions.⁵ Second, the 12P approach specifies a priori acoustic feature categories that were selected based on their theoretical potential to differentiate children with ASD from other children. Such a condition limits consideration to a small set of fairly well-defined features. Accordingly, the predictive power of the more limited model may be reduced, though from an explanatory perspective it may be advantageous.

In contrast, the DS approach is a statistical one. Its emphasis is not on validating the theoretical justification for the models or on enhancing the clarity of the feature set. Instead, the goal is to maximize statistical power to achieve the highest levels of accuracy. In other words, the

5 Child vocalization segments represent all sounds originating from a child's vocal tract and so typically comprise both speech-related and nonspeech-related information (e.g., breath and vegetative sounds). The 12P approach used acoustic energy information to identify those portions of vocalization segments most likely to correspond to speech-related sounds and excluded everything else from further analysis. The DS approach made no such distinction and retained the entire segment for analysis.

DS approach is intended foremost to optimize the acoustic feature set that can be identified to maximally differentiate children with ASD from other children. Clarification of structural details of the resulting feature set and their correspondence to theoretical bases and readily quantifiable physical or acoustic components remains a topic for future investigation.

3.0 RESULTS

3.1 Performance Metrics

AAS performance was evaluated following two approaches: 1) 2 x 2 classification tables were constructed to compare AAS classification with a priori diagnostic categories; and 2) AAS posterior classification probabilities (i.e., the likelihood of classification to the target group for each comparison) were correlated with independently collected measures of symptomatology. Classification performance for the first approach was evaluated after fixing sensitivity and specificity at the Equal Error Rate (EER) threshold determined by a Receiver Operating Characteristic (ROC) analysis.⁶ For the second approach, the original posterior probabilities (p) were transformed to a linear scaling using the logit link function ($\text{logit}[p] = \ln[p/(1-p)]$). Additional details and performance results are provided below.

3.2 Classification Performance

Classification performance is reported here for three comparisons: 1) ASD vs. Non-ASD (TD + LD) samples; 2) ASD vs. TD samples; 3) ASD vs. LD samples. All comparisons shown were conducted at the *child-level* by collapsing results within-child across individual recordings. Child-level classification probabilities were computed to be the geometric average of each child's recording-level probabilities.

6 In a 2x2 classification task based on a continuous variable, it is necessary to set a threshold for detection of the target group of interest. The threshold value determines the number of correct detections and the number of incorrect acceptances (false positives) and incorrect rejections (false negatives). There is an inherent trade-off between the two types of error; i.e., as the false positive rate decreases the false negative rate increases and vice versa. A commonly used index of classification performance is the Equal Error Rate (EER), the error rate at the threshold point at which the false positive and false negative rates are equal. Thus, the lower the EER, the fewer classification errors of any kind. For a specific classification task it may be desirable to favor one error type over another.

Performance was evaluated along the six criteria summarized in Table 3 (using identification of the ASD group as an example): sensitivity, specificity, positive predictive power (PPP), negative predictive power (NPP), overall accuracy, and Cohen’s kappa. All criterion values were computed relative to the classification threshold value obtained at the EER point.

Table 3: Classification Performance Criteria for ASD Group Identification

Criterion	Definition
Sensitivity	Percentage of the ASD group correctly classified
Specificity	Percentage of the non-ASD group correctly classified
Positive Predictive Power	Percentage of those classified to the ASD group from the ASD group
Negative Predictive Power	Percentage of those classified to the non-ASD group from the non-ASD group
Overall Accuracy	Percentage of the overall sample correctly classified
Cohen’s Kappa	An adjusted accuracy measure that compensates for potential inflation related to distributional differences between the target and non-target groups

Tables 4, 5, and 6 detail classification and performance metrics on the three 2-way comparisons of interest; additional comparisons are provided in Appendix E. Classification performance was evaluated both on the full participant sample and on a subset created by restricting recording age to 24 months to 48 months of age. The target age range for the AAS was 24 months to 48 months; however, recording data contributed by children younger than 24 months was utilized in the modeling process.

Table 4: Child-Level Classification Performance: ASD vs. Non-ASD^a

4a: All participants

		AAS Classification		Total		
		ASD	Non-ASD			
Criterion Classification	ASD	68	9	77	Sensitivity:	.88
	Non-ASD	18	137	155	Specificity:	.88
	Total	86	146	232	Positive P.P.:	.79
					Negative P.P.:	.94
					Overall Accuracy:	.88
					Cohen's Kappa:	.75

4b: Participants 24 – 48 months of age^b

		AAS Classification		Total		
		ASD	Non-ASD			
Criterion Classification	ASD	66	7	73	Sensitivity:	.89
	Non-ASD	13	110	123	Specificity:	.89
	Total	79	117	196	Positive P.P.:	.84
					Negative P.P.:	.94
					Overall Accuracy:	.90
					Cohen's Kappa:	.79

- a Child-level classification probabilities were computed as the geometric average of within-child recording-level classification probabilities. The Non-ASD comparison group combines participants from both the Typically Developing and Language Delay samples.
- b The AAS is intended to be applicable to children age 24 months to 48 months. Recording data contributed by children younger than 24 months were utilized in the model training process but are excluded from this table.

Table 5: Child-Level Classification Performance: ASD vs. Typically Developing^a

5a: All participants

		AAS Classification		
		ASD	TD	Total
Criterion Classification	ASD	71	6	77
	TD	8	98	106
Total		79	104	183

Sensitivity:	.92
Specificity:	.92
Positive P.P.:	.90
Negative P.P.:	.94
Overall Accuracy:	.92
Cohen's Kappa:	.84

5b: Participants 24 – 48 months of age^b

		AAS Classification		
		ASD	TD	Total
Criterion Classification	ASD	67	6	73
	TD	7	74	81
Total		74	80	154

Sensitivity:	.92
Specificity:	.92
Positive P.P.:	.91
Negative P.P.:	.93
Overall Accuracy:	.92
Cohen's Kappa:	.83

- a Child-level classification probabilities were computed as the geometric average of within-child recording-level classification probabilities. Participants from the Language Delay samples were excluded from this comparison.
- b The AAS is intended to be applicable to children age 24 months to 48 months. Recording data contributed by children younger than 24 months were utilized in the model training process but are excluded from this table.

Table 6: Child-Level Classification Performance: ASD vs. Language Delay^a

6a: All participants

		AAS Classification		Total
		ASD	LD	
Criterion Classification	ASD	66	11	77
	LD	7	42	49
Total		73	53	126

Sensitivity:	.86
Specificity:	.86
Positive P.P.:	.90
Negative P.P.:	.79
Overall Accuracy:	.86
Cohen's Kappa:	.70

6b: Participants 24 – 48 months of age^b

		AAS Classification		Total
		ASD	LD	
Criterion Classification	ASD	62	11	73
	LD	6	36	42
Total		68	47	115

Sensitivity:	.85
Specificity:	.85
Positive P.P.:	.91
Negative P.P.:	.77
Overall Accuracy:	.85
Cohen's Kappa:	.69

- a Child-level classification probabilities were computed as the geometric average of within-child recording-level classification probabilities. Participants from the Typically Developing samples were excluded from this comparison.
- b The AAS is intended to be applicable to children age 24 months to 48 months. Recording data contributed by children younger than 24 months were utilized in the model training process but are excluded from this table.

Ordinal Scoring

For interpretive convenience, continuous LDA classification probabilities were reduced to seven ordinal categories using variable thresholds derived from the sensitivity, specificity, and EER point of the training data models. Threshold values were computed at the child-level for all participants. Table 7 shows the relationship between AAS ordinal scores, the criterion indices on which each is based, and their corresponding probability and criterion threshold values. For example, the lower threshold probability value for an AAS score of 4 is 0.08, which corresponds to a sensitivity value of 0.95 (i.e., a 5% false negative rate). Threshold values for all participants computed at the recording-level are provided in Appendix F.

Table 7: Child-Level Criterion Threshold Values for AAS Ordinal Scores

AAS Ordinal Score	Criterion Index	Lower Threshold Probability Value	Lower Threshold Criterion Value
1		0	1
2	Sensitivity	.01	.99
3		.04	.97
4		.08	.95
5		EER Point	.18
6	Specificity	.28	.95
7		.94	.99

Figure 4 depicts the relationship between sensitivity and specificity as a function of AAS classification threshold probability. Ordinal score boundaries (vertical lines) reflect the lower threshold probability values shown in Table 7. Data markers indicate child-level averages from all participants.

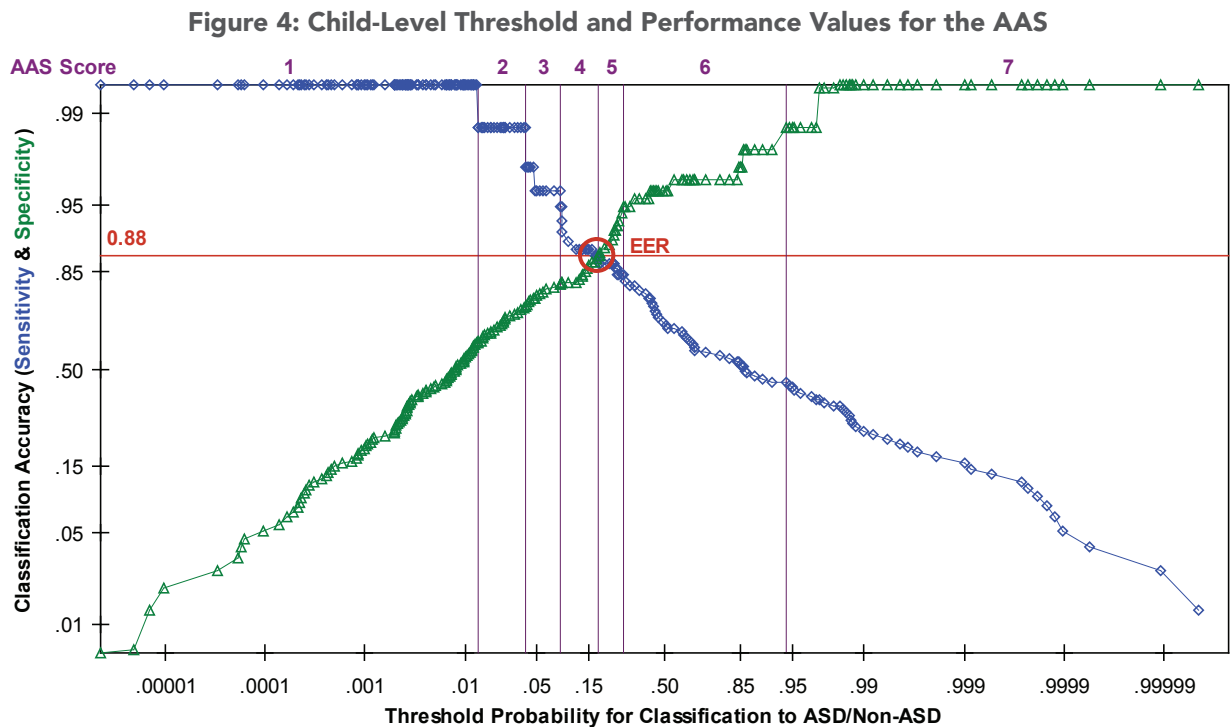
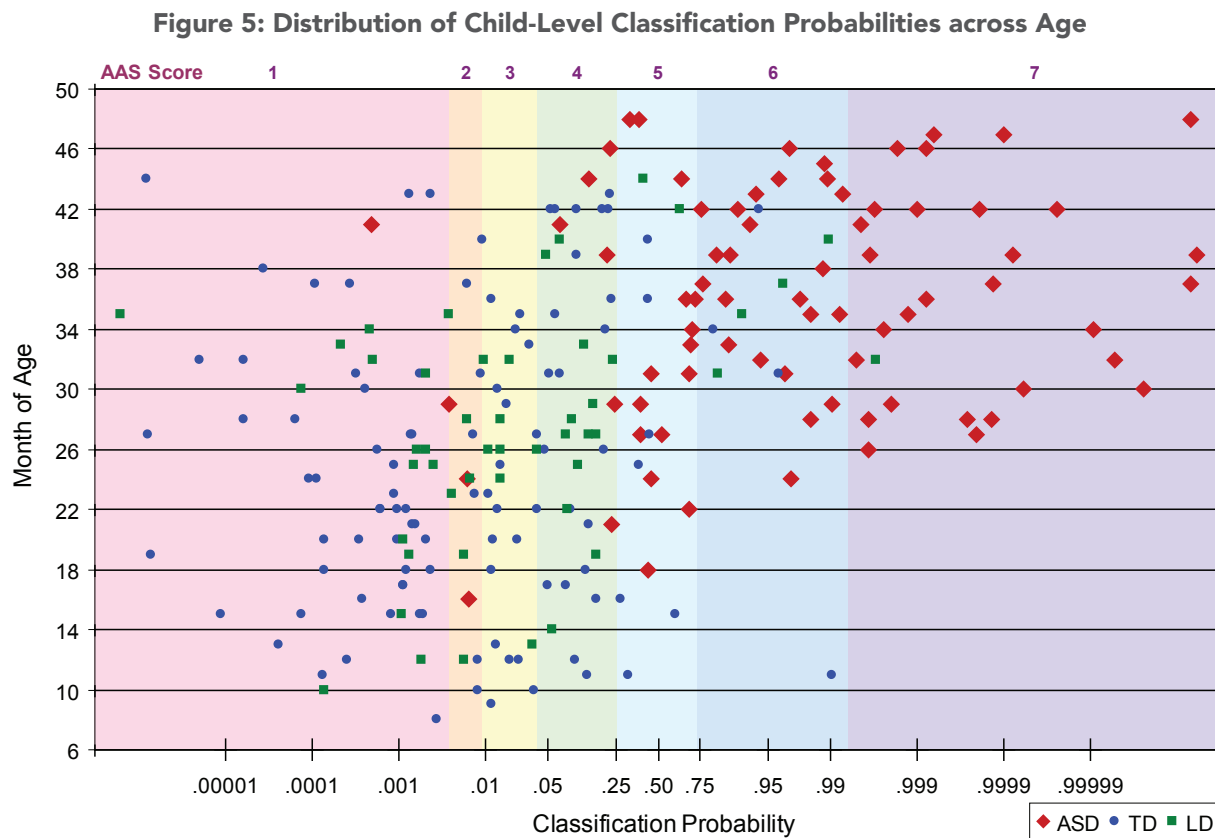


Figure 5 plots child-level classification probabilities for the three diagnostic groups (Samples 1 and 2 combined for each) by month of age. Color bands demark threshold boundaries for ordinal AAS scoring. Note that although the TD and LD samples are shown separately, they were combined in the analysis from which these data were obtained (i.e., the ASD vs. non-ASD classification matrix shown in Table 4a).



3.3 Correlations with ASD Symptom Assessments

The AAS score derives from a classification analysis based on acoustic feature sets, not on more traditionally assessed symptoms. Thus, it is not necessarily the case that higher posterior classification probabilities should indicate greater ASD symptom severity. Statistically, higher AAS scores need only reflect a closer fit to the acoustic pattern observed for the ASD sample. Conceivably, a behavior-based ASD symptom severity dimension could be orthogonal to a vocalization-based likelihood dimension.

In fact, correlational results are consistent with this hypothesis. Three parent self-report measures of ASD symptomatology were collected concurrently with the audio recordings for which AAS probabilities were obtained: the M-CHAT, the SCQ, and the CBCL.⁷ AAS scores did not correlate with scores on either the M-CHAT [$r(75) = -.08, p = .49$] or the SCQ [$r(73) = -.01, p = .92$]. AAS scores also did not correlate with the Total Score of the CBCL, which measures child competencies and behavioral problems [$r(74) = -.13, p = .27$], or with any of its subscales. Compared with traditional measures for which higher scores typically indicate increased symptomatology, the acoustic pattern-based DS approach does not itself yield a measure that indicates greater or lesser ASD-related symptom severity.

3.4 Test-Retest Reliability

To examine consistency in AAS scores across recordings within-child, the dataset was restricted to the first three recordings for each child, and 11 participants with fewer than three recordings were dropped from further analysis.⁸ The maximum difference between the three recording scores was computed. Across three recordings AAS ordinal scores differed by an average of 1.5 points ($SD = 1.2$) on the 7-point scale. The Typically Developing sample ($M = 1.6, SD = 1.2$) did not significantly differ from the Language Delay sample ($M = 1.8, SD = 1.0$). However, the ASD sample ($M = 1.2, SD = 1.3$) demonstrated a significantly smaller (by approximately one half point) average maximum difference than the non-ASD samples [$t(218) = 2.86, p = .005$]. That is, the distribution of acoustic features from ASD participants were the most consistent across recording days.

4.0 DISCUSSION

Performance results presented here indicate that the automated classification of vocalization data from children diagnosed with ASD may be reliably accomplished. Children with ASD were distinguished consistently from typically developing children and from children diagnosed with language disorders but without ASD. Automatically identifiable information in the vocalizations of children in our samples could be used to identify those at risk for ASD with close to 90% sensitivity and specificity. Thus, the automatic identification approach

7 M-CHAT is the Modified Checklist for Autism in Toddlers (Robins, Fein & Barton, 1999). SCQ is the Social Communication Questionnaire (Rutter, Bailey & Lord, 2003). CBCL is the Child Behavior Checklist (Achenbach & Rescorla, 2000).

8 Although Sample 1 participants across groups contributed more recordings, Sample 2 participants in each group recorded a maximum of three days. To simplify interpretation of results, all participants were limited to three recording days.

allows us to utilize machine capabilities to accomplish what a human cannot do efficiently — collect and sift through hours of audio data looking across multiple dimensions for subtle acoustic information.

The acoustic-based information related to the discrimination of children with ASD from typically developing children and children diagnosed with language disorders may relate both to differences in language development and to restricted and repetitive vocalization behavior in some children, as well as to differences in vocalization motor behavior. Furthermore, naturalistic audio recordings are likely to include social interaction and emotion-sharing data that may contribute to successful discrimination in ways yet to be explored.

Performance for this data-driven DS approach on some comparisons surpasses those previously reported for the thematically similar but statistically more limited 12P approach that utilized acoustic features derived from a language theory perspective (Oller et al., 2010). The 12P approach established the feasibility of the methodology; the current DS approach refines and improves on the method.

The potential benefits of this or other automated approaches to larger-scale or even universal screening of children for ASD are significant. Using a relatively simple, unobtrusive and low-involvement design, audio recording data may be obtained from a child's daily language and sound environment in quantities sufficient to accomplish screening. The automated analysis method provides the opportunity to collect this data from many families simultaneously and to obtain results quickly. By incorporating such an approach into a services triage system, human labor-intensive assessment and intervention could then be directed more efficiently toward children identified as being at higher risk for ASD or more likely to benefit from immediate intervention. Also, the automated nature of the screening provides a measure of objectivity that is difficult to obtain from human assessments. A fully objective child measure offers potential for use in, for example, clinical tracking and assessing parental adherence to intervention protocols.

As noted previously, early identification of children at risk for ASD is considered to be a beneficial, if not crucial, component to successful long-term outcomes. Research suggests that identification may be achieved reliably by 24 months or earlier; however, the necessity of advanced clinical training for professionals to make valid diagnoses combined with

typically limited resources results in the average age of assessment and diagnosis of ASD being closer to 60 months (AAP, 2006). The automated approach described in this report has the potential to reduce this intervention lag. Speculating further, although most children in the present study were older than 24 months, a lower bound is not yet known for the age at which discriminative acoustic features may reliably be identified in a child's vocalizations. Perhaps such features develop in conjunction with expressive language production or appear as precursors; if so, the identification of at-risk children could be made earlier. Additional research can be directed to explore this possibility.

As promising as these initial results appear to be, there remain significant limitations to the present study. Although portions of the TD and LD samples (Sample 1 in each case) were evaluated on site by certified SLPs to establish each child's language and developmental status, none of the ASD sample was evaluated on site by a trained and certified professional using standard tools (e.g., the ADOS [Lord, Rutter, DiLavore & Risi, 1999]). We required that parents provide documentation of diagnosis by a professional, and our participant data are consistent on symptom measures with published results from well-established samples, but it would be preferable to replicate these results using a novel sample for whom the diagnostic status has been more clearly established.

It is likely that the participating ASD sample is not sufficiently representative of the full distribution of children with ASD; for example, parental effects related to self-selection biases cannot be ruled out. In addition, the ASD participant samples were limited regarding demographic factors such as age, gender, and socioeconomic status. The catchall approach used for this first effort effectively limited our ability to examine diagnostic subgroups of ASD; however, such could be done with larger samples.

Moreover, current performance results are limited in their interpretability by the very nature of the automated and data-driven approach that has been applied. Again, we may speculate that those acoustic characteristics that account for the successful discrimination of children with ASD from other children are likely to correspond directly or indirectly to known and already identified features, such as those explored previously and described in Oller et al. (2010). However, additional work remains before such correspondences could be established to provide a more clear theoretical basis for the level of diagnostic group discrimination we have described in this report.

Despite such limitations, these results suggest important directions for future research. Foremost is the need to collect recording data from a well-established sample of children diagnosed with ASD using standard assessments. Prospective longitudinal studies of children or infants at high risk (e.g., younger siblings of children already diagnosed with ASD) would be particularly instructive to identify potential acoustic indicators of autism present in very early vocalization and language behavior.

Other applications of the current system may prove to be beneficial. As mentioned above, the approach could be used to help develop protocols for fast-track screening procedures to aid clinical service providers. Additional research could be undertaken to explore the system's applicability to non-English languages and to adapt it for other cross-linguistic applications when indicated. Ultimately, to clarify and explain more fully the classification accuracy of the system, a detailed analysis of the specific spectral properties most indicative of ASD could be conducted.

5.0 SUMMARY

Although currently evaluations and diagnostic criteria for ASD do not incorporate the identification of atypical vocalizations, prior research suggests that such features may indeed be present and detectable in children with ASD. Major obstacles to the clinical utilization of acoustic features have included the difficulty of collecting adequately representative samples of child vocalizations and the correspondingly modest research efforts into their usability, as well as there being few demonstrations of their potential for informing the ASD diagnosis. Presented here is an approach that takes advantage of new technology that eliminates the sampling obstacle and a statistical methodology that strongly supports the viability of using vocalization-based acoustic information to reliably classify and distinguish children with ASD from typically developing children and from other children with language-related disorders. Though future research efforts may clarify and extend this work, the current results constitute clear evidence of the soundness of the automated approach described.

REFERENCES

- Achenbach, T. M., & Rescorla, L. A. (2000). *Manual for the ASEBA preschool forms & profiles*. Burlington, VT: University of Vermont, Research Center for Children, Youth, & Families.
- American Academy of Pediatrics: Council on Children With Disabilities, Section on Developmental Behavioral Pediatrics, Bright Futures Steering Committee, and Medical Home Initiatives for Children With Special Needs Project Advisory Committee. (2006). Identifying infants and young children with developmental disorders in the medical home: an algorithm for developmental surveillance and screening. *Pediatrics*, 118, 405-420.
- American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (Revised 4th ed.). Washington, DC: Author.
- Ford, M., Baer, C.T., Xu, D., Yapanel, U., & Gray, S. (2008). *The LENA™ language environment analysis system: Audio specifications of the DLP-0121* (Technical report LTR-03-2). Boulder, CO: LENA Foundation. Retrieved August 31, 2010 from http://www.lenafoundation.org/TechReport.aspx/Audio_Specifications/LTR-03-2
- Gilkerson, J. & Richards, J.A. (2008). *The Infoture Natural Language Study* (Technical report LTR-02-2). Boulder, CO: LENA Foundation. Retrieved August 31, 2010 from: http://www.lenafoundation.org/TechReport.aspx/Natural_language_Study/LTR-02-2
- Lord, C., Rutter, M., DiLavore, P. C. & Risi, S. (2002) *Autism Diagnostic Observation Schedule*. Los Angeles, CA: Western Psychological Services.
- Oller, D.K., Niyogi, P., Gray, S., Richards, J.A., Gilkerson, J., Xu, D., Yapanel, U., & 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*, 107(30), 13354-13359.
- Richards, J.A., Gilkerson, J. Paul, T. & Xu, D. (2008). *The LENA™ automatic vocalization assessment* (Technical report LTR-08-1). Boulder, CO: LENA Foundation. Retrieved August 31, 2010 from: <http://www.lenafoundation.org/TechReport.aspx/AVA/LTR-08-1>
- Robins, D., Fein, D., & Barton, M. (1999). *Modified Checklist for Autism in Toddlers*. Atlanta, GA: Author. Retrieved August 31, 2010 from: http://www2.gsu.edu/~psydlr/Diana_L._Robins,_Ph.D._files/M-CHAT_new.pdf

- Rutter, M., Bailey, A., & Lord, C. (2003). *The social communication questionnaire*. Los Angeles, CA: Western Psychological Services.
- Sikora, D.M., Hall, T.A., Hartley, S.L., Gerrard-Morris, A.E. & Cagle, S. (2008). Does parent report of behavior differ across ADOS-G classifications: Analysis of scores from the CBCL and GARS. *Journal of Autism and Developmental Disorders*, 38, 440-448.
- Watt, N., Wetherby, A.M. & Barber, A. (2008). Repetitive and stereotyped behaviors in children with autism spectrum disorders in the second year of life. *Journal of Autism and Developmental Disorders*, 38, 1518-1533.
- Wetherby, A.M., & Prizant, B.M. (2002). *Communication and symbolic behavior scales developmental profile*. Baltimore, MD: Brookes.
- Xu, D., Richards, J.A., Gilkerson, J., Yapanel, U., Gray, S., & Hansen, J. (2009, November). *Automatic childhood autism detection by vocalization decomposition with phone-like units*. Paper presented at the 2nd Workshop on Child, Computer and Interaction, Cambridge, Massachusetts.
- Xu, D., Yapanel, U., & Gray, S. (2008). *Reliability of the LENA™ language environment analysis system in young children's natural language home environment* (Technical report LTR-05-2). Boulder, CO: LENA Foundation. Retrieved August 31, 2010 from: <http://www.lenafoundation.org/TechReport.aspx/Reliability/LTR-05-2>

APPENDIX A: SELECTION CRITERIA FOR DIAGNOSTIC SAMPLES

Fixed Criteria – All Participants

1. English is the primary language in the home
2. Child's age is \leq 48 months

Additional Criteria by Group – All Participants

1. TD Samples 1 & 2:
 - a. No indication of developmental disorder
2. LD Samples 1 & 2:
 - a. Parent report of a diagnosis of language delay by a qualified professional
 - b. No indication of autism in prior diagnosis
3. ASD Samples 1 & 2:
 - a. Parental report that a diagnosis of ASD had been given by a qualified professional
 - b. Written diagnostic documentation supplied by parents to our staff from the professional(s) who had evaluated the child
 - c. Asperger syndrome excluded

Additional Criteria by Group – Specific to Sample 2 Participants

4. TD Sample 2:
 - a. Age 18-36 months
 - b. M-CHAT passed on both scoring options (no ASD)
 - c. LENA Developmental Snapshot Standard Score Between 80 – 110 (midrange language levels)
 - d. No sibling with developmental delay diagnosis
 - e. No sibling with autism
 - f. No other symptoms of autism on intake questionnaire (e.g., frequently repeated motions, lack of eye contact)

APPENDIX A: SELECTION CRITERIA FOR DIAGNOSTIC SAMPLES (CONT.)

5. LD Sample 2:
 - a. M-CHAT passed on both scoring options (no ASD)
 - b. Written diagnostic documentation supplied by parents to our staff from the professional(s) who had given the diagnosis of language delay to the child
 - c. LENA Developmental Snapshot Standard Score ≥ 1.5 SD below the mean (low language level)
 - d. No sibling with autism
 - e. No other symptoms of autism on intake questionnaire (frequently repeated motions, lack of eye contact)

6. ASD Sample 2:
 - a. M-CHAT failed on at least one of two scoring options

APPENDIX B: DEMOGRAPHIC INFORMATION

Table B1: Maternal Attained Education Level and Gender Distribution of Sample Participants

Mother's education	TD Sample			LD Sample			ASD Sample			All Samples		
	F	M	Total	F	M	Total	F	M	Total	F	M	Total
Some high school	8	4	12	3	0	3	0	2	2	11	6	17
GED	1	3	4	0	0	0	0	2	2	1	5	6
Trade school	2	0	2	1	0	1	1	3	4	4	3	7
High school diploma	6	7	13	1	7	8	1	7	8	8	21	29
Some college	19	14	33	5	12	17	1	11	12	25	37	62
Associate's degree	5	1	6	1	3	4	0	1	1	6	5	11
Bachelor's degree	14	11	25	1	7	8	5	24	29	20	42	62
Graduate degree	3	8	11	4	4	8	5	14	19	12	26	38
Total	58	48	106	16	33	49	13	64	77	87	145	232

APPENDIX C: LANGUAGE ASSESSMENT SUMMARY

Table C1 details parent self-report and SLP-administered language assessments that were completed by sample participants.

Table C1: Language Assessment Completion by Diagnostic Sample

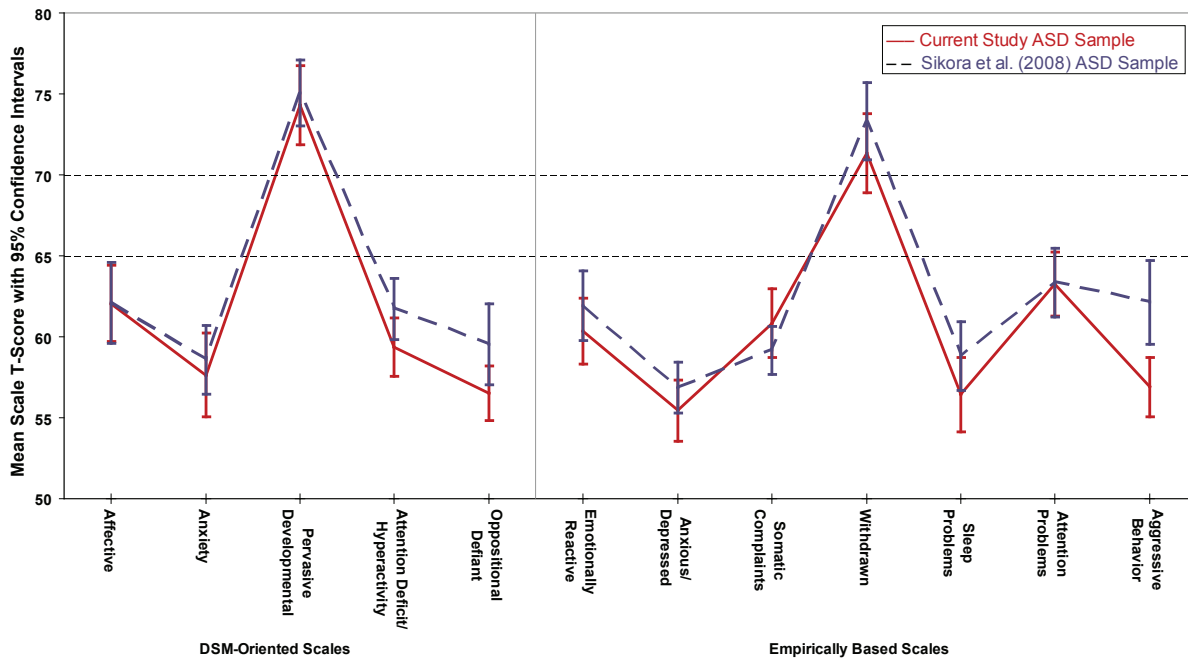
Parent Questionnaire	Typically Developing Samples	Language Delay Samples	ASD Samples
Brief-P	1		1
CDI	1,2	1,2	1,2
CBCL/LDS	1,2	2	1,2
CSBS-CQ	2	2	1,2
M-Chat	2	2	1,2
MacArthur	1,2	1,2	1,2
SCQ	2	2	1,2
Snapshot	1,2	1,2	1,2

SLP Assessment	Typically Developing Samples	Language Delay Samples	ASD Samples
CAT/CLAMS	1	1	
GFTA	1	1	
PLS-4	1	1	
PPVT	1	1	
REEL-3	1	1	

APPENDIX D: DIAGNOSTIC SAMPLE COMPARISONS

Figure D1 demonstrates the consistency across average CBCL subscale scores of current ASD participants (combined Samples 1 and 2) with scores reported previously for an independently collected and published ASD sample (Sikora et al., 2008).

Figure D1: CBCL Subscale Scores for ASD Samples

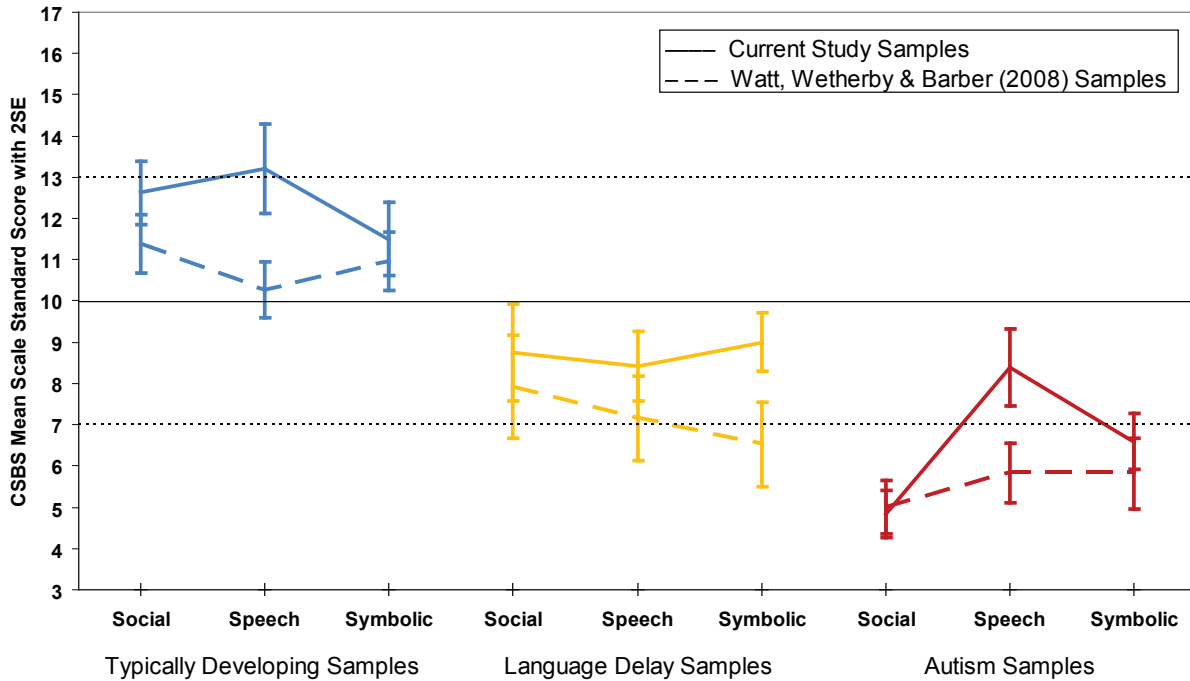


Adapted from Oller et al. (2010)

APPENDIX D: DIAGNOSTIC SAMPLE COMPARISONS (CONT.)

Figure D2 compares current participants (ASD: Combined Samples; TD and LD: Sample 2) to an independently collected and reported sample of children on CSBS subscale scores (Watt, Wetherby & Barber, 2008).

Figure D2: CSBS Subscale Scores for ASD Samples



Adapted from Oller et al. (2010)

APPENDIX E: ADDITIONAL CHILD-LEVEL CLASSIFICATION PERFORMANCE SUMMARIES

Table E1: Typically Developing vs. Other (Language Delay & ASD)

E1a: All participants

		AAS Classification		
		TD	Other	Total
Criterion Classification	TD	85	21	106
	Other	24	102	126
Total		109	123	232

Sensitivity:	.80
Specificity:	.80
Positive P.P.:	.78
Negative P.P.:	.83
Overall Accuracy:	.81
Cohen's Kappa:	.61

E1b: Participants 24–48 months of age

		AAS Classification		
		TD	Other	Total
Criterion Classification	TD	66	15	81
	Other	21	94	115
Total		87	109	196

Sensitivity:	.82
Specificity:	.82
Positive P.P.:	.76
Negative P.P.:	.86
Overall Accuracy:	.82
Cohen's Kappa:	.63

APPENDIX E: ADDITIONAL CHILD-LEVEL CLASSIFICATION PERFORMANCE SUMMARIES (CONT.)

Table E2: Typically Developing vs. Language Delay

E2a: All participants

		AAS Classification		Total
		TD	LD	
Criterion Classification	TD	78	28	106
	LD	13	36	49
Total		91	64	155

Sensitivity:	.74
Specificity:	.74
Positive P.P.:	.86
Negative P.P.:	.56
Overall Accuracy:	.74
Cohen's Kappa:	.44

E2b: Participants 24–48 months of age

		AAS Classification		Total
		TD	LD	
Criterion Classification	TD	63	18	81
	LD	9	33	42
Total		72	51	123

Sensitivity:	.79
Specificity:	.79
Positive P.P.:	.88
Negative P.P.:	.65
Overall Accuracy:	.78
Cohen's Kappa:	.54

APPENDIX E: ADDITIONAL CHILD-LEVEL CLASSIFICATION PERFORMANCE SUMMARIES (CONT.)

Table E3: Language Delay vs. Other (Typically Developing & ASD)

E3a: All participants

		AAS Classification			Total		
		LD	Other				
Criterion Classification	LD	33	16		49	Sensitivity:	.68
	Other	58	125		183	Specificity:	.68
	Total	91	141		232	Positive P.P.:	.36
						Negative P.P.:	.89
						Overall Accuracy:	.68
						Cohen's Kappa:	.27

E3b: Participants 24–48 months of age

		AAS Classification			Total		
		LD	Other				
Criterion Classification	LD	30	12		42	Sensitivity:	.71
	Other	44	110		154	Specificity:	.71
	Total	74	122		196	Positive P.P.:	.41
						Negative P.P.:	.90
						Overall Accuracy:	.71
						Cohen's Kappa:	.34

APPENDIX F: RECORDING-LEVEL PERFORMANCE

Classification performance data presented in the main text are reported at the child-level. Results from all recordings available for each child are combined as the geometric average to produce one score per child. Model training also may be conducted and performance assessed at the recording-level, ignoring non-independence effects. Table F1 and Figure F1 show classification performance for the full set of 1486 recordings, conducted at the recording-level.

Table F1: Recording-Level Criterion Values for AAS Thresholds

AAS Ordinal Score	Criterion Index	Lower Threshold Probability Value	Lower Threshold Criterion Value
1		0	1
2	Sensitivity	.004	.99
3		.009	.97
4		.039	.95
5	EER Point	.251	.87
6	Specificity	.738	.95
7		.994	.99

Figure F1: Recording-Level Threshold and Performance Values for the AAS

