An Assessment Battery for Identifying Language Impairments in African American Children

Holly K. Craig
Julie A. Washington
University of Michigan
Ann Arbor, MI

This investigation compares the performances of 24 African American children, diagnosed as language impaired (LI) and receiving school-based language therapy, to 2 groups of typically developing peers ($N = 48$) on 5 traditional types of language assessment measures. Three of the measures were derived from child-centered free play language sample analyses and included average length of communication units (MLCU), frequencies of complex syntax, and numbers of different words. Two of the measures examined language comprehension and included responses to requests for information in the form of Wh-questions and responses to probes of active and passive sentence constructions. The performances of the group of children with language impairments were significantly lower on each measure than that of chronological age matched African American children who were typically developing. Sensitivity and specificity of the battery appeared excellent. The findings are discussed in terms of the potential of these informal language measures to contribute to a culturally fair assessment protocol for young African American children.

KEY WORDS: African American, children, language, assessment

Young African American children are at considerable risk for being misidentified as language impaired because of a dearth of culturally fair language assessment methods. Culturally appropriate speech-language assessment protocols must be capable of distinguishing the child with atypical language development from the typically developing African American child who is a speaker of African American English (AAE). To be most effective, these protocols should assess the range of communication skills defined by the areas of articulation, language production, and language comprehension; include both formal and more child-centered assessment tasks; and be standardized on AAE speaking populations.

In the absence of a battery of culturally fair assessment methods for use with children from minority groups, a relatively new approach, dynamic assessment, has been proposed (Kayser, 1996; Lidz & Pena, 1996). This approach just recently has been found promising for African American children (Fagundes, Haynes, Haak, & Moran, 1998) who reportedly may benefit from multiple changes in tasks and formats (Boykin, 1977; Miller-Jones, 1989). In contrast to the development of new approaches, however, Dollaghan, Campbell, and colleagues (Campbell, Dollaghan, Needleman, & Janosky, 1997; Dollaghan & Campbell, 1998) offer an excellent example of a new way to apply an older method, the nonword repetition task (Archer, 1960), to the study of the African American child.

In the nonword repetition task, performance dependencies on prior world knowledge are reduced, resulting in a nonbiased assessment task. It seems important not to abandon time-tested approaches to language assessment prematurely if they can continue to serve an important assessment role.

Two examples of the retention of longstanding approaches to the assessment of African American children have resulted from fresh examination of the Arizona Test of Articulation Proficiency, Second Edition (Fudala & Reynolds, 1986) and the Peabody Picture Vocabulary Test, (PPVT–III; Dunn & Dunn, 1997). The former successfully distinguishes speech impaired from nonspeech impaired AAE speaking children (Washington & Craig, 1992). A system of scoring adjustments has been developed for use with AAE speaking children who are Southern Dialect users (Cole & Taylor, 1990), enhancing the breadth of application appropriate for this standardized measure. In addition, the PPVT (Dunn, 1959) has been revised twice (Dunn & Dunn, 1981, 1997), and the recently published third edition is more culturally fair (Washington & Craig, 1999). These two tests appear appropriate for assessing the areas of articulation and receptive vocabulary and therefore can make important contributions as standardized measures to an assessment protocol for African American children. More child-centered informal methods must be developed as well to make a culturally fair language battery more comprehensive.

At the University of Michigan for the past few years our research program has focused on characterizing the language development of young, typically developing, African American children who are speakers of AAE and has focused on both dialectal and nondialectal aspects (Craig, Washington, & Thompson-Porter, 1998a; Craig, Washington, & Thompson-Porter, 1998b; Washington & Craig, 1998; Washington, Craig, & Kushmaul, 1998). A number of the measures we have been using in our study of typically developing African American children appear to be good candidates for inclusion in a culture-fair assessment battery for African American children. These are: average length of C-units (Craig et al., 1998b), frequencies of complex syntax (Craig & Washington, 1994), responding to requests for information and to probes of the active/passive voice in sentence constructions (Craig et al., 1998a). Features that recommend these measures for assessment purposes are the following: (a) all are age and/or grade sensitive; (b) performance distributions on these measures can be quantified and characterized independently from the child's level of AAE use; and (c) mean C-unit lengths and frequencies of complex syntax can be sampled during low-structured and child-centered language sampling contexts. An assessment battery that incorporates measures like these of oral expression and comprehension, reflects low-structured discourse as well as elicited responding, and shows systematic changes with the child's chronological and academic growth would offer both clinicians and researchers valuable tools for evaluating the language of African American children. If these measures do not vary systematically with a child's socioeconomic status and gender, they would have relatively broad applicability across the population.

Although these measures appear promising, the critical next step in determining their potential as assessment procedures is to examine performances of children with language impairments on the targeted tasks compared to those of typically developing peers. The purpose of this study, therefore, was to examine the potential of a selected set of informal, well-established types of language measures, grounded in our prior research with AAE-speaking children, to distinguish African American children with language impairments from typically developing peers. The following questions were posed.

1. Are the performances of African American children with language impairments statistically different from those of peers matched for chronological age or mean C-unit length on selected expressive and receptive language measures?
2. What patterns characterize language production and comprehension, and how do the children with language impairments compare to those who are typically developing?
3. What are the sensitivity and specificity of this set of measures for detecting language impairment and for avoiding misidentifications of African American children who are typically developing?

Method

Participants

The participants were 72 African American children residing in metropolitan Detroit. All were speakers of African American English (AAE). African American children comprised more than 75% of the student body in the participating school districts in this research project. Amounts of dialect evidenced in the children's discourse varied widely. Discourse was scored for AAE using Washington and Craig's definitions (Washington & Craig, 1994, 1998). Typically developing children in this age range use two forms of AAE more than others (Washington & Craig, 1994; 1998): the zero copula/auxiliary (e.g., "this thing the hardest") and subject/verb agreement (e.g., "ok, food go in there"), and this was the case for the current participants. Tables 1 and 2 report two measures of dialect usage: a dialect density measure (DDM), and the percentage frequencies of C-units in a standard corpus of 50 C-units that included one or more

instances of AAE. The former was developed in order to control for the relationship between increased C-unit lengths and opportunities for AAE (Craig et al., 1998b). The DDM was derived by dividing the frequencies (tokens) of AAE by the number of words (tokens) in standard length 50 C-unit samples. The CLAN programs of the Children's Data Exchange System, CHILDES (MacWhinney, 1994) automatically generated the frequencies and types of AAE using the frequency command (FREQ). In earlier work with preschoolers (Washington & Craig, 1994; Craig & Washington, 1994), we reported percentage frequencies of utterances in which children produced one or more AAE tokens. For the convenience of the reader in relating the current data to our earlier work, both the DDM and percentage frequency measures are reported here in Tables 1 and 2.

The participants were assigned to one of three groups (see Table 1). Twenty-four of the children were diagnosed as language impaired and enrolled on the speech and language caseloads of their school districts. The 48 remaining children were selected from our larger research program to create two control groups of typically developing participants matched to the children with language impairments based on either their chronological ages or their average C-unit lengths (MLCU). Although none of the children with LI had been included in our prior published studies, 19 (40%) of the typically developing children had been. The children were considered to be typically developing based upon teacher and parental judgments, and none were enrolled in special education or speech and language services of any type. All participants passed a hearing screening at 25 dB for 500, 1,000, and 4,000 Hz (ANSI, 1989).

Table 1. Numbers of participants in the language-impaired (LI), chronological-age in months (CA), and mean length of C-units in words (MLCUw) groups; their means and standard deviations on the CA and MLCUw matching variables and on the Dialect Density Measure (DDM); and the percentage of C-units coded for one or more instances of African American English (C-units/AAE).

	LI	CA	MLCUw
Socioeconomic status			
Low	14	14	14
Middle	10	10	10
Total	24	24	24
Gender			
Male	18	19	16
Female	6	5	8
Total	24	24	24
Chronological age			
M	82.5	82.4	nm
SD	27.6	26.2	
MLCUw			
M	2.89	nm	2.85
SD	.75		.64
DDM			
M	.057	.060	.064
SD	.032	.033	.030
C-unit/AAE			
M	15.2	17.7	15.7
SD	8.9	8.8	7.4

Note. nm = not matched.

Language Impaired Group (LI)

All children in the school district enrolled in the elementary grades who were receiving language intervention services were identified and contacted by their speech-language pathologists and invited to participate. Qualification for intervention services required referral from their classroom teachers; scores more than 1 standard deviation below the mean on a standardized language test, most typically the Clinical Evaluation of Language Fundamentals, Revised (Semel, Wiig, & Secord, 1987); and clinical judgment by the school speech-language pathologist that the child evidenced a language impairment. Thirty-four children were referred to the project over the course of the 4-year project span. Language documentation was current in all files, except for 2 children so these children were excluded from further consideration. Review of the files revealed that 5 of the children evidenced primary phonological deficits and that therapy emphasized articulation, and another 3 had primary emotional problems with diagnoses in the autism spectrum. The remaining 24 children were receiving therapy that emphasized language goals and became the participants in the current project.

Chronological age was not a subject selection variable. Participants in the LI group ranged in age from 4 to 11 years, with eleven 4- and 5-year-olds, four 6- and 7-year-olds, six 8- and 9-year-olds, and three 10- and 11-year-olds. There were 18 males and 6 females. Fourteen of the children were from low (LSES) and 10 from middle socioeconomic status (MSES) homes. Socioeconomic status was determined by the demographics of the children's communities, and each child's participation or nonparticipation in a free or reduced-price federal lunch program, as described elsewhere (Washington & Craig, 1998). All of the participants were verbal and intelligible. They were heterogeneous, however, in terms of the nature of their language impairments. This heterogeneity was designed to reflect the diversity of language disorders apparent on the caseloads of many public-school speech-language pathologists and allowed us to evaluate whether the protocol would be able to distinguish a broad range of language problems. Potential within group differences were explored by dividing

the LI group into three subgroups and comparing performances on the five measures of interest across the following subgroups:

1. children with language impairments secondary to a primary developmental or medical disorder (DIS, $n = 6$)
2. children who presented a primary language deficit and a K-ABC Triangles score of 7 or more were considered children with SLI ($n = 11$)
3. children with low cognition who presented a primary language deficit and a Triangles scaled score of 6 or less (LOW; $n = 7$)

The 6 participants in the DIS subgroup had language impairments secondary to another primary deficit, including fetal alcohol syndrome ($n = 1$), closed head injury ($n = 1$), Down syndrome ($n = 1$), neurological impairment ($n = 1$), and children classified as educable mentally impaired (EMI; $n = 2$). Three were 4 or 5 years of age, and three were 8 or 9 years old. Their cognitive scores on the Triangles subtest of the Kaufman Assessment Battery for Children (K-ABC; Kaufman & Kaufman, 1983) were highly variable, ranging from 1 to 11, with a mean of 6.2 and a standard deviation of 3.6. Triangles was selected as a measure of cognition because it is culturally fair and examines a generalized cognitive skill (for additional discussion see Craig et al., 1998a).

The remaining 18 participants had language impairments that were primary in nature, occurring in the absence of any clinically significant behavioral, emotional, sensory, or neurological problems. Eleven of the children with primary language impairments achieved scaled scores of 7 or above on Triangles, scores within 1 standard deviation of the subtest's mean scaled score of 10. By considering a performance of 7 or above on this generalized task of cognition as evidence of cognitive skills within normal limits, these 11 participants met criteria consistent with those established for the diagnosis of Specific Language Impairment (SLI; Aram, Morris, & Hall, 1993; Stark & Tallal, 1981). Six of these children were 4- and 5-year-olds; one was 7; three were 8 or 9; and one was 10-years-old. The other 7 participants in the LOW subgroup achieved scaled scores of 4 through 6 on Triangles, performances 2 standard deviations below the scaled score mean of 10. Two of these children were 5-year-olds, three were 6- or 7-year-olds, and two were 10 or 11 years of age.

Chronological Age Mates (CA)

The 24 children in the CA group were selected retrospectively on a continuous enrollment basis from the program's research files of typically developing children to match as closely as possible the chronological ages, the distribution of LSES and MSES, and the gender of the 24 children with language problems. The chronological ages of the LI and CA groups in months were not significantly different [$t(46) = .011, p > .05$], and the same distribution of 4- and 5-year-olds, 6- and 7-year-olds, and so forth, as that of the LI group was created (see Table 1). Of these three criteria, gender was the last subject-selection criterion considered because in prior work with typically developing children, gender was not a source of systematic variation for the nondialectal measures used in this study (Craig et al., 1998a, 1998b). When more than one match was identified who met all three criteria, the participant was selected from the pool using randomized procedures.

MLCU Mates (MLCUw)

The 24 children in the MLCUw group also were selected retrospectively from the program's research files of typically developing children, again on a continuous enrollment basis to match the mean C-unit lengths in words (MLCUw) of the 24 children with language problems. The MLCUw's of the LI and MLCUw groups were not significantly different [$t(46) = 1.32, p > .05$]. The distribution of LSES and MSES were matched, and the gender of the LI children also was matched when possible. When more than one child was identified as a potential match, the selection was made using randomized procedures. See Table 1. The group mean for the chronological ages of these children was approximately 20 months younger than that obtained for the LI group. There were nineteen 4- and 5-year-olds and five 6- and 7-year-olds.

Data Collection and Scoring

The data collection protocol consisted of three randomly ordered components: collection of a spontaneous language sample, administration of a task designed to elicit responses to Wh-questions, and administration of a task probing the active/passive voice distinction. From the spontaneous language samples, two quantitative measures: mean length of C-units and number of different words, were selected as metrics of grammatical development and expressive vocabulary, respectively. Mean length of C-unit data have been published previously for young African American children (Craig et al., 1998b), but data for the numbers of different words were new to this investigation. Both metrics are frequently used to identify oral language disorders resulting from a variety of etiologies and to evaluate progress within clinical settings (Miller, 1981; Watkins, Kelly, Harbers, & Hollis, 1995). Amount of complex syntax was selected as well because problems in the acquisition of advanced syntactic relationships are central to most developmental language problems (Aram & Nation, 1975; Johnston &

Kamhi, 1984; Stark & Tallal, 1981). In addition to production, the children's responses to requests for information in the form of Wh-questions and their ability to distinguish active from passive voice constructions were examined. Tasks like these relate to language growth (Chapman, 1988; Craig et al., 1998a; Parnell, Patterson, & Harding, 1984) and have potential, therefore, to contribute to the comprehension portion of a culturally fair assessment protocol.

The language samples were 15 to 20 min in duration, and the comprehension tasks required approximately 10 min of administration time each. The comprehension tasks each required only approximately 5 min to score. The transcription and scoring of the 50 C-unit samples was more time intensive and varied with the child's intelligibility. Using CHILDES and an audio transcriber with foot pedal, transcription and coding time were approximately 2 hours per child. Spontaneous language sampling has been the core child-centered data collection component of our research program, but other tasks have been added or removed over time as questions and priorities have changed. Accordingly, spontaneous language samples were collected and analyzed for all participants, but the responses to the Wh-questions task was unavailable for 4 of the participants (1 LI, 2 CA, 1 MLCUw), and the active/passive voice distinction task for 5 (1 LI, 3 CA, 1 MLCUw). Each task is discussed below.

Spontaneous Language Sampling and Analysis

Spontaneous language samples were collected during dyadic interaction with an unacquainted female African American examiner who spoke AAE to the children. As reported elsewhere, a number of African American examiners have participated in data collection, and variations in their use of AAE have not affected the children's use of dialect in systematic ways (Craig et al., 1998a; Washington & Craig, 1998; Washington et al., 1998). Language samples were collected during free play using action figures or dolls and associated props, the Fisher-Price School, or Legos, which the children selected from three possible sets. The samples were transcribed orthographically using the segmentation criteria of Loban (1976) for C-units. As discussed elsewhere (Craig et al., 1998b), these included independent clauses plus their modifiers, nonclausal units that were responses to adult questions, and single-word stereotypical acknowledgments to prior adult comments or "what" functioning as a contingent query. C-units were transcribed into CHAT files consistent with the conventions of CHILDES. The first 50 wholly intelligible C-units were identified. Most of the children produced 50 C-units in the 15- to 20-min sampling periods. Two of the 72 children did not produce 50 C-units in this time period (one of the LSES children with SLI, and one of the MSES preschoolers). Their free-play samples were included in the data set so that the findings would be maximally representative and not reflect only highly verbal children (as recommended by Miller & Chapman, 1981; Craig et al., 1998b).

Average C-Unit Lengths

The mean length in words and in morphemes of the first 50 wholly intelligible C-units (MLCUw and MLCUm, respectively) were calculated for each participant using the mean length of turn (MLT) and mean length of utterance (MLU) CLAN programs of CHILDES.

Expressive Vocabulary

The number of different words (NDW) in the 50 C-unit samples was calculated as a measure of lexical diversity and expressive vocabulary. The FREQ command of CLAN automatically generated word lists. These lists were then edited so that morphological variations in the form of number and tense markers on regular nouns and verbs were ignored. For example, *girl*, *girls*, for nouns and *walks*, *walked*, and *walking* for verbs were treated as the same noun or verb lexical root. These morphological forms may be variably included or excluded when a child is speaking AAE, so it seemed appropriate to represent this aspect of production as dialectal, rather than lexical.

Syntactic Complexity

Each C-unit was scored for the presence of complex syntax (CSyn), using Craig and Washington's (1994) taxonomy. The codes ranged from simple infinitival relationships (e.g., "he tryin' to put it in his hand" coded as an infinitive with same subject) to clauses joined by a variety of cognitively more advanced subordinate conjunctions (e.g., "then he came back after they blew up Madison-Square Garden"). More than one complex syntax code was possible per C-unit. For example, "that's why she always take me to Toys-R-Us all the time when she pick me up" was coded as two variants of the noninfinitive Wh-clause, resulting in two tokens and one type for computation purposes. See Craig et al., 1998b for additional discussion. Complex syntax codes were entered in the CHAT files as dependent coding tiers and analyzed using the FREQ commands of CLAN for the first 50 C-units.

Comprehension of Requests for Information

This task presented two activity pictures to the children, each depicting barbecuing and snow shoveling—

pictures number 33 and 35, respectively, from the Bracken Concept Development Program (Bracken, 1986)—and probed responses to Wh-questions. Earlier work demonstrated that performances were not different statistically on the two pictures, so raw scores can be combined and then reflect more than a single trial on each of the request types (Craig et al., 1998a). For this task, each picture is accompanied by 12 questions presented by the examiner using AAE. The prompts are designed to elicit a range of responses from simple naming (i.e., What this?) to cognitively more demanding probes of manner (How he + *perform action*?) and time (When this happenin'?). The order of presentation of each prompt for each picture was randomized. Examiners recorded on a scoring form the response each child made to each prompt, in the form of a spoken utterance or picture point. Subsequent scoring of the examiner's records assigned 0 points if the child said something unrelated, "I don't know," or did not respond (Type A error); 1 point if the child responded but to a potentially different Wh-question (Type B error); 2 points if the child responded to the pragmatic intent of the specific request for information but used a nonspecific referent or misnamed it (Type C error), and full credit (3 points) if the child produced the target response. The total possible score was 72 points. (See Craig et al., 1998a.)

Comprehension of Active/Passive Sentences

This task explored children's word-order strategies for comprehension of reversible active and passive sentence constructions (RevS) using a forced-choice picture pointing task, a task format selected for its high ecological validity to many traditional types of assessment that are more formal in nature. The stimulus set of picturable agents and actions was pretested with each subject for familiarity, and all participants readily identified the referents. RevS consisted of 30 spoken prompts, three for each of 10 pairs of black line drawings. The prompts used common nouns and verbs that were selected for their likelihood to occur bidirectionally in real life experiences. An example was the mom hug the baby (target active voice trial), the baby hug the mom (foil active voice trial), the baby was hugged by the mom (passive trial). Both the order of the picture pairs and the order of the spoken prompts were randomly determined. The examiner recorded each picture point made by the child on the score form for later point assignment. A match between the prompt and the child's picture selection on the passive voice trial was credited only when the child also correctly matched both active voice trials. In order to maintain balance between the contributions of the active and passive trials to the total scores, scoring assigned only one point to correct responding on both active trials. The total possible score was 20 points. For additional discussion see Craig et al. (1998a).

Reliability

Transcription reliabilities were established for all participants. Approximately 10% of each transcript was retranscribed by an independent observer. A point-to-point comparison at the level of the morpheme was high, 91%, when the number of agreements was divided by the number of agreements plus disagreements. Seven transcriptions, representing approximately 10% of the participants, were randomly selected and recoded by a second observer. Point-to-point comparisons were calculated for each scoring system by dividing the number of agreements by the number of agreements plus disagreements. The percentages of agreement for AAE tokens was 90%, for CSyn types was 94%, for CSyn tokens was 89%, and for C-unit segmentation was 99%.

Scoring reliabilities of the two comprehension tasks were examined as well. Seven of the response forms for each task were randomly selected. The records of spoken responses and nonverbal points on each response form were rescored by an independent examiner as a check on the scoring decisions made by the original examiner. Scoring agreements were high at 98% for the Wh-questions task and 97% for the RevS task when the number of points in agreement was divided by the number of agreements plus disagreements.

Results

Potential effects of SES and gender on the five types of measures were examined and found to be nonsignificant. In addition, dialect density showed nonsignificant relationships across the LI, CA, and MLCUw groups and across the LI subgroups [DDM: $F(2, 21) = 0.57, p > .05$]. The data were collapsed, therefore, across SES, gender, and dialect density in subsequent analyses. In addition, across the three subgroups of LI, nonsignificant relationships ($p > .05$) were found for each of the five types of measures in the assessment protocol [MLCU: in words $F(2, 21) = 0.77, p > .05$; in morphemes $F(2, 21) = 0.73, p > .05$; CSyn: $F(2, 21) = 0.43, p > .05$; NDW: $F(2, 21) = 1.11, p > .05$; Wh-q: $F(2, 20) = 0.17, p > .05$; RevS: $F(2, 20) = 0.47, p > .05$]. See Tables 1 and 2. Whereas the LI subgroups varied in size, the minima and maxima are presented as well in Table 2, as a check on the nonsignificant ANOVA values. The extent of overlap further indicates that performances were not distinctive by subgroup.

Each measure was examined for its potential to distinguish the group of children with language impairments from the two groups of typically developing children. A significance level of .01 was established to reduce

Table 2. Means, standard deviations, and Minima and Maxima (Min–Max) for the specifically language impaired (SLI), low cognition (LOW), and combined disorders (DIS) subgroups on mean length of C-units in words (MLCUw) and morphemes (MLCUm), frequencies of complex syntax (CSyn), numbers of different words (NDW), points on the responses to Wh-questions (Wh-q), and responses to reversible sentences (RevS) tasks on the Dialect Density Measure (DDM) and percentage of C-units coded for instances of African American English (C-units/AAE).

	SLI (n = 11)	LOW (n = 7)	DIS (n = 6)
MLCUw			
M	3.00	3.01	2.56
SD	.68	.66	.68
Min-Max	2.16–5.24	2.06–3.90	1.76–3.54
MLCUm			
M	3.26	3.28	2.78
SD	.94	.75	.74
Min–Max	2.37–5.72	2.24–4.34	1.92–3.88
CSyn			
M	3.7	3.6	2.2
SD	3.6	3.6	3.1
Min–Max	0–11	1–11	0–7
NDW			
M	70.8	72.7	60.8
SD	15.5	17.1	13.3
Min–Max	51–106	51–102	44–77
Wh-q			
M	50.7	47.4	46.8
SD	14.3	12.9	17.4
Min–Max	29–68	28–60	18–64
RevS			
M	11.2	9.4	12.4
SD	4.7	4.5	7.4
Min–Max	5–19	1–14	1–20
DDM			
M	.057	.065	.046
SD	.029	.042	.023
Min–Max	.013–.110	.030–.126	.014–.075
C-units/AAE			
M	16.3	17.4	10.7
SD	8.4	11.9	4.8
Min–Max	6–32	8–40	4–16

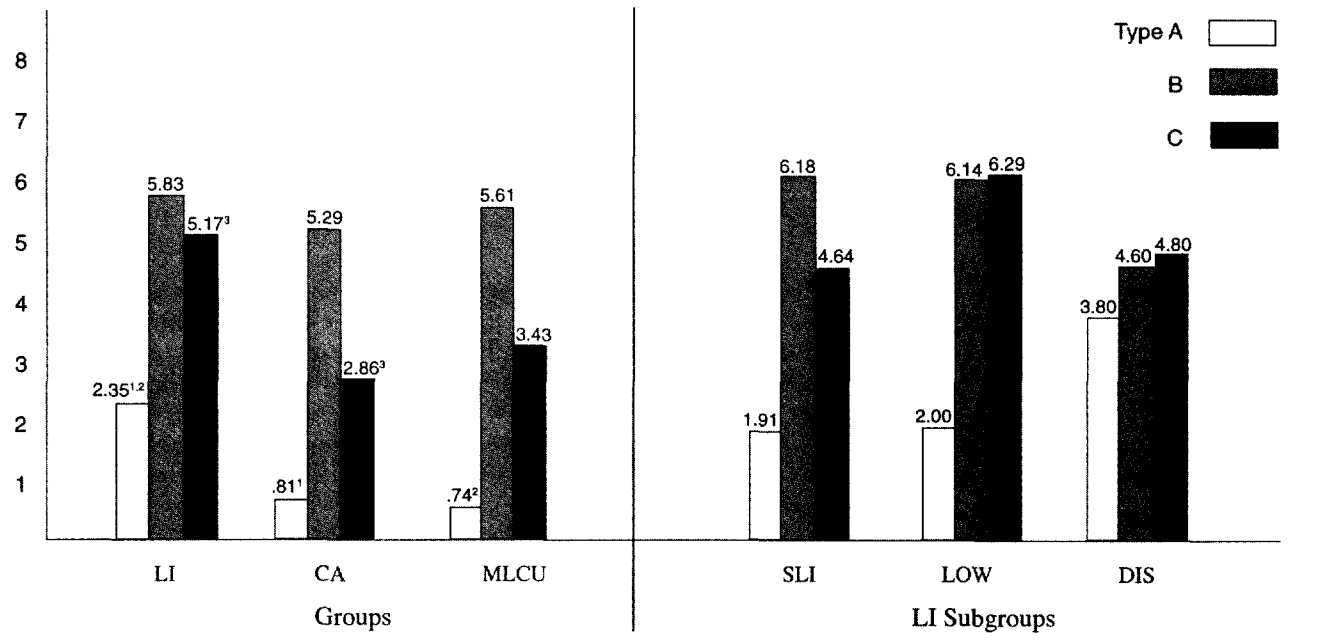
the effects of any significant intercorrelations (Craig et al., 1998a, 1998b) by dividing the experimentwise alpha level of .05 by the five types of measures in the assessment protocol ($.05/5 \leq .01$). The quantitative metrics of language growth revealed significant differences between the LI group and their CA mates. The mean length of C-units in words was significantly less [$t(46) = 3.00, p = .004$] for the LI group ($M = 2.89$) than their chronological age mates ($M = 3.57$). The mean length of C-units in morphemes varied significantly across the three groups [$F(2, 69) = 7.84, p = .001$]. A Tukey-HSD post hoc analysis revealed significant differences between the LI and CA groups and also between the CA and MLCUw groups. The number of different words also varied significantly across the three groups, [$F(2, 69) = 6.95, p = .002$]. A Tukey-HSD post hoc analysis revealed significant differences between the LI and CA groups, and also between the CA and MLCUw groups. See Table 3. All of these quantitative measures were successful in identifying the children in the LI group as functioning lower linguistically than their same-aged peers who were typically developing.

The amounts of complex syntax varied significantly by group [$F(2, 69) = 5.11, p = .009$] and a Tukey-HSD post hoc analysis once more revealed that the LI group produced significantly less complex syntax than the CA group ($p < .05$). This corresponded to a significant difference in the number of different types [$F(2, 69) = 4.21$; $p = .019$] between the LI and CA groups, whereas no other group differences were apparent ($p > .05$). See Table 3. As a group, the chronological age controls produced almost twice ($M = 4.00$; $SD = 2.0$) the number of different types of complex syntax compared to that of the LI group ($M = 2.42$; $SD = 2.1$). The use of conjunctions, infinitives with same subjects, noninfinitive Wh-clauses, and noun phrase complements, were produced by at least half of the CA controls and were some of the types of complex syntax most widely dispersed across the participants in the LI and MLCUw groups as well. Examination of the distributions across the subgroups within the LI group revealed that again conjunctions and infinitives with same subjects were the most widely used types of complex syntax regardless of subgroup.

The responses to the Wh-questions task revealed significant total point score differences by group [$F(2, 65) = 7.52, p = .001$]. A Tukey-HSD post hoc analysis revealed significantly lower performances ($p < .05$) by the LI group ($M = 48.9$) compared to both the CA group and the MLCUw group ($M = 59.5$ and 57.4, respectively). An analysis of the children's errors revealed that it was rare, less than once per child on average, for the typically developing children to fail to respond to the question probes or to produce an unrelated reply. However, the LI group differed significantly from the typically developing children in their frequency of no-response and unrelated response type errors [$F(2, 64) = 4.28, p = .02$]. Tukey HSD post hoc testing indicated that the significant differences ($p < .05$) were between the LI and both groups of typically developing children. See Figure 1.

The errors of the two groups of typically developing children were primarily factual ones. In contrast, the LI group evidenced factual errors and errors in the form of responses to a different question, at comparable levels. Responding to a potentially different question probe

Figure 1. Percentage of participants in each group and each LI subgroup with an unrelated or no-response (Type A error), a response that answers a potentially different question (Type B error), and a response that is nonspecific or incorrectly names the referent (Type C error) on the Responses to Wh-questions task. 1–3, $p \leq .05$



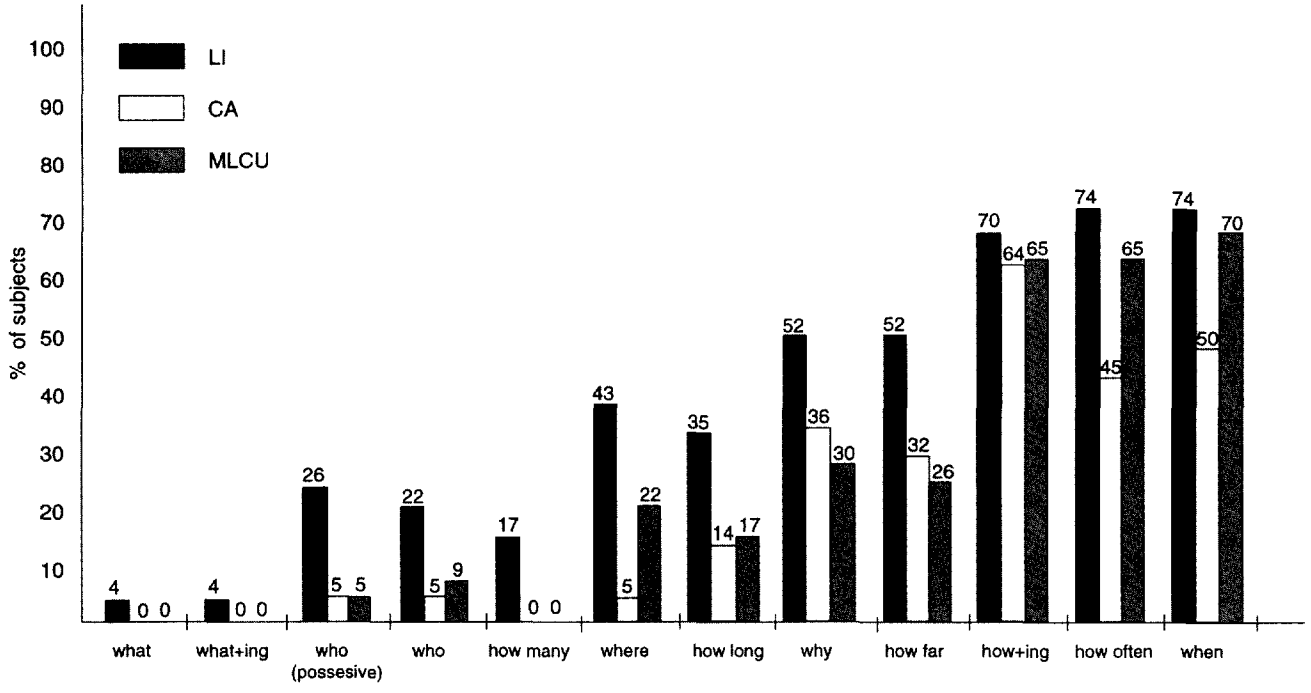
(Type B) showed significant group differences [$F(2, 64) = 3.25$, $p = .045$] and Tukey HSD post hoc testing indicated that errors of this type were significantly greater ($p < .05$) by the LI group than the CA group. Considered together, the CA group and the MLCUw group rarely failed to be responsive and primarily made errors that were factual or imprecise in nature. In contrast, errors of all types characterized the performances of the LI group, and those affecting response meaning (no response or "I don't know" and responding to a different question) were more characteristic of the LI group.

In order to further probe the nature of the LI group differences, the error profiles for the subgroups within the LI group were examined and revealed interesting differences (see Figure 1). The profile for the SLI subgroup was quite similar to that of the CA and MLCUw groups. Unlike the group or subgroup profiles, the DIS subgroup showed a flatter distribution of error types with substantially more no-response or "I don't know" (Type A) errors. The LOW subgroup showed a mixed profile with relatively few no-response or "I don't know" (Type A) errors like the SLI subgroup, and the CA and MLCUw groups, but approximately comparable levels of errors involving responses to a potentially different Wh-question (Type B) and errors involving nonspecific referents or misnaming (Type C), more like the DIS subgroup.

Errors also were examined relative to the type of Wh-question posed. Figure 2 displays the performances of the three groups relative to the form of the Wh-question. Overall, the LI group had more difficulty with most types of probes, but all three groups made more errors in response to the *how + ing*, *how often*, and *when* types of requests for information. The children's performances on the reversible sentences task also varied significantly by group [$F(2, 64) = 4.81$, $p = .011$. A Tukey-HSD post hoc analysis revealed that the LI group ($M = 10.9$) performed significantly lower ($p < .05$) than the CA group ($M = 14.9$). See Table 3.

Individual performance profiles were explored as well in order to evaluate the ability of the battery to distinguish children with LI from typically developing peers. Figure 3 presents each of the 24 pairs of children in the LI group and their peer in the CA group, organized by ascending chronological age in months. The ability of the battery to identify children as LI was probed by searching for children in the LI group who scored lower than their CA control. This was accomplished by visual inspection of the pairs and in the absence of normative means and standard deviations by age, used a very strict criterion of *any lower score of any magnitude* showing a discrepancy from the CA control. Further, these lower performances had to be apparent on two or more of the measures, consistent with standard clinical practice that does not rely upon a single low performance for identification purposes. Using these strict criteria, all of the children in the LI group (24 of 24) evidenced performances lower than that of their CA match on at least two of the five types of measures. Accordingly, the battery has a sensitivity of

Figure 2. Percentage of participants in each group having both trials wrong on the Responses to Wh-questions task.



1.00, where sensitivity is calculated as the number of participants identified as language impaired with these criteria, divided by the number of LI participants. One of these pairs had no comprehension data (Pair 7); nevertheless, he met the criteria for identification as LI on the battery because again, at least two of his (expressive) scores were lower than those of the CA control. Another three pairs (Pairs 11, 14, and 20) were missing one or both of their comprehension data, this time for the CA control. In these cases, the performance of the LI member was compared to that of his or her ascending CA neighbor on the figure as an alternative point of reference. This point of reference did not change the identification of the LI member of the pair as LI, using the criterion of lower performances on at least two language measures.

There was considerable variation in the CA performances on this battery. Two of the CA peers had no comprehension data available and so were removed from this examination (Pairs 11 and 14). Examination of the distribution of CA scores per measure relative to the CA distribution as a whole, suggests that despite expected variability and a trend for increasing slopes with age, the CA controls in Pairs 3, 5, and 17 are outliers on at least two of the five measures. Using a criterion of two or more low measures compared to CA peers, a specificity for the battery of .86 (19/22) was determined, where specificity was calculated as the number of CA participants identified as typically developing using these criteria, divided by the number of CA participants.

Discussion

The major goal of this study was to explore whether a set of child-centered measures derived from spontaneous free play language samples, and from responses to fairly traditional types of comprehension probes, would distinguish African American children identified as language disordered from typically developing peers. The findings were that all five types of measures were successful in identifying the group of children enrolled in language intervention from a group of chronological age mates. Each measure, therefore, appears to be a candidate for continued use and need not be abandoned. MLCU (in either words or morphemes) as a measure of average C-unit length, number of different words as a measure of expressive vocabulary, frequencies of complex syntax as a measure of grammatical growth, responses to requests for information and to active and passive sentence constructions as measures of comprehension, can be combined with the AAPS and the PPVT–III for the assessment of school age African American children. Together they offer both low and high structured procedures, more than a single look at the major language domains, and a comparison of expressive and receptive modalities. Lack of systematic variation of the measures with SES and gender indicates that they may have good broad generalizability across the population of African American children. Further, using low performances on at least two of the measures as criterion, the sensitivity of the battery was 1.00 and specificity

Table 3. Means and standard deviations in the mean length of C-units in words (MLCUw) and morphemes (MLCUm), frequencies of complex syntax (CSyn) tokens and types, numbers of different words (NDW), responses to Wh-questions (Wh-q), and responses to reversible sentences (RevS) for the language impaired (LI), chronological-age (CA), and MLCUw matched groups.

	LI	CA	MLCUw
MLCUw			
M	2.89[1]	3.60[1]	na
SD	.75	.87	
MLCUm			
M	3.15[2]	3.97[2, 3]	3.12[3]
SD	.84	1.00	.68
CSyn (tokens)			
M	3.3[4]	6.8[4]	4.6
SD	3.4	4.1	3.9
CSyn (types)			
M	2.4[5]	4.0[5]	2.7
SD	2.1	2.0	1.9
NDW			
M	68.9[6]	83.7[6, 7]	67.5[7]
SD	15.6	19.9	14.1
Wh-q			
M	48.9[8, 9]	59.5[8]	57.4[9]
SD	14.0	7.4	5.8
RevS			
M	10.9[10]	15.0[10]	12.6
SD	5.2	3.9	3.4

Note. na = not appropriate—was a subject-selection variable.
[1-10], $p \leq .01$

was .86, generally considered excellent levels (Meisels, 1989). Unfortunately, in the absence of normative data, Sensitivity and Specificity had to be estimated relative to the CA control group. This required application of a very strict criterion and conservative approach to the estimations. Nevertheless, these ratios indicate that the measures will be successful in identifying children with language impairments without falsely identifying typically developing African American children. It is noteworthy that the battery was successful with this cohort of LI children, 11 of whom were SLI. Identification of SLI in AAE-speaking children is especially difficult because of the overlap at a surface level in the major forms of AAE and the morphosyntactic characteristics of SLI.

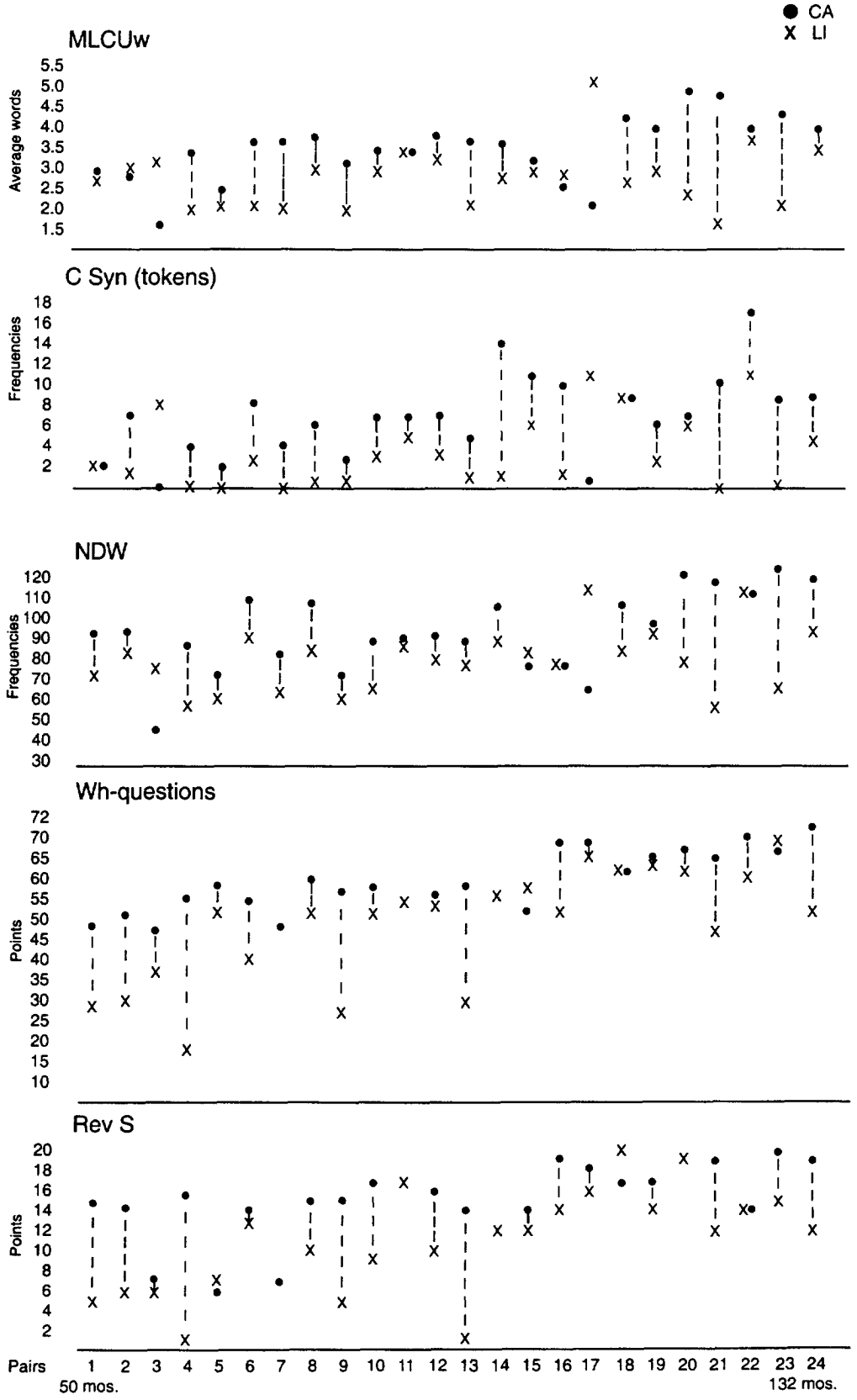
Our previous work with these types of measures (Craig et al., 1998a, 1998b) focused on the performances of typically developing children. The present study extended this work by determining whether each measure could distinguish children with language impairments against the "gold standard" of clinician judgment (see for example, Aram et al., 1993; Tomblin et al.,1996). The present investigation only indicates that these measures are good candidates for assessment purposes with African American children. In order to become part of current clinical practice, norm-referenced statements now need to be developed for each measure. Considerable future research will be needed to determine means and standard deviations by age, to define normal distributions for each measure, and to test the sensitivity of these distributions against clinician judgment. Although this set of activities represents considerable additional future work, the outcomes should yield culturally fair assessment methods for African American children and thus is well worth the investment of resources and effort.

In the present study, the MLCU, NDW, complex syntax, responses to Wh-questions and to reversible sentences consistently differentiated the LI from the CA groups, indicating good validity for identification purposes using age-discrepancy criteria. Currently, age-discrepancy criteria are recommended strongly over other criteria such as cognitive referencing for identifying children with language problems (Lahey, 1988; Stark & Tallal, 1981; Tomblin et al., 1997). This recommendation is particularly important when appropriate cognitive, IQ, and achievement measures remain largely unavailable for minority children. A comparison to peers of the same age is a conservative and ecologically valid standard for comparison.

Although these five measures consistently differentiated children with language impairments from their chronological age mates, all but one did not distinguish the LI from the MLCUw group. Considerable prior research with majority children also has failed to show significant differences on many receptive and expressive measures of these types between children with language impairments and typically developing children matched for language levels (Bishop, 1979; Morehead & Ingram, 1973). Performances on the Wh-questions task was the only measure that yielded significantly lower scores for the LI group from both the CA and the MLCUw-matched peers. Error analyses on this task also yielded qualitatively different response profiles from those of the typically developing children. Interestingly, this was also the only task that yielded discernible LI subgroup variations, specifically in the profiles of error types. Future research with the Wh-questions task appears warranted.

The five measures did not reveal statistical differences among the three subgroups within the LI group. For the purposes of this study, the findings that the five measures identified the LI group from age mates, despite heterogeneity of language profiles, underscores the utility of a protocol of this type for clinical practice. However, the failure to find statistical differences may be due at least in part to the small and unequal sample

Figure 3. Individual performances of pairs of children in the LI group and their matches in the CA group on the five types of measures. MLCUw and MLCUm yielded similar profiles, so just one measure is included for the purposes of this analysis.



sizes for each subgroup. Indeed, given the heterogeneous nature of the individuals in the subgroup of children with language impairments secondary to another medical condition (DIS subgroup), the mean values on these measures may be meaningless. The considerable overlap in the distributions of scores obtained by each subgroup (see minima and maxima in Table 2) suggest that the failure to find statistically significant subgroup differences on these measures will not be resolved by simply increasing the size of the participant groups in additional research. Other factors may need to be pursued.

The measures overall are fairly gross and quantitative. They may *identify* children with language impairments from typically developing children of the same age but fail to provide in-depth diagnostic information about the nature of the disorders. More qualitative analyses will be necessary to answer questions about the nature of language impairments in this population, just as they have been important to advance understanding of the language impairments of children who are members of the majority culture (Bedore & Leonard, 1998; Conti-Ramsden & Jones, 1997; Rice & Wexler, 1996; Rice, Wexler & Cleave, 1995).

Overall, it is not clear whether different measures would have revealed systematic differences in the subgroups or whether the subgroupings reflected in part some false distinctions. Perhaps this similarity in outcomes across LI subgroups is another example of the gross quantitative nature of these measures. Alternatively, however, Leonard (1998) cogently discusses the arbitrariness of cognitive cutoff scores used to define SLI. Perhaps the failure to distinguish subgroups in the present study reflected the artificial distinction that cutoffs in cognitive scores render in research designs. If, for example, two standard deviations below the mean on Triangles had been used as an acceptable lower range, all of the children in the LOW subgroup would have moved into the SLI designation. Plante (1998) recently recommended ruling out frank mental retardation when determining SLI, but then not further restricting the IQ range. If we could consider as SLI children with a primary deficit in language who are not diagnosed as educably mentally impaired (EMI), then researchers would be able to increase subject sample sizes and relate more closely to the caseloads of most practicing speech-language pathologists. This issue is of special import for the study of African American children because IQ and achievement tests discriminate against minorities. Accordingly, a –1 standard deviation cut-off on most IQ tests may be too restrictive and may artificially reduce prevalence estimates of SLI in African American children while inflating the numbers of children considered borderline or low normal intellectually. Considerable future research is needed to address prevalence issues in this population.

In summary, this investigation explored whether a set of fairly traditional measures, grounded in our prior research program, would distinguish dialect from disorder in a population of school-age African American children. The findings indicate that all measures were sensitive to language impairment and suggest that they should be pursued in future research. This study represents an important first step in the identification of African American children with language impairments, including those with SLI, generally considered a baseline language impairment (Leonard, 1998). In future research it will be important to establish normative statements to improve our ability to interpret individual performance data. Findings from the present investigation will also permit researchers to pursue the underlying nature of language disorders identified with this battery for this population. Future research addressing these issues will significantly improve our ability to include typically developing African American children in research programs and contribute to the clinical management of those with language disorders.

Acknowledgments

This investigation was supported by Research Grant 1 RO1 DC 02313-01A1 from the National Institute on Deafness and Other Communication disorders. The authors thank the administrators, staff, and students in the Oak Park, MI, public schools for their participation and Carol McDonald Connor for helpful comments on the manuscript.

References

American National Standards Institute. (1989). *Specifications for audiometers* (ANSI S3.6-1989). New York: Author.

Aram, D., & Nation, J. (1975). Patterns of language behavior in children with developmental language disorders. *Journal of Speech and Hearing Research, 18,* 229–241.

Aram, D., Morris, R., & Hall, N. (1993). Clinical and research congruence in identifying children with specific language impairment. *Journal of Speech and Hearing Research, 36,* 580–591.

Archer, E. J. (1960). A re-evaluation of the meaningfulness of all possible CVC trigrams. *Psychological Monographs, 74*(Whole No. 497).

Bedore, L., & Leonard, L. (1998). Specific language impairment and grammatical morphology: A discriminant function analysis. *Journal of Speech, Language, and Hearing Research, 41,* 1185–1192.

Bishop, D. (1979). Comprehension in developmental language disorders. *Developmental medicine and child neurology, 21,* 225–238.

Boykin, A. (1977). Experimental psychology from a Black perspective: Issues and examples. In W. Cross (Ed.), *Final report from the third conference on empirical research in*

Black psychology (pp. 53–77). Washington, DC: National Institute of Education.

Bracken, B. A. (1986). *Bracken concept development program.* San Antonio, TX: Psychological Corporation.

Campbell, T., Dollaghan, C., Needleman, H., & Janosky, J. (1997). Reducing bias in language assessment: Processing-dependent measures. *Journal of Speech, Language, and Hearing Research, 40,* 519–525.

Chapman, R. (1988). Language acquisition in the child. In N. Lass, L. McReynolds, J. Northern, & D. Yoder (Eds.), *Handbook of speech-language pathology and audiology* (pp. 120–124). Toronto: B. C. Decker.

Cole, P.A., & Taylor, O. L. (1990). Performance of working-class African American children on three tests of articulation. *Language, Speech, and Hearing Services in Schools, 21,* 171–176.

Conti-Ramsden, G., & Jones, M. (1997). Verb use in specific language impairment. *Journal of Speech, Language, and Hearing Research, 40,* 1298–1313.

Craig, H., & Washington, J. (1994). The complex syntax skills of poor, urban, African American preschoolers at school entry. *Language, Speech, and Hearing Services in Schools, 25,* 181–190.

Craig, H., Washington, J., & Thompson-Porter, C. (1998a). Performances of young African American children on two comprehension tasks. *Journal of Speech, Language, and Hearing Research, 41,* 445–457.

Craig, H., Washington, J., & Thompson-Porter, C. (1998b). Average C-unit lengths in the discourse of African American children from low-income, urban homes. *Journal of Speech, Language, and Hearing Research, 41,* 433–444.

Dollaghan, C., & Campbell, T. F. (1998). Nonword repetition and child language impairment. *Journal of Speech, Language, and Hearing Research, 41,* 1136–1146.

Dunn, L. (1959). *The Peabody Picture Vocabulary Test.* Circle Pines, MN: American Guidance Service.

Dunn, L., & Dunn, L. (1997). *The Peabody Picture Vocabulary Test III (PPVT–III).* Circle Pines, MN: American Guidance Service.

Dunn, L., & Dunn, L. (1981). *The Peabody Picture Vocabulary Test–Revised.* Circle Pines, MN: American Guidance Service.

Fagundes, D. D., Haynes, W. O., Haak, N. J., & Moran, M. J. (1998). Task variability effects on the language test performance of southern lower soicioeconomic class African American and Caucasian five-year-olds. *Language, Speech, and Hearing Services in Schools, 29,* 148–157.

Fudala, J. B., & Reynolds, W. M. (1986). *Arizona Articulation Proficiency Scale: Second Edition.* Los Angeles, CA: Western Psychological Services.

Johnston, J., & Kamhi, A. (1984). Syntactic and semantic aspects of the utterances of language-impaired children: The same can be less. *Merrill-palmer quarterly, 30,* 65–85.

Kaufman, A., & Kaufman, N. (1983). *Kaufman Assessment Battery for Children: Interpretive manual.* Circle Pines, MN: American Guidance Service.

Kayser, H. (1996). Cultural/linguistic variation in the United States and its implications for assessment and intervention in speech-language pathology: An epilogue. *Language, Speech, and Hearing Services in Schools, 27,* 385–386.

Lahey, M. (1988). *Language disorders and language development.* New York: MacMillan.

Leonard, L. (1989). Language learnability and specific language impairment in children. *Applied Psycholinguistics, 10,* 179–202.

Leonard, L. (1998). *Children with specific language impairment.* Cambridge: MIT Press.

Lidz, C., & Pena, E. (1996). Dynamic assessment: The model, its relevance as a nonbiased approach, and its application to Latino American preschool children. *Language, Speech, and Hearing Services in Schools, 27,* 367–372.

Loban, W. (1976). *Language development: Kindergarten through grade twelve.* Urbana, IL: National Council of Teachers of English.

MacWhinney, B. (1994). *The CHILDES project: Tools for analyzing talk* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Meisels, S. (1989). Can developmental screening tests identify children who are developmentally at risk? *Pediatrics, 83,* 578–585.

Miller, J. (1981). *Assessing language production in children.* Baltimore, MD: University Park Press.

Miller, J., & Chapman, R. (1981). The relation between age and mean length of utterance in morphemes. *Journal of Speech and Hearing Research, 24,* 154–161.

Miller-Jones, D. (1989). Culture and testing. *American Psychologist, 44,* 360–366.

Morehead, D., & Ingram, D. (1973). The development of base syntax in normal and linguistically deviant children. *Journal of Speech and Hearing Research, 16, 330–352.*

Parnell, M., Patterson, S., & Harding, M. (1984). Answers to Wh-questions: A developmental study. *Journal of Speech and Hearing Research, 27,* 297–305.

Plante, E. (1998). Criteria for SLI: The Stark and Tallal legacy and beyond. *Journal of Speech and Hearing Research, 41,* 951–957.

Rice, M. L., & Wexler, K. (1996). Toward tense as a clinical marker of specific language impairment in English-speaking children. *Journal of Speech and Hearing Research, 39,* 1239–1257.

Rice, M. L., Wexler, K., & Cleave, P. L. (1995). Specific language impairment as a period of extended optional infinitive. *Journal of Speech and Hearing Research, 38,* 850–863.

Semel, E., Wiig, E., & Secord, W. (1987). *Clinical Evaluation of Language Fundamentals–Revised.* San Antonio: The Psychological Corporation.

Stark, R., & Tallal, P. (1981). Selection of children with specific language deficits. *Journal of Speech & Hearing Disorders, 46,* 114–122.

Tomblin, J. B., Records, N. L., Buckwalter, P., Zhang, X., Smith, E., & O'Brien, M. (1997). Prevalence of specific language impairment in kindergarten children. *Journal of Speech, Language, and Hearing Research, 40,* 1245–1260.

Tomblin, J. B., Records, N. L., & Zhang, X. (1996). A system for the diagnosis of specific language impairment in kindergarten children.. *Journal of Speech, Language, and Hearing Research, 39*, 1284–1294.

Washington, J., & Craig, H. (1992). Articulation test performance of low-income, African American preschoolers with communication impairments. *Language, Speech, and Hearing Services in Schools, 23*, 203–207.

Washington, J., & Craig, H. (1994). Dialectal forms during discourse of urban, African American preschoolers living in poverty. *Journal of Speech and Hearing Research, 37*, 816–823.

Washington, J., & Craig, H. (1998). Socioeconomic status and gender influences on children's dialectal variations. *Journal of Speech, Language, Hearing Research, 41*, 618–626.

Washington, J., & Craig, H. (1999). Performances of at-risk, African American preschoolers on the Peabody Picture Vocabulary Test–III. *Language, Speech, and Hearing Services in Schools, 30*, 75–82.

Washington, J., Craig, H., & Kushmaul, A. (1998). Variable use of African American English across two language sampling contexts. *Journal of Speech, Language, Hearing Research, 41*, 1115–1124.

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, Language, Hearing Research, 38*, 1349–1355.

Received November 12, 1998

Accepted September 15, 1999

Contact author: Holly K. Craig, PhD, University of Michigan, 1111 E. Catherine St. Ann Arbor, MI 48109-2054. Email: hkc@umich.edu