The Relationship Between Standardized Measures of Language and Measures of Spontaneous Speech in Children With Autism

Karen Condouris
Boston University School of Medicine

Echo Meyer
Division TEACH, University of North Carolina

Helen Tager-Flusberg
Boston University School of Medicine

This study investigated the relationship between scores on standardized tests (Clinical Evaluation of Language Fundamentals [CELF], Peabody Picture Vocabulary Test–Third Edition [PPVT-III], and Expressive Vocabulary Test) and measures of spontaneous speech (mean length of utterance [MLU], Index of Productive Syntax, and number of different word roots [NDWR]) derived from natural language samples obtained from 44 children with autism between the ages of 4 and 14 years old. The children with autism were impaired across both groups of measures. The two groups of measures were significantly correlated, and specific relationships were found between lexical–semantic measures (NDWR, vocabulary tests, and the CELF lexical–semantic subtests) and grammatical measures (MLU, and CELF grammar subtests), suggesting that both standardized and spontaneous speech measures tap the same underlying linguistic abilities in children with autism. These findings have important implications for clinicians and researchers who depend on these types of language measures for diagnostic purposes, assessment, and investigations of language impairments in autism.

Key Words: autism, spontaneous speech measures, standardized language tests, MLU, NDWR

Researchers and clinicians regularly rely on a variety of measures of language to assess and chart developmental changes in language in a variety of populations. Clinicians typically depend on measures of language to diagnose children with language impairments, to assess a range of language skills, and to design and monitor treatment programs. Researchers use language measures to define their participant populations, to document their participants' language status, to match groups of participants, or to investigate aspects of language impairment in different populations. Typically, two classes of measures are used for these purposes: (a) standardized psychometric tests and (b) measures of spontaneous speech derived from natural language samples, which can be collected in a variety of ways in different contexts.

Both types of measures may be used to assess a range of language skills, including phonology, lexical knowledge, semantics, morphosyntax, and pragmatics, in children at different ages. Generally, standardized language measures assess both receptive and expressive abilities, whereas measures of spontaneous speech are used to tap expressive language. Psychometric tests are norm-referenced, and when administered according to the standardized procedures defined for them, they provide a relatively quick means for comparing a child to age-matched peers. When tests have been normed on similar samples, they also allow one to compare a child's performance across different tests to yield a profile of language performance across language domains. In contrast, measures derived from natural language samples require a significant investment of time. These measures provide an index of the child's use of language in everyday informal settings and are especially useful for assessing a variety of pragmatic and discourse skills.

The focus of this study is on children with autism, a disorder characterized by delays and deficits in language.

The primary type of language impairment that defines autism is in the area of pragmatics, including limited uses of language and deficits in discourse (see Lord & Paul, 1997; Tager-Flusberg, 2000b; and Wilkinson, 1998, for recent reviews). In addition, a significant proportion of children with autism also have impairments in other aspects of language, including lexical–semantic and grammatical development (Kjelgaard & Tager-Flusberg, 2001). Given the broad range of impairments found in autism, both clinicians and researchers need to depend on measures that span both standardized tests and natural language samples, but we do not know how measures derived from these different types of assessment compare in this population. This study begins to address this important clinical and research issue in a group of relatively high functioning verbal children with autism.

There are advantages and limitations in the use of standardized and spontaneous speech measures. The administration of standardized tests provides a portrait of a child's language abilities across a prespecified set of language skills. However, in this structured context, factors such as children's test-taking skills, attention, or motivation to interact with the examiner may also contribute to language scores. Measures from natural language samples offer an assessment of a child's real-time language performance. Such measures thus reveal the influence of the dynamic interaction among a child's individual linguistic knowledge, internal processing factors, and external processing constraints on verbal performance (Evans, 1996).

A small number of studies have compared these two methods of measuring language. Bornstein and Haynes (1998) examined the relationship between measures derived from standardized assessments and measures of spontaneous speech through their investigation of the relationships among different language measures taken from 184 normally developing 20-month-old children. They compared the expressive and receptive scales from the Reynell Developmental Language Scales to mean length of utterance (MLU) and the number of different word roots (NDWR) derived from mother–child free play sessions. They found that all these measures correlated significantly with one another, suggesting that both standardized and spontaneous speech measures tap the same language competence in normally developing toddlers.

A second study compared standardized vocabulary test scores to spontaneous speech measures in 28 normally developing preschoolers (Ukrainetz & Blomquist, 2002). Specifically, Ukrainetz and Blomquist investigated the relationship between four vocabulary tests: the Peabody Picture Vocabulary Test–Third Edition (PPVT-III), the Expressive Vocabulary Test (EVT), the Receptive One Word Picture Vocabulary Test, and the Expressive One Word Picture Vocabulary Test–Revised, and the following measures derived from spontaneous speech samples: MLU, NDWR, and the total number of words (TNW). They found significant intercorrelations between the vocabulary tests and NDWR. Weaker correlations were obtained between the vocabulary tests and MLU and TNW, which are assumed to be less specifically related to vocabulary knowledge. These findings were taken as empirical support that the four vocabulary tests and NDWR were measuring the same construct.

The diagnosis of disorders such as specific language impairment (SLI) is usually made on the basis of standardized test measures, with researchers or clinicians identifying some arbitrary cut-off (e.g., more than 1.25 *SD*s below the mean) for defining language impairment. A study by M. Dunn, Flax, Sliwinski, and Aram (1996) compared the sensitivity of standardized test measures to measures derived from natural language samples for diagnosing SLI. They found that measures from natural language samples, specifically MLU and percentage of utterances containing structural errors, were better at defining SLI than were the psychometric tests that had been given to the children in their study. Rescorla, Roberts, and Dahlsgaard (1997) conducted a follow-up study of toddlers identified as late talkers. Although there was considerable variability in both the standardized test measures and the measures from natural language samples, MLU and the Index of Productive Syntax (IPSyn; Scarborough, 1990) were more sensitive to language delays at follow-up than any of the standardized tests. Similarly, Goffman and Leonard (2000) recommended using measures from natural language samples to assess language growth in children with SLI. Other studies have also highlighted the usefulness of spontaneous speech measures for diagnosing children with SLI who come from Spanish-speaking families (Restropo, 1998) or African-American families (Craig & Washington, 2000). Botting, Conti-Ramsden, and Crutchley (1997) investigated the sensitivity of standardized psychometric tests to different types of language disorders in a sample of more than 240 children ages 6–8 years old. They found that although such tests were good at discriminating children with structural language impairments, none of the tests could identify children with semantic–pragmatic disorders. They concluded that psychometric measures cannot be used for diagnosing these kinds of language impairments, which are prevalent in children with autism spectrum disorders.

Thus far, no one has compared the use of standardized test and spontaneous speech measures of language in children with autism, though both types of measures are widely used in research and clinical practice. There are significant challenges in assessing the language of children with autism (Sparrow, 1997; Tager-Flusberg, 2000a). Because of the core social deficits in autism and high rates of echolalia, found especially in younger children, it may be difficult for them to provide an adequate natural language sample in the context of a conversational interaction. On the other hand, perhaps the unique behavior, motivation, and attentional problems found in many children with autism interfere with the demands of the formal testing situation required for standardized tests. Some researchers have questioned whether standardized tests can be used to describe language functioning in children with autism (e.g., Koegel, Koegel, & Smith, 1997), and others have suggested that the highly structured testing situation in fact enhances the performance of children with autism, whose rigid behavioral styles might be well suited to standardized test assessments (Paul & Cohen, 1995).

The main goal of this study was to investigate the relationship between standardized and spontaneous speech measures of language in children with autism. Specifically, we explored the children's performance across both types of measures in comparison to normative data on the same measures and used correlational and regression analyses to assess the relationships between the two types of measures in this population. We focused on measures of lexical–semantics and morphosyntax because they can readily be assessed in both standardized testing and natural language samples, thus allowing us to examine general and specific relationships between these methods of measuring language skills in children with autism.

Method

Participants

A group of 44 children with autism, including 7 girls and 37 boys ages 4 to 14 years old, participated in this cross-sectional study. The participants were part of a larger project investigating language functioning in autism that was previously described by Kjelgaard and Tager-Flusberg (2001). The children for this study were selected from the larger cohort of children with autism on the basis of their ability to complete all the language testing within age level. Diagnosis of autism was made on the initial visit, based on the Autism Diagnostic Interview–Revised (Lord, Rutter, & Le Couteur, 1994), and the Autism Diagnostic Observation Schedule–Generic (ADOS; Lord et al., 2000), administered by trained examiners. The children's scores on these instruments are presented in Table 1, together with other descriptive statistics for the sample. An expert clinician confirmed that all the children met *Diagnostic and Statistical Manual for Mental Disorders–Fourth Edition* (American Psychiatric Association, 1994) criteria for autism, including qualitative impairments in social functioning and communication and repetitive behaviors and interests. IQ scores were assessed using the Differential Ability Scales (DAS; Elliot, 1990). Children were administered either the Preschool or School-Age version of the DAS depending on their age and ability level. The DAS yielded full scale IQ, verbal IQ, and nonverbal IQ subscores. The socioeconomic status (SES) of the sample was estimated using maternal educational level, which is the most significant SES predictor of language functioning in children (Dollaghan et al., 1999). We obtained this information from 40 of the 44 mothers, among whom 20% (*n* = 8) had 12–15 years of education and 80% (*n* = 32) had 16 or more years of education.

TABLE 1. Participant characteristics (*N* = 44).

	M	*SD*	Range
Chronological age (years;months)	7;3	2;4	4–14;2
Full scale IQ	85.3	19.0	52–141
Verbal IQ	83.7	19.2	53–133
Nonverbal IQ	90.0	20.7	49–153
ADI-R social domain	21.3	5.12	10–28
ADI-R communication domain	17.5	3.67	9–24
ADI-R repetitive behavior domain	6.32	2.55	2–12
ADOS social score	9.2	2.1	4–12
ADOS communication score	5.5	2.1	1–9

Note. ADI-R = Autism Diagnostic Interview–Revised; ADOS = Autism Diagnostic Observation Schedule.

Standardized Language Test Measures

The PPVT-III (L. M. Dunn & Dunn, l997) and EVT (Williams, l997) were administered to assess lexical knowledge and word retrieval. We chose these two vocabulary tests because they (or their British equivalent) are widely used language tests in published studies on children with autism and were standardized on the same large sample that included children with a broad range of abilities levels. The Clinical Evaluation of Language Fundamentals–Preschool (CELF-P, Wiig, Secord, & Semel, l992) or CELF-3 (Semel, Wiig, & Secord, l995) was included as our omnibus measure of higher order lexical–semantic, grammatical, and verbal memory abilities. The CELF was selected because it has excellent psychometric properties, it covered the age range of children in our study, and based on our clinical experience, its inclusion of engaging visual stimuli helped to focus and maintain the attention of children with autism. Moreover, we could use the CELF both as an omnibus language measure and for more specific measures of lexical–semantic and grammatical skills, using select subtests that tap these language domains, based on information in the test manuals.

PPVT-III. The PPVT-III is a standardized test of receptive lexical knowledge. The child must select the line drawing from four choices on a page that matches a word spoken by the examiner.

EVT. The EVT measures lexical knowledge and word retrieval by asking the child either to name pictures, or, as the test gets more advanced, to provide a synonym for the spoken and pictured target word.

CELF-P and CELF-3. The CELF-P and CELF-3 are omnibus tests of language ability. Each includes six subtests, three in the receptive domain and three in the expressive domain. Individual subtest scores are calculated as standard scores (*M* = 10, *SD* = 3), and the Receptive and Expressive composite scores and Total Language standard scores are calculated as standard scores (*M* = 100, *SD* = 15). The CELF-P is suitable for children between the ages of 3;0 (years;months) and 6;11 and the CELF-3 covers the age range of 6;0 to 21;11. Table 2 lists the main subtests on the CELF-P and CELF-3. Across the broad age range from 3 to 21, the same tests cannot be used to assess language performance; nevertheless, there are some parallel subtests included in the two versions of the CELF. We used information provided in the test manuals to select subtests that specifically tapped lexical–semantic and grammatical domains of language, rather than more integrated language

TABLE 2. Summary of subtests on the Clinical Evaluation of Language Fundamentals–Preschool (CELF-P) and Third Edition (CELF-3).

CELF-P	CELF-3	
Ages 3–6;11	Ages 6–8	Ages 9+
	Receptive subtests	
Linguistic Concepts[a]	Concepts and Directions[a]	(same)
Basic Concepts	Word Classes	(same)
Sentence Structure[b]	Sentence Structure[b]	—
		Semantic Relationships
	Expressive subtests	
Formulating Labels	Formulating Sentences	(same)
Word Structure[b]	Word Structure[b]	—
		Sentence Assembly
Recalling Sentences	Recalling Sentences	(same)

[a] Tests tapping lexical–semantic skills. [b] Tests tapping grammatical skills.

abilities or verbal memory. We selected subtests that were significantly correlated across both the CELF-P and the CELF-3 (Semel et al., 1995) as our measures of lexical–semantic and grammatical ability on these tests. Thus, Linguistic Concepts (CELF-P) and Concepts and Directions (CELF-3) tap lexical–semantic skills and correlate significantly (r = .49). Both Sentence Structure (r = .36 between CEL-P and CELF-3) and Word Structure (r = .41 between CELF-P and CELF-3) measure morphology and syntax.

Testing Procedures

A certified speech-language pathologist administered the standardized language measures. Administration was typically conducted over two 60-min sessions scheduled on different days within a 1-month period. Breaks were provided when needed, following the testing procedures outlined for each test. The examiners actively worked at ensuring that the children were always engaged in the test and attending to the stimuli. When needed, reinforcers such as stickers or stars were used to maintain the child's motivation. Typically, the examiner followed the children's lead during the assessment in order to maximize their performance (cf. Koegel et al., 1997). In most cases, tests were given in the following order: PPVT-III, EVT, and CELF. All the children were tested within age level on the CELF. The tests were scored by a certified speech-language pathologist and then checked by a trained coder.

Natural Language Samples

Natural language samples were collected from the children while they interacted with one of their parents (almost always the mother) for 30 min in the laboratory. The children and parents were provided with a standard set AQ1 of toys. For the 4- to 10-year-olds, the set included Play-Doh with cookie cutters; figurines with bedroom, kitchen, and dentist office furnishings; blocks; colored markers and paper; a train set; puppets; Uniset Picture Making booklets; and a barrel filled with monkeys, dinosaurs, and vehicles. The 11- to 14-year-olds were provided with action figures, battling robots, a magic set, playing cards, dominoes, and a domino-rally racing game. Participants were asked to play and interact with each other as they would at home. The session was video- and audio-recorded.

Transcription. The language samples were transcribed using the Systematic Analysis of Language Transcripts (SALT) transcription format (Miller & Chapman, 2000) by a team of research assistants trained in transcription procedures. Utterance segmentation and the identification of bound morphemes were based on the guidelines specified by Miller and Chapman. Transcripts were prepared by one person and checked by a second trained transcriber using both the audio- and video-recordings. All transcription disagreements were resolved through consensus.

After omitting the first 10 child utterances from the transcript, a corpus of 100 consecutive, complete, and intelligible child utterances was selected. All full or partial imitations of adult utterances that were within five transcript lines of the child's utterance were excluded from analyses if the child failed to add to or modify at least one constituent of the adult's sentence. Verbatim songs or nursery rhymes were also excluded from analyses. Four participants produced fewer than 100 complete and intelligible utterances; for these transcripts, were calculated AQ2 following procedures outlined for reduced samples (Miller, 1991; Miller & Chapman, 2000; Scarborough, 1990).

Measures of Spontaneous Speech. MLU, measured in morphemes (Brown, 1973), the IPSyn (Scarborough, 1990), and the NDWR in a 50 utterance sample were selected as quantitative measures of the children's language. These three measures were selected because of their sensitivity in indicating developmental changes in children's language abilities and their wide use.

MLU. MLU is a measure of utterance length used as an index of children's grammatical complexity (Brown, 1973). In typically developing children, MLU correlates significantly with age up to approximately MLU 2.5–3.0 (Klee, 1992; Rondal, Ghiotto, Bredart, & Bachelet, 1987). With MLUs greater than 3.0, the association between age and MLU is less reliable; however, it continues to be a valid predictor of syntactic complexity and diversity up to approximately MLU 4.0 (Klee, 1992; Rollins, Snow, & Willett, 1996). MLU has been used as a diagnostic measure to differentiate between normally developing children and language impaired populations (e.g., Klee, Schaffer, May, Membrino, & Mougey, 1989; Rondal, Ghiotto, Bredart, & Brachelet, 1988; Scarborough, Rescorla, Tager-Flusberg, & Fowler, 1991; Scarborough, Wyckoff, & Davidson, 1986).

IPSyn. The IPSyn provides an alternative index of syntactic and morphological development. Unlike MLU, the IPSyn is a type-based measure. It assesses the child's emergent use of specified morphological and syntactic structures, rather than their use in obligatory contexts (Scarborough, 1990). There are 59 items on the IPSyn, each worth a maximum of 2 points, resulting in a Total IPSyn score ceiling of 118. On each item, a child can earn 0 points (*no exemplar produced*), 1 point (*1 exemplar produced*), or 2 points (*2 different exemplars produced*). The index includes four subscales: Noun Phrase, Verb Phrase, Questions and Negations, and Sentence Structure. Scoring is based on the analysis of a 100-utterance corpus. For corpora with fewer than 100 utterances, a conversion table of estimated IPSyn scores is provided (Scarborough, 1990).

A trained coder scored all the transcripts for the IPSyn measure from the same 100-utterance corpora that were used for the MLU analyses. Two scorers independently calculated IPSyn scores for 25% of the language samples. Interrater reliability was calculated using the Pearson correlation statistic: $r(10) = .99$.

NDWR. NDWR is a measure of lexical diversity, for which developmental and diagnostic validity and temporal reliability have been demonstrated (Klee, 1992; Miller, 1991). NDWR is calculated on the number of different word roots (bare stem) found in language samples of a fixed length. This measure was calculated on the first 50 utterances of each corpus, which has been shown to be a reliable measure of lexical diversity (Watkins, Kelly, Harbers, & Hollis, 1995).

MLU and NDWR were computed using SALT-based analyses (Miller & Chapman, 2000). The SALT reference database Version 6.1 includes spontaneous speech data collected on a large sample of normally developing children that includes these measures. We selected for each participant in this study an age- (and transcript length) matched comparison sample of 15–35 children from the database to compare the scores obtained from the children with autism to this SALT normative sample.

Results

Children's Performance on Language Tests and Spontaneous Speech Measures

Of the 44 participants, 26 were tested on the CELF-P and 18 were tested on the CELF-3. Among the children tested on the CELF-3, 9 were given the tests for the oldest age range and thus did not receive either of the grammatical subtests, Sentence Structure and Word Structure. Table 3 presents the children's test scores and the spontaneous speech measures derived from the natural language samples. As shown in Table 3, mean standard scores on the standardized tests and subtests reflecting vocabulary and morphosyntax were generally more than 1 *SD* below the

TABLE 3. Children's scores on the standardized language tests and spontaneous speech measures.

	M	*SD*	Range	*N*
CELF subtests				
Linguistic concepts/directions	5.27	3.51	3–17	44
CELF lexical-semantic score	5.27	3.51	3–17	44
Sentence structure	5.03	2.80	3–12	35
Word structure	5.91	3.75	3–14	35
CELF grammar score	5.47	3.0	30–12	35
CELF receptive language score	70.98	20.4	50–116	44
CELF expressive language score	74.55	19.30	50–116	44
CELF total language score	72.30	18.93	50–113	44
PPVT-III	85.59	19.19	55–134	44
EVT	84.02	17.61	40–136	44
Vocabulary (combined PPVT+EVT)	84.80	16.90	50–136	44
Spontaneous speech measures				
MLU (in morphemes)	3.38[b]	0.89	1.63[c]–5.21	44
NDWR	71.66[b]	17.90	40[c]–108[a]	44
IPSyn	71.98	15.54	37–100	44

Note. CELF = Clinical Evaluation of Language Fundamentals; PPVT-III = Peabody Picture Vocabulary Test–Third Edition; EVT = Expressive Vocabulary Test; MLU = mean length of utterance; NDWR = number of different word roots; IPSyn = Index of Productive Syntax.

[a]Over 1 *SD* below the Systematic Analysis of Language Transcripts (SALT) age-referenced database mean. [b]Over 2 *SD* below the SALT age-referenced database mean. [c]Over 3 *SD* below the SALT age-referenced database mean.

mean; however, as expected, there was wide variability among the children. Mean scores on two of the spontaneous speech measures, NDWR and MLU, were 2 *SD*s below the SALT reference database mean, and the group mean for IPSyn scores was below the level expected for the children's age based on data presented by Scarborough (1990). Thus, as a group, vocabulary and morphosyntactic abilities were significantly below age-level expectations on both methods of assessment.

We compared scores on the standardized test and spontaneous speech measures at an individual participant level to investigate, at an exploratory level, how children performed relative to the mean. For the lexical–semantic measures, we included NDWR from the natural language sample and compared this measure to Linguistic Concepts/Concepts and Directions from the CELF and a combined vocabulary score from the PPVT-III and EVT. For the grammatical measures we included MLU and a combined score from Sentence Structure and Word Structure. Each child's spontaneous speech score was compared to the SALT reference database mean, following the same procedures used by Ukrainetz and Blomquist (2002). The data are presented in Table 4 in a series of contingency tables. Because too many of the cells in the tables were empty, it was not possible to conduct statistical analyses. Nevertheless, the overall trend shown in Table 4, illustrated by the numbers that are presented in bold, is that the spontaneous speech measures were more deviant from the mean than were the scores from the standardized tests. Thus, comparing performance on the vocabulary tests and NDWR (see the top section of Table 4), there were 33 children whose scores on NDWR were more than 1 *SD* lower than their scores on the vocabulary tests, compared to only 1 child whose combined vocabulary test score was significantly lower than the NDWR. The parallel comparison for the CELF lexical–semantic score and NDWR (see the middle section of Table 4) was 17 children with NDWR < CELF compared to 3 children with CELF < NDWR; in the comparison of MLU and CELF Grammar (see the bottom section of Table 4), 18 children had scores MLU < CELF and 2 children had scores CELF < MLU.

Relationships Between Standardized Test Scores and Spontaneous Speech Measures

The correlations among the measures that tap lexical–semantic and grammatical skills were calculated partialing out the effects of age and nonverbal IQ, and the data are presented in Table 5. The data presented here are from the 35 children who completed the grammatical subtests on the CELF. The measures include all of those derived from the natural language sample (MLU, IPSyn, and NDWR), the two vocabulary tests (PPVT-III and EVT), and the subtests from the CELF-P and CELF-3 that assess lexical–semantic and grammatical abilities (individually and combined). The total CELF score, as an omnibus language measure, and the combined vocabulary score (PPVT + EVT) are also included in the correlation matrix. Because of the relatively large number of correlations, only those reaching a more conservative *p* value of .01 were considered statistically significant.

All the correlations were positive, and most reached statistical significance. The correlations among the measures

TABLE 4. Comparison of performance on standardized spontaneous speech measures of lexical–semantic and grammatical skills.

AQ3

NDWR and Vocabulary Tests

Vocabulary (PPVT-III + EVT)	NDWR normal	NDWR < 1 *SD*	NDWR < 2 *SD*	Total
Vocabulary normal	0	**5**	**12**	17
Vocabulary < 1 *SD*	0	2	**16**	18
Vocabulary < 2 *SD*	0	1	8	9
Total	0	8	36	44

NDWR and CELF Lexical–Semantic

CELF lexical–semantic	NDWR normal	NDWR <1 *SD*	NDWR <2 *SD*	Total
Lexical–semantic normal	0	**4**	**8**	12
Lexical–semantic < 1 *SD*	0	1	**5**	6
Lexical–semantic < 2 *SD*	0	3	23	26
Total	0	8	36	44

MLU and CELF Grammar

CELF grammar	MLU normal	MLU < 1 *SD*	MLU < 2 *SD*	Total
CELF grammar normal	1	**4**	**6**	11
CELF grammar < 1 *SD*	0	2	**8**	10
CELF grammar < 2 *SD*	1	1	12	14
Total	2	7	26	35

Note. NDWR = number of different word roots; PPVT-III = Peabody Picture Vocabulary Test–Third Edition; EVT = Expressive Vocabulary Test; CELF = Clinical Evaluation of Language Fundamentals; MLU = mean length of utterance.

TABLE 5. Correlations (*df* = 31) between standardized tests and spontaneous speech measures with age and nonverbal IQ partialed out.

	Measure	1	2	3	4	5	6	7	8	9	10
1.	NDWR										
2.	MLU	.77**									
3.	IPSyn	.70**	.80**								
4.	PPVT-III	.41*	.31	.37							
5.	EVT	.33	.31	.31	.51*						
6.	Vocabulary combined	.42*	.36	.39*	.87**	.87**					
7.	Concepts and Directions[a]	.54**	.55**	.35	.24	.50*	.43*				
8.	Sentence Structure[b]	.36	.32	.23	.61**	.40*	.58**	.37			
9.	Word Structure[b]	.40*	.42*	.21	.40*	.58**	.56**	.56**	.53**		
10.	CELF grammar	.43*	.43*	.25	.55**	.58**	.65**	.52**	.81**	.92**	
11.	CELF total	.51*	.49*	.36	.58**	.68**	.73**	.72**	.59**	.84**	.86**

Note. NDWR = number of different word roots; MLU = mean length of utterance; IPSyn = Index of Productive Syntax; PPVT-III = Peabody Picture Vocabulary Test–Third Edition; EVT = Expressive Vocabulary Test.

[a] CELF measure of lexical–semantic skills. [b] CELF measure of grammatical skills

*$p < .01$. **$p < .001$.

derived from the natural language samples were all highly significant, ranging between .7 and .8, even after age and nonverbal IQ were partialed out. The correlation between the two vocabulary tests was significant, and these tests correlated significantly with the CELF Total score, both individually and when combined. The CELF subtest scores correlated significantly with CELF Total, and the sets of subtests that tap grammatical skills (Sentence Structure and Word Structure) were also intercorrelated. The correlations between the CELF lexical–semantic measure and the vocabulary tests were moderately high; however, the PPVT-III was not significantly correlated with the CELF measure (Linguistic Concepts/Concepts and Directions), despite the fact that they are both receptive measures.

Looking at the correlations across the standardized and spontaneous speech measures, Table 5 shows that the combined vocabulary score correlated significantly only with NDWR, not with MLU or IPSyn. The CELF Total score correlated significantly with NDWR and MLU, but not IPSyn. Indeed, none of the correlations between the IPSyn and the standardized test scores reached significance. NDWR correlated significantly with the CELF lexical–semantic measure, and MLU correlated significantly with both the CELF lexical–semantic measure and with CELF Grammar. Thus, two of the three spontaneous speech measures, NDWR and MLU, were significantly correlated with standardized test measures that assess performance in the same language domain.

To further explore the relationships between standardized and spontaneous speech measures, we conducted hierarchical regression analyses. Separate analyses were run for the two spontaneous speech measures as the dependent variables, MLU and NDWR, each of which was significantly correlated with the standardized measures. Because age was correlated with both of these dependent variables, $r(33) = .28$, $p = .051$ for MLU and $r(33) = .40$, $p = .008$ for NDWR, it was entered as a control variable in the first step of the regression analyses. Nonverbal IQ was not significantly correlated with either of the dependent variables, $r(33) = .15$ for MLU and $r(33) = .16$ for NDWR, and was therefore not entered into the regression models. For MLU, age was entered first, and then CELF grammar, CELF lexical–semantic, and vocabulary scores were entered into the regression model. For NDWR, vocabulary, CELF lexical–semantic, and then CELF grammar scores were entered into the model, after age.

For MLU, age accounted for 8% of the variance, but this was not significant, $F_{\text{change}}(1, 33) = 2.84$, *ns*. When CELF grammar was entered into the model, the R^2 change was .149, which was significant, $F_{\text{change}}(1, 32) = 6.18$, $p < .02$. The addition of the lexical semantic scores from the CELF and the vocabulary tests contributed an additional 10% to the variance, but this was not significant, $F_{\text{change}}(2, 30) = 2.32$, *ns*). Thus, for MLU only, the CELF grammar score contributed unique significant variance. For NDWR, age accounted for 16% of the variance, $F_{\text{change}}(1, 33) = 6.36$, $p < .02$; the R^2 change was .14 for vocabulary, $F_{\text{change}}(1, 32) = 6.43$, $p < .02$; and at the next step, the R^2 change was .09 for CELF lexical–semantic, $F_{\text{change}}(1, 31) = 4.38$, $p < .05$. CELF grammar did not account for any additional variance in the final step of the model. Thus, in addition to age, both lexical–semantic measures, from the standardized vocabulary tests and the CELF subtests, contributed unique variance to NDWR derived from the natural language sample.

Discussion

The main aim of this study was to investigate whether standardized tests and measures derived from natural language samples provide comparable assessments of language skills in children with autism. We specifically focused on measures that tapped lexical–semantic and morphological–syntactic abilities in these children, and, overall, our findings provided support for the view that both kinds of assessment are measuring the same linguistic abilities in this population.

Comparing our participants with autism to normative data on the standardized and spontaneous speech measures, we found that as a group, the children in this study performed lower than age expectations on all the measures. These findings suggest that the majority of, though not all, verbal children with autism have impairments in formal aspects of language as assessed by both kinds of measures included in this study, and confirm other data on language deficits in children with autism (e.g., Cantwell, Baker, & Rutter, 1978; Conti-Ramsden, Botting, Simkin, & Knox, 2001; Kjelgaard & Tager-Flusberg, 2001; Lord & Pickles, 1996; Stevens et al., 2000).

Our individual participant analysis, summarized in Table 4, indicated that the children were more impaired relative to normative data on the spontaneous speech measures than on the standardized tests. This analysis must be interpreted with caution because the norms to which we compared the children with autism were obtained in very different ways for the spontaneous speech measures and the standardized tests. Strictly speaking, the SALT reference database (Miller & Chapman, 2000), used to obtain means and standard deviations for MLU and NDWR, provides a comparison sample but not psychometrically defined norms. Thus, the difference in performance relative to the mean for the spontaneous speech measures and standardized measures may simply be a function of differences in the norms. On the other hand, it may be that this disparity between the different measures reflects genuine differences in the children's *use* of their linguistic knowledge in discourse versus highly structured testing situations. Because of their primary impairments in pragmatics and social reciprocity, children with autism may not use the range of vocabulary and grammatical constructions that they have acquired in everyday conversation, even with their mothers. This would suggest that measures of lexical–semantic and grammatical abilities obtained from natural language samples are influenced by pragmatic factors. More research is needed to develop comparable norms for spontaneous speech and standardized measures, which could be used to develop profiles of performances for children with autism or other language impaired populations.

The second set of analyses investigated the relationship between the different types of measures using both correlational and regression statistical methods. We found strong positive correlations among the language measures, even after partialing out age and nonverbal IQ. One exception was the IPSyn, a measure of grammatical knowledge that was derived from the natural language sample; this measure only correlated with the other spontaneous speech measures, but not with any of the standardized test scores. Scarborough et al. (1991) found that relative to MLU, IPSyn scores significantly underestimated linguistic knowledge in children with autism. The low scores obtained on the IPSyn were because the children with autism used a narrow range of grammatical constructions and had especially low scores on the Question/Negation subscale of the IPSyn. The data from this study are consistent with these earlier findings and suggest that IPSyn may be a less useful spontaneous speech measure of grammatical ability than MLU for children with autism.

In contrast to the findings for the IPSyn, MLU provided a good measure of grammatical ability for the children in this study, as shown by the significant correlations with the standardized test scores. In the regression analysis, the CELF grammar subtests (Sentence Structure and Word Structure) were the strongest significant predictors of the children's MLU scores. Nevertheless, performance on these CELF subtests accounted for only 15% of the variance in MLU beyond the variance explained by age, suggesting that MLU reflects more than the skills assessed on the standardized tests of grammatical knowledge. Our findings also indicated that spontaneous speech and standardized test measures of lexical–semantic skills were highly related in children with autism. NDWR from the natural language sample correlated significantly with the combined vocabulary score from the PPVT and EVT, as well as with the subtests on the CELF that assess this domain (Linguistic Concepts/Concepts and Directions). These standardized tests each contributed significant unique variance to NDWR, together accounting for 23% of the variance in NDWR beyond the variance explained by age. These findings suggest that for children with autism, measures derived from spontaneous speech and standardized tests are tapping the same specific abilities in lexical–semantic and morphological syntactic domains of language. These findings confirm earlier studies with typically developing children (Bornstein & Haynes, 1998; Ukrainetz & Blomquist, 2002), which had also found strong general and specific correlations across language measures derived from natural language samples and standardized psychometrically based language tests.

These findings are important for several reasons. The data presented confirm the utility of both standardized and spontaneous speech measures for assessing language in children with autism. Given that these kinds of measures are useful for different purposes, clinicians and researchers may feel more confident as they continue to use these measures in their work with children with autism. The strong correlations found among the different measures suggest that for children with autism, a relatively consistent picture of language abilities may be obtained, both in structured settings where standardized tests are administered and in language measures derived from more informal everyday conversational interactions. Despite the significant social, behavioral, and communicative impairments that characterize children with autism, language assessments may be obtained in both contexts.

Research on language in autism has focused extensively on the profile of performance across different domains of language, with considerable attention paid to the relative strengths in structural language in contrast to significant impairments in pragmatics and discourse (Tager-Flusberg, 1994; Wilkinson, 1998). Thus, for example, Tager-Flusberg and Anderson (1991) found dissociation between growth in structural language abilities as measured by MLU and the absence of developmental change in discourse in a small group of children with autism. One concern might be that the assessment of structural language abilities in natural language samples may not be commensurate with more valid assessments of such abilities using standardized measures, calling into question whether there really is a dissociation between these language domains in autism (or other populations). The findings from this study lend some support for taking measures such as MLU or other measures from natural language samples as a good proxy for assessing grammatical and lexical knowledge in children with autism, thus supporting the conclusions that in autism there may be significant dissociations between structural and functional aspects of language acquisition, especially in children with intact linguistic abilities.

The strong relationships we found among the CELF, vocabulary tests, and measures of spontaneous speech have important implications for the use of these kinds of measures. The findings provide empirical support to researchers' and clinicians' reliance on both types of measures as useful tools for identifying language impairments and quantifying linguistic skills of children with autism, as well as for matching groups in research studies and documenting developmental changes in language in this population. Given the very broad range of impairments that are found in autism, covering both linguistic and pragmatic deficits, clinicians may take advantage of these measures for diagnosing deficits in different language domains.

Acknowledgments

This research was supported by grants from the National Institute on Deafness and Other Communication Disorders and the National Institute of Neurological Disorders and Stroke (PO1 DC 03610; RO1 NS 38668). We are extremely grateful to Susan Bacalman, Laura Becker, Courtney Hale, Robert Joseph, and Jenny Roberts for their help in collecting some of the data reported in this article. We offer special thanks to the children and families who participated in this study.

References

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Bornstein, M. H., & Haynes, O. M. (1998). Vocabulary competence in early childhood: Measurement, latent construct, and predictive validity. *Child Development, 69*, 654–671.

Botting, N., Conti-Ramsden, G., & Crutchley, A. (1997). Concordance between teacher/therapist opinion and formal language assessment scores in children with language impairment. *European Journal of Disorders of Communication, 32*, 317–327.

Brown, R. (l973). *A first language: The early stages*. Cambridge, MA: Harvard University Press.

Cantwell, D., Baker, L., & Rutter, M. (1978). A comparative study of infantile autism and specific developmental receptive language disorder: IV. Analysis of syntax and language function. *Journal of Child Psychology and Psychiatry, 19*, 351–362.

Conti-Ramsden, G., Botting, N., Simkin, Z., & Knox, E. (2001). Follow-up of children attending infant language units: Outcomes at 11 years of age. *International Journal of Language and Communication Disorders, 36*, 207–219.

Craig, H. K., & Washington, J. A. (2000). An assessment battery for identifying language impairments in African American children. *Journal of Speech, Language, and Hearing Research, 43*, 366–379.

Dollaghan, C., Campbell, T., Paradise, J., Feldman, H., Janosky, J., Pitcairn, D., & Kurs-Lasky, M. (1999). Maternal education and measures of early speech and language. *Journal of Speech, Language, and Hearing Research, 42*, 1432–1443.

Dunn, L. M., & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test* (3rd ed.). Circle Pines, MN: American Guidance Service.

Dunn, M., Flax, J., Sliwinski, M., & Aram, D. (1996). The use of spontaneous language measures as criteria for identifying children with specific language impairment: An attempt to reconcile clinical and research incongruence. *Journal of Speech and Hearing Research, 39*, 643–654.

Elliott, C. D. (1990). *Differential Ability Scales: Introductory and technical handbook*. New York: The Psychological Corporation.

Evans, J. (1996). SLI subgroups: Interaction between discourse constraints and morphological deficits. *Journal of Speech and Hearing Research, 39*, 655–660.

Goffman, L., & Leonard, J. (2000). Growth of language skills in preschool children with specific language impairment: Implications for assessment and intervention. *American Journal of Speech-Language Pathology, 9*, 151–161.

Kjelgaard, M., & Tager-Flusberg, H. (2001). An investigation of language impairment in autism: Implications for genetic subgroups. *Language and Cognitive Processes*, *16*, 287–308.

Klee, T. (1992). Developmental and diagnostic characteristics of quantitative measures of children's language production. *Topics in Language Disorders, 12*, 28–41.

Klee, T., Schaffer, M., May, S., Membrino, I., & Mougey, K. (1989). A comparison of the age-MLU relations in normal and specifically language impaired preschool children. *Journal of Speech and Hearing Disorders, 54*, 226–233.

Koegel, L. K., Koegel, R. L., & Smith, A. (1997). Variables related to differences in standardized test outcomes for children with autism. *Journal of Autism and Developmental Disorders, 27*, 233–243.

Lord, C., & Paul, R. (1997). Language and communication in autism. In D. J. Cohen & F. R. Volkmar (Eds.), *Handbook of autism and pervasive developmental disorders* (2nd ed., pp. 195–225). New York: Wiley.

Lord, C., & Pickles, A. (1996). Language level and nonverbal social-communicative behaviors in autistic and language-delayed children. *Journal of the American Academy of Child and Adolescent Psychiatry, 35*, 1542–1550.

Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Lenventhal, B. L., DiLavore, P. S., et al. (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*, 205–223.

Lord, C., Rutter, M., & Le Couteur, A. (l994). Autism Diagnostic Interview–Revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of Autism and Developmental Disorders, 24*, 659–685.

Miller, J. (1991). Quantifying productive language disorders. In J. Miller (Ed.), *Research on child language disorders: A decade of progress* (pp. 211–220). Austin, TX: Pro-Ed.

Miller, J., & Chapman, R. (2000). Systematic Analysis of Language Transcripts (SALT) [Computer software, SALT for Windows, Research Version 6.1]. Madison: University of Wisconsin, Language Analysis Lab.

Paul, R., & Cohen, D. (1985). Comprehension of indirect requests in adults with autistic disorders and mental retardation. *Journal of Speech and Hearing Research, 28*, 475–479.

Rescorla, L., Roberts, J., & Dahlsgaard, K. (1997). Late talkers at 2: Outcome at age 3. *Journal of Speech and Hearing Research, 40*, 556–566.

Restropo, M. A. (1998). Identifiers of predominantly Spanish-speaking children with language impairment. *Journal of Speech, Language, and Hearing Research, 41*, 1398–1411.

Rollins, P. R., Snow, C., & Willett, J. (1996). Predictors of MLU: Semantic and morphological developments. *First Language, 16*, 243–259.

Rondal, J. A., Ghiotto, M., Bredart, S., & Bachelet, J. F. (1987). Age-relation, reliability, and grammatical validity of measures of utterance length. *Journal of Child Language, 14*, 433–446.

Rondal, J. A., Ghiotto, M., Bredart, S., & Bachelet, J. F. (1988). Mean length of utterance of children with Down syndrome. *American Journal of Mental Retardation, 93*, 64–66.

Scarborough, H. S. (1990). Index of Productive Syntax. *Applied Psycholinguistics, 11*, 1–22.

Scarborough, H. S., Rescorla, L., Tager-Flusberg, H., & Fowler, A. E. (1991). The relation of utterance length to grammatical complexity in normal and language-disordered groups. *Applied Psycholinguistics, 12*, 23–45.

Scarborough, H. S., Wyckoff, J., & Davidson, R. (1986). A reconsideration of the relation between age and mean utterance length. *Journal of Speech and Hearing Research, 29*, 394–399.

Semel, E., Wiig, E. H., & Secord, W. A. (l995). *Clinical Evaluation of Language Fundamental* (3rd ed.). San Antonio, TX: The Psychological Corporation/Harcourt Brace.

Sparrow, S. (1997). Developmentally based assessments. In D. J. Cohen & F. R. Volkmar (Eds.), *Handbook of autism and pervasive developmental disorders* (2nd ed., pp. 411–447). New York: Wiley.

Stevens, M. C., Fein, D. A., Dunn, M., Allen, D., Waterhouse, L., Feinstein, C., & Rapin, I. (2000). Subgroups of children with autism by cluster analysis: A longitudinal examination. *Journal of the American Academy of Child and Adolescent Psychiatry, 39*, 346–352.

Tager-Flusberg, H. (1994). Dissociations in form and function in the acquisition of language by autistic children. In H. Tager-Flusberg (Ed.), *Constraints on language acquisition: Studies of atypical children* (pp. 175–194). Hillsdale, NJ: Erlbaum.

Tager-Flusberg, H. (2000a). The challenge of studying language development in children with autism. In L. Menn & N. Bernstein Ratner (Eds.), *Methods for studying language production* (pp. 313–331). Mahwah, NJ: Erlbaum.

Tager-Flusberg, H. (2000b). Understanding the language and communicative impairments in autism. In L. M. Glidden (Ed.), *International review of research on mental retardation* (Vol. 20, pp. 185–205). San Diego, CA: Academic Press.

Tager-Flusberg, H., & Anderson, M. (1991). The development of contingent discourse ability in autistic children. *Journal of Child Psychology and Psychiatry, 32*, 1123–1134.

Ukrainetz, T. A., & Blomquist, C. (2002). The criterion validity of four vocabulary tests compared with a language sample. *Child Language Teaching and Therapy, 18,* 59–78.

Watkins, R. V., Kelly, D. J., Harbers, H. M., & Hollis, W. (1995). Measuring children's lexical diversity: Differentiating typical and impaired language learners. *Journal of Speech and Hearing Research, 38*, 1349–1355.

Wiig, E. H., Secord, W., & Semel, E. (1992). *Clinical Evaluation of Language Fundamentals–Preschool*. San Antonio, TX: The Psychological Corporation/Harcourt Brace.

Wilkinson, K. M. (1998). Profiles of language and communication skills in autism. *Mental Retardation and Developmental Disabilities Research Reviews, 4*, 73–79.

Williams, K. T. (1997). *Expressive Vocabulary Test*. Circle Pines, MN: American Guidance Service.

Received January 1, 2002
Accepted February 7, 2003
DOI: 10.1044/1058-0360(2003/xxx)

Contact author: Helen Tager-Flusberg, PhD, Department of Anatomy and Neurobiology, Boston University School of Medicine, 715 Albany Street, L-814, Boston, MA 02118-2526. E-mail: htagerf@bu.edu

Author Queries

AQ1: I have change "play dough" to "Play-Doh." Edit ok, or was a generic product used?

AQ2: It seems like a word is missing in this sentence: "Four participants produced fewer than 100 complete and intelligible utterances; for these transcripts, were calculated following procedures outlined for reduced samples…." Please clarify wording.

AQ3: In the note for Table 4, please explain what bold text represents for the bolded values.