Exploring the Utility of Narrative Analysis in Diagnostic Decision Making: Picture-Bound Reference, Elaboration, and Fetal Alcohol Spectrum Disorders

John C. Thorne
Truman E. Coggins
Heather Carmichael Olson
Susan J. Astley
University of Washington

Purpose: To evaluate classification accuracy and clinical feasibility of a narrative analysis tool for identifying children with a fetal alcohol spectrum disorder (FASD).
Method: Picture-elicited narratives generated by 16 age-matched pairs of school-aged children (FASD vs. typical development [TD]) were coded for semantic elaboration and reference strategy by judges who were unaware of age, gender, and group membership of the participants. Receiver operating characteristic (ROC) curves were used to examine the classification accuracy of the resulting set of narrative measures for making 2 classifications: (a) for the 16 children diagnosed with FASD, low performance (n = 7) versus average performance (n = 9) on a standardized expressive language task and (b) FASD (n = 16) versus TD (n = 16).
Results: Combining the rates of semantic elaboration and pragmatically inappropriate reference perfectly matched a classification based on performance on the standardized language task. More importantly, the rate of ambiguous nominal reference was highly accurate in classifying children with an FASD regardless of their performance on the standardized language task (area under the ROC curve = .863, confidence interval = .736–.991).
Conclusion: Results support further study of the diagnostic utility of narrative analysis using discourse level measures of elaboration and children's strategic use of reference.

KEY WORDS: assessment procedures, pragmatics, discourse analysis, language sample analysis, diagnostics

Narrative discourse permeates our social lives from an early age, making it a critical area to address in the measurement of language abilities. For more than 2 decades, narrative analysis has been recommended as an ecologically valid way to assess the production of meaningful language in socially integrated discourse (see Owens, 1999). Underlying this recommendation is the assumption that narrative analysis provides a more integrated appraisal of a child's communicative abilities than is possible via standardized language measures (Adams, Lloyd, Aldred, & Baxendale, 2006; Botting & Adams, 2005; Culatta, Page, & Ellis, 1983; Norbury & Bishop, 2003; Wagner, Nettelbladt, Sahlen, & Nilholm, 2000). Thus, narrative analysis should be able to identify children with meaningful communicative impairments that might be missed using conventional standardized assessment instruments. The current study examined this largely untested assumption by retrospectively comparing narratives produced by school-aged children with a fetal alcohol spectrum disorder (FASD) with those produced by age- and gender-matched typically developing (TD) peers. Comparisons were made via the Semantic Elaboration Coding System[1] (Thorne, 2004), which systematically measures the use of pragmatically appropriate strategies of reference and the semantic elaboration of concepts.

FASDs and Language

Children with prenatal alcohol exposure exhibit a wide range of abilities across all body systems (Astley & Clarren, 2000; Carmichael-Olson, Morse, & Huffine, 1998; Streissguth, 1997). When specific growth, facial, and central nervous system impairments are present within a well-specified range, a diagnosis from within the continuum of FASD can be rendered (see Astley, 2004; Chudley et al., 2005). Because the development and use of language have been reported to be affected by high levels of prenatal alcohol exposure (Mattson & Riley, 1998; Streissguth, Barr, Kogan, & Bookstein, 1996), measurement of language ability has been an important feature of interdisciplinary assessment of these individuals. The preponderance of evidence regarding language behavior in children with an FASD has been gathered using standardized, norm-referenced tests (Abkarian, 1992; Becker, Warr-Leeper, & Leeper, 1990; Carney & Chermak, 1991; Church, Eldis, Blakley, & Bawle, 1997; Church & Kaltenbach, 1997; Gentry et al., 1998; Janzen, Nanson, & Block, 1995; Weinberg, 1997). The goal of these studies has been to establish how well children with an FASD comprehend and/or produce language structures in standardized contexts. Typically, these contexts measure language using discrete responses at or below the level of single-sentence utterances. Although the results have revealed an array of performance deficits, no core deficit profile has emerged.

Because no recognizable deficit profile has resulted from research using standardized language tests, researchers have begun to look at suprasentential discourse in school-aged children diagnosed with an FASD. In preliminary research, discourse level deficits have been documented in children with an FASD including reduced ability to provide sufficient information for listeners both during conversations (Hamilton, 1981) and in narratives (Coggins, Friet, & Morgan, 1998; Coggins, Olswang, Carmichael-Olsen, & Timler, 2003). In addition, caregivers report that children with an FASD often fail to accommodate the perspectives of others during interaction (Timler, Olswang, & Coggins, 2005). This early research suggests that despite widely variable performance on standardized tests, children with an FASD may have difficulty producing integrated extended discourse that requires them to balance linguistic and social–cognitive task demands (Coggins et al., 2003). This emerging profile, when coupled with our ability to identify an FASD independent of communication ability, makes this heterogeneous group of children an ideal population to test the discriminative utility of a narrative analysis system.

Narrative Analysis

As a primary form of extended discourse, narratives provide children with a means of verbally recapitulating experiences (Bishop & Edmundson, 1987; Feagans & Appelbaum, 1986; Feagans & Short, 1984) and are an important source of knowledge about inference, social cognition, and perspective taking (Owens, 1999). The ability to produce contextually integrated extended discourse is difficult to measure using the discrete responses typical of standardized tests. Analysis of narrative samples offers a viable alternative.

Unlike standardized measures, narrative analysis allows for measurement of discourse level parameters of communication that result directly from the pragmatics of a relatively communicative interaction (Owens, 1999). These parameters of behavior manifest in the history of concepts as they are developed across sentences in the narrative text and should provide information regarding language ability that is unavailable in the noncommunicative context of standardized testing. Arguably, the most informative context in which to sample children's narrative ability is one that obligates them to organize and generate narratives without an adult model or other contextual supports (Curenton & Justice, 2004; Juncos-Rabadan, Pereiro, & Rodriguez, 2005; Norbury & Bishop, 2003). This decontextualized narrative discourse stresses the language system by limiting the nonlinguistic tools available during discourse. These limitations determine the type of discourse breakdowns that can be predicted in children with compromised cognitive systems. The Semantic Elaboration Coding System (Thorne, 2004) was designed to capture these predictable discourse level behaviors in school-aged children.

The Semantic Elaboration Coding System

The Semantic Elaboration Coding System implements a framework for narrative analysis based upon cognitive linguistics (Croft & Cruse, 2004; Langacker, 1991; Talmy, 2000b; Tomasello, 2003). *Cognitive linguistics* seeks to account for structural properties of language in terms of its relation to more general conceptual structures and functions. It has, therefore, examined "the linguistic structuring of basic ideational and affective categories attributed to cognitive agents, such as attention, perspective, volition, and intention, and expectation and effect" (Talmy, 2000b, Vol. I, p. 3).

[1]An unpublished training manual for the Semantic Elaboration Coding System is available from the first author via jct6@u.washington.edu.

In a detailed look at the conceptual structuring of narrative, Talmy (2000a) identified a series of conceptual parameters in narrative that constitute a "set of organizing principles that apply in common across all major cognitive systems" (p. 422). The Semantic Elaboration Coding System is organized along two of these parameters that would be expected to vary monotonically with quality in a decontextualized narrative. The first involves the strategic use of linguistic reference to assure that concepts are explicit and uniquely identifiable in the text. The second involves the degree to which semantic concepts are elaborated or well specified in the text.

The strategic use of reference in narrative. In any narrative, it is essential that the concepts involved (both entities and events) are kept distinct from each other to reduce ambiguity. The linguistic strategies used to make distinctive reference to various concepts in a narrative may vary in form from semantically complete phrases and clauses to semantically ambiguous forms like pronouns dependent upon the presuppositions the narrator has about the listener's current knowledge and attention state regarding those concepts. Wong and Johnston (2004) identify three basic *reference functions* in narrative tasks related to presuppositions about the knowledge and attention state of the listener: (a) the *introduction* of new concepts into the discourse (presupposes no knowledge of the concept), (b) the *maintenance* of foreground/in-focus concepts (presupposes both knowledge of and attention to the concept), and (c) the *reintroduction* of previously introduced background/out-of-focus concepts in the discourse (presupposes knowledge of, but limited attention to, the concept). The distinction between adequate and inadequate use of various strategies for meeting these discourse functions cannot be made without consideration of both the textual and extratextual context of the particular instance of use (see Cornish, 1999, for a discussion; see also Levine & Klin, 2001; Maratsos, 1976; van Hoek, 1997; Wong, 2001; Wong, Au, & Stokes, 2004).

In decontextualized narrative discourse, strategies for meeting all three basic reference functions are restricted to the linguistic code (see Halliday & Hasan, 1976). To maintain unambiguous reference, storytellers must use discourse strategies that do not presuppose unwarranted knowledge or attention on the part of their listener or require extralinguistic support to be interpreted meaningfully. This can be particularly challenging for younger storytellers as they continuously adapt their narratives to the ever-changing knowledge and attention states of their listeners (Coggins et al., 1998; Cornish, 1999; Lewis, 2004; Wong & Johnston, 2004).

School-aged children are learning to effectively incorporate a variety of reference strategies into their language production (Stephens, 1988), allowing differentiation between children with typical and delayed language development (see Liles, Duffy, Merritt, & Purcell, 1995; Wong, 2001, for example). Measurement of the strategic use of reference in narrative serves as a primary component of the Semantic Elaboration Coding System.

Semantic elaboration. Decontextualized discourse also demands a greater density of ideas, or *semantic elaboration,* from the storyteller. An analysis of narrative elaboration must account for the contribution of particular words or syntactic structures to the listener's growing conceptualization of a concept in a way that accounts for the history of that concept in the preceding discourse (Croft & Cruse, 2004; Fauconnier, 2004; Klin, Weingartner, Guzman, & Levine, 2004; Talmy, 2000b). This requires that the measurement of elaboration be integrated with the measurement of successful reference to those concepts because a structure cannot contribute to the elaboration of a concept if it does not unambiguously make reference to that concept.

Investigators have heretofore used a variety of lexically based and syntactically based measures to capture elaboration in discourse. These measures have differentiated children with different overall language ability as measured by standardized language tests (Condouris, Meyer, & Tager-Flusberg, 2003; Hammer, Yont, & Tomblin, 2005; Loban, 1976). Most approaches treat the structures they quantify independent of the textual history of the concepts they describe, much as is done with standardized tests of lexical knowledge or syntactic competence. Consequently, they are unlikely to provide information regarding language ability beyond that available through standardized testing (a largely untested assumption, but see Hesketh, 2004). A system that integrates measurement of elaboration and reference strategy may provide information about integrated language abilities that is inaccessible to the more traditional approaches.

Purpose

The Semantic Elaboration Coding System is designed to be used with decontextualized narratives produced for a naive listener by school-aged children as they look through the wordless picture book, *Frog Where Are You?* (Mayer, 1969). It integrates analysis of two narrative discourse parameters that would be difficult to quantify using standardized measures: (a) *ambiguity,* the use of inappropriate strategies of reference, and (b) *elaboration,* the semantic elaboration of concepts as they develop across the narrative (see Thorne & Coggins, 2004; see also Thorne & Coggins, 2005).

With respect to children diagnosed with an FASD, the Semantic Elaboration Coding System must be able to (a) identify children with an FASD who perform poorly on a standardized test to establish a level of concurrent validity with that standardized test and (b) identify those children with an FASD who perform like typically developing (TD) children on the standardized language test to establish superior classification accuracy when compared with that test. This study was designed to test the potential for the Semantic Elaboration Coding System to make these key discriminations. Specifically, the research questions under study were as follows:

1. Can narrative analysis using the Semantic Elaboration Coding System correctly classify a group of school-aged children into separate groups based on typical development versus an identified FASD?
2. Can narrative analysis using the Semantic Elaboration Coding System accurately predict which children with an identified FASD have either average or low performance on a standardized language task?
3. Which specific measure or combinations of measures from within the Semantic Elaboration Coding System are most accurate in performing these discrimination tasks, and, therefore, reasonably warrant further development?

Method

Participants

Thirty-two school-aged children from two previous studies (Carmichael-Olson & Astley, 2005; Coggins, 1995) participated. They ranged in age from 8;5 years to 11;7 years (M = 9;11 years) and presented a range of socioeconomic and ethnic profiles. Sixteen of the children presented key clinical features consistent with an FASD while the remaining children were considered TD.

Children with an FASD. The 16 FASD participants had a diagnosis of either (a) full or partial fetal alcohol syndrome or (b) a confirmed alcohol exposure accompanying static encephalopathy or neurobehavioral disorder. Diagnosis was performed by an interdisciplinary team at the University of Washington Fetal Alcohol Syndrome Diagnostic and Prevention Network. All children were originally diagnosed using the 1999 version of the 4-Digit Diagnostic Code (Astley & Clarren, 1999). All codes were translated into the 2004 version of the 4-Digit Diagnostic Code (Astley, 2004) to provide an up-to-date diagnostic standard for comparison. Participants reflected “the true diversity and continuum of disability associated with prenatal alcohol exposure” (Astley, 2004, p. 13). Table 1 specifies each participant’s 4-Digit Diagnostic Code, which provides information regarding their growth, facial morphology, brain development, and alcohol exposure (see Astley, 2004, for details of code interpretation).

Existing nonverbal and verbal measures provided an additional basis for selection and are also provided in Table 1. The Matrices subtest from the Kaufman Brief Intelligence Test (Kaufman & Kaufman, 1990) provided an overall measure of nonverbal problem solving, with participants excluded based on a score 1.5 standard deviations below the mean (range: 79–130; M = 101). The 16 participants in the FASD group were dichotomized into two performance groups based on their scores from the Re-Creating Sentences subtest of the Test of Language Competence (RS-TLC; Wiig & Secord, 1989): (a) an average-performance group (n = 9) with standard scores within one standard deviation of the mean (between 7 and 10) and (b) a low-performance group (n = 7) with standard scores two or more standard deviations below the mean (between 3 and 4).

The resulting sample included 9 females and 7 males. Family income for the group ranged from $15,000 to $220,000 per annum ($M$ = $88,000, Mdn = $75,000). The group included 11 children identified as Caucasian, 3 as bi- or multiracial, and 1 each as African American and Native American. Only 3 of the children were still living with their biological parent(s) at the time data were collected. The remaining 13 were in adoptive or legal guardianship placements (5 with relatives).

TD peers. Each participant with FASD was paired with a TD peer matched on chronological age (±12 months, mean difference = 3.5 months). Thirteen TD age-matched peers also matched the gender of their FASD counterpart (15 females, 17 males). Table 1 displays age and gender for all participant pairs.

The TD aged-matched peers were recruited from elementary schools representing two school districts in the greater metropolitan Seattle area. Median family incomes were similar across school districts ($61,435–$62,195). The sample included 12 children identified as Caucasian, 2 as Asian, and 1 each as African American and Hispanic (representative of the home county for both districts).

No intelligence or standardized language measures were available for TD participants. However, a school psychologist familiar with the 16 children and with the profile of FASD screened school records for each child with respect to school performance, social ability, and general behavior. Based on this review of available records, each was judged to be following a typical developmental course due to their unremarkable behavior and adequate school achievement. The TD participants did not undergo the same interdisciplinary assessment as the children with FASD.

Materials

Self-generated, decontextualized narratives were selected from two independent databases: one from a study

Table 1. Participant characteristics: diagnosis, language performance group, test scores, age, and gender.

Diagnostic Code & Category[a]	Re-Creating Sentences—Test of Language Competence[b]	Kaufman Brief Intelligence Test—Matrices[c]	FASD Group		TD Group	
			Gender	Age	Gender	Age
2443 = A	3: low	79	F	8;9	F	8;8
1234 = F	3: low	90	M	9;2	M	8;11
3432 = B	3: low	95	F	8;11	F	9;3
2444 = A	3: low	80	M	10;8	M	10;11
4234 = E	4: low	80	F	10;6	M	9;9
4344 = C	3: low	96	M	11;1	M	11;4
1324 = G	3: low	87	F	11;2	M	11;6
1124 = H	7: average	92	F	8;5	F	8;4
1124 = H	7: average	128	M	8;8	M	8;9
1224 = H	7: average	130	M	8;10	M	8;11
3344 = C	7: average	101	M	9;3	M	9;1
1224 = H	8: average	114	F	9;5	F	9;2
1223 = H	9: average	113	F	10;6	F	9;6
1223 = H	9: average	122	F	10;6	F	10;10
1224 = H	7: average	98	F	11;2	M	11;5
3233 = E	10: average	105	M	11;5	M	11;7
Mean	Average: 8; low: 3	101		9;11	9;10	

Note. Diagnostic code provides information, from left to right, regarding growth, facial morphology, brain development, and alcohol exposure. Scores range from 1 (*unremarkable*) to 4 (*severe*). FASD = fetal alcohol spectrum disorder; TD = typical development; F = female; M = male; FASD categories: A = FAS (alcohol exposed); B = FAS (alcohol exposure unknown); C = partial FAS (alcohol exposed); E–H indicate the remaining FASD categories (with confirmed alcohol exposure). Details for interpretation of the 4-Digit Diagnostic Code of FASD can be found in Astley (2004).

[a]Astley (2004). [b]*M* = 10, *SD* = 3. [c]*M* = 100, *SD* = 15.

involving children diagnosed with an FASD (Carmichael-Olson & Astley, 2005) and the second from a normative study of TD school-aged children (Coggins, 1995).

Procedures

All narratives from the two respective databases were elicited using *Frog Where Are You?* (Mayer, 1969). In both studies, participants were tested individually and received the same instructions. Each child was instructed to look through Mayer's book to become familiar with the story line. When the child completed previewing the story, the examiner exhorted the participant to tell the best story possible while using the picture book as a visual prompt. In each case, examiners were seated across the room from the child, with the storybook out of their line of sight.

Transcription and Coding

Narratives were recorded on audiocassette and orthographically transcribed by trained graduate students. The second author supervised the narrative collection and transcription process and then stripped all transcripts of identifying information while assigning each a random code so that relevant information could be retrieved for later data analysis.

Transcripts were coded by the first author using the Semantic Elaboration Coding System (Thorne, 2004). The system assigns codes along the parameters of (a) ambiguity, the consequence of inappropriate reference strategies, and (b) elaboration of concepts.

Ambiguity. Operationally, measurement of ambiguity involved coding references to concepts as either unambiguous or ambiguous. To reduce the number of coding categories, ambiguous anaphoric reference strategies for maintenance and reintroduction of concepts and ambiguous introduction strategies that make unwarranted presuppositions about listener knowledge and attention were collapsed into just two ambiguity categories (nominal and pronominal) because their use has similar impact on a listener's discourse processing (Cornish, 1999; van Hoek, 1997).

Elaboration. Measurement of elaboration involved core lexical items that unambiguously introduced concepts into the story that were coded as either schematic (i.e., minimally characterized) or elaborated. Additional words that helped elaborate concepts were also coded.

For this study, each word in a transcript was assigned 1 of 10 mutually exclusive scoring codes along these two parameters, or a null code. The 10 scoring codes are presented later with a brief definition of each category. Further details can be found in the Semantic Elaboration Coding System (Thorne, 2004).

Ambiguity codes. Two codes identified unambiguous anaphoric reference to concepts and made a distinction between nominal forms and pronominal forms serving maintenance and reintroduction functions.

1. *Nominal reference (NR):* an unambiguous nominal form used to maintain or reintroduce a concept previously introduced into the discourse (e.g., *a dog* introduced into the narrative later referred to unambiguously as *the dog*).
2. *Pronoun reference (PR):* an unambiguous pronominal form used to maintain or reintroduce a concept previously introduced into the discourse (e.g., *a dog* later referred to unambiguously as *he* or *it*).

Two additional codes were used to identify cases in which reference was ambiguous.

3. *Ambiguous nominal reference (ANR):* an ambiguous use of a nominal form attempting to introduce, maintain, or reintroduce a concept.
4. *Ambiguous pronoun reference (APR):* an ambiguous use of a pronominal form attempting to introduce, maintain, or reintroduce a concept.

Semantic elaboration codes. Six codes were used to quantify semantic elaboration. Four of these codes dichotomized core lexical items as either schematic or elaborated. Two codes identified core lexical units as schematic.

5. *Schematic verb (SV):* a word that introduced basic information regarding an event into the discourse (i.e., the *fact* that something happened: *went, got, is, going*).
6. *Schematic nominal (SN):* a word that introduced basic information regarding an entity into the discourse (e.g., *boy, dog, frog, jar, animal, thing*).

Two additional codes identified those core lexical items that were relatively elaborated.

7. *Elaborated verb (EV):* a word that introduced elaborated information regarding an event into the discourse (i.e., the manner in which something happened: *ran, fell, chased, yelled*).
8. *Elaborated nominal (EN):* a word that introduced elaborated information regarding an entity unambiguously into the discourse (e.g., *Timmy, elk, bullfrog, wife, bumblebee*).

Two final codes identified word forms associated with these core lexical items that provided additional semantic elaboration. These two codes are the most frequent in the Semantic Elaboration Coding System, making up a significant portion of the information the system gathers.

9. *Verb satellite (VS):* a word providing elaborating information about a verb (e.g., went *away,* ran *quickly, in the morning when* he got *up*).
10. *Nominal modifier (NM):* a word providing elaborating information regarding a nominal (e.g., *big mad* owl, frog *that ran away*).

A null code was used to indicate that a word did not fit any of the 10 scoring categories.

Null code (null): a word not meeting operational definitions for any ambiguity or elaboration code category in the system.

Analysis

Intercoder agreement. A graduate student in speech and hearing sciences was recruited and trained to function as a secondary coder. Coder competence was established when intercoder agreement between the primary (the first author) and secondary coder reached a kappa of .7 or better for each code in the system on a set of five training narratives taken from the CHILDES databank (MacWhinney, 2000).

The primary coder then scored all 32 of the study narratives while the secondary coder independently scored 25% of the narratives (n = 8) randomly selected using SPSS for Windows (SPSS, 1998). Random selection and all coding were completed before either coder knew the diagnostic status, age, or gender of the storytellers. The resulting sample contained 5 narratives from the TD group and three from the FASD group.

For both training and study narratives, kappa was calculated as a measure of agreement between coders for each of the 10 Semantic Elaboration Coding System codes and for the null code. Calculation of kappa was conducted separately for each code, with agreement based on a binary decision—every word in the narratives was identified as carrying the designated code or not (following Kraemer, n.d.; see also Bakeman & Gottman, 1997). All words not coded by both judges with the designated code were treated as disagreements in the calculation of kappa for that code. Because it reveals performance for individual codes and avoids overly optimistic and difficult-to-interpret estimates of overall agreement that can occur in multicode calculations of kappa, this is a conservative method of estimating intercoder agreement (Acklin, McDowell, Verschell, & Chan, 2000; see also Kraemer, Periyakoil, & Noda, 2004). Results are presented in Table 2.

Eleven kappa statistics were computed. The precision of 10 kappa scores ranged from substantial (i.e., .6–.8) to almost perfect (i.e., >.8; Kraemer, Periyakoil, & Noda,

Table 2. Interrater agreement (kappa) for all Semantic Elaboration Coding System codes.

Code	κ
VS	.836
SV	.893
EV	.935
NM	.653
NR	.876
SN	.793
EN	.792
PR	.767
ANR	.764
APR	.540
Null (no code)	.836

Note. Calculation of the kappa was conducted separately for each code, with agreement based on a binary decision: Every word in the narratives was identified as carrying the designated code or not. All words not coded by both judges with the designated code were treated as disagreements in the calculation of the kappa for that code. VS = verb satellite; SV = schematic verb; EV = elaborated verb; NM = nominal modifier; NR = nominal reference; SN = schematic nominal; EN = elaborated nominal; PR = pronoun reference; ANR = ambiguous nominal reference; APR = ambiguous pronoun reference.

2004; Landis & Koch, 1977). The kappa statistic for the coding of ambiguous pronoun reference was in the moderate range of precision (i.e., .4–.6). Because it increases chances of Type II error, the moderate level of intercoder agreement on this single code might be considered insufficient for clinical application (cf. Bakeman & Gottman, 1997; Cicchetti & Sparrow, 1981), highlighting the need for attention to coder training and measurement stability in subsequent development of the system (see the Discussion section for more on pronouns and intercoder agreement).

Data preparation. Several preparatory steps were taken to ready the data for analysis. First, the total number of words (TW) was calculated for each narrative (excluding mazes, range: 154–796; M = 308) using SALT (Miller, 2004). Second, the raw frequencies of each of the 11 scoring codes in the 32 narratives were calculated also using SALT. Next, the raw frequencies of codes for APR and ANR were combined to create a summary ambiguity score, while the raw frequencies of codes for EV forms, EN forms, VS forms, and NM forms were combined to create a summary elaboration score.

Next, code and score rates were computed by dividing each measure by the TW in the narrative. Because the Semantic Elaboration Coding System examines narratives on a word-by-word basis, the TW in a story was considered to best represent the length of the story and thereby became the denominator used in the calculation of code and score rates. Using a common denominator for the calculation of all rates also facilitated comparison between rates.[2] This process resulted in 26 narrative measures for analysis: 11 code frequencies, 2 summary scores, and 13 associated rates as a function of narrative length. Table 3 presents all 26 measures.

Classification accuracy analysis. To explore the accuracy of classification for each of the 26 measures, empirical classification rates including sensitivity (true positive rate), specificity (true negative rate), and efficiency (overall accuracy rate) were examined. Methods from signal detection theory based on these three measurement parameters were implemented to judge the relative potential of each of the 26 measures to match the classification of participants provided by the appropriate reference standard (Kraemer, 1988, 1992; Kraemer, Noda, & O'Hara, 2004; McFall & Treat, 1999). The two classifications of particular interest were as follows: *Classification 1*—the accuracy of each measure in classifying participants as members of the FASD or TD group; *Classification 2*—the accuracy of each measure in classifying participants diagnosed with an FASD as a member of the group with average performance or low performance on the RS-TLC.

More specifically, for both classifications, each of the 26 measures was analyzed using an empirical receiver operating characteristics (ROC) curve based on values obtained from the 32 narratives (for recent applications of this method, see Dwolatzky et al., 2003; see also Heilmann, Weismer, Evans, & Hollar, 2005). This analysis compared the classification of each measure against the classification by the appropriate reference standard, and provided an ROC curve plotting the sensitivity against 1 minus the specificity of each measure for all obtained values. The ROC curve, the area under the ROC curve (AUC), and the asymptotic significance, standard error, and 95% AUC confidence intervals were all calculated using SPSS.

Results

Criteria for a Reasonable Measure

The AUC is a widely accepted measure of overall accuracy (McFall & Treat, 1999). An index of effect size, the AUC can distinguish between tests that are random (AUC = 0.5), poorly accurate (0.5–0.7), moderately

[2]It is common to calculate proportional rates for referential terms based on total number of referential opportunities rather than total words. In the case of the Semantic Elaboration Coding System, this would mean dividing each ambiguity code frequency by the total number of NRs + PRs + ANRs + APRs + SNs + ENs. To assure that results using total word rates were not substantially skewed when compared with those using total opportunity rates, correlations and ROC curves were run for ambiguity code rates calculated using both methods. Rates were substantially correlated ($r > .9$), and AUC results were not significantly different ($p > .4$). AUC data are reported using rates calculated with TW to facilitate comparison between elaboration rates and ambiguity rates.

Table 3. Semantic Elaboration Coding System measures evaluated for classification accuracy.

Raw Code Frequencies	Summary Scores and Formula	Rates as a Function of TW
Ambiguity measure		
NR		NR/TW
PR		PR/TW
ANR		ANR/TW
APR		APR/TW
	AS = ANR + APR	AR = AS/TW
Elaboration measure		
EV		EV/TW
EN		EN/TW
VS		VS/TW
NM		NM/TW
	ES = EV + EN + VS + NM	ER = ES/TW
Schematic code		
SV		SV/TW
SN		SN/TW
Null		Null/TW

Note. TW = total words; AS = ambiguity score; AR = ambiguity rate; ES = elaboration score; ER = elaborate rate; Null = no code.

accurate (0.7–0.9), and highly accurate (0.9–1.0; see Swets, 1988). To be considered reasonable in the current analysis, a measure required an AUC with an asymptotic significance better than .02 and an AUC 95% confidence interval with a lower bound above 0.7. These criteria assured not only that the AUC was significantly different from a random test (i.e., AUC = 0.5) but also that it had at least a moderate chance of accurately classifying cases.

Sensitivity, specificity, and efficiency values were calculated at the best possible cut-points for measures obtaining an AUC with an asymptotic significance better than .02. A "best cut-point" was defined as the obtained value along the ROC curve with the highest efficiency (overall accuracy rate). If multiple cut-points had equivalent efficiency, the one with the highest sensitivity was chosen.

Data Presentation

Figure 1 presents AUC data for both of the tested classifications. The top portion of Figure 1 displays results for Classification 1—FASD versus TD group membership. The lower portion of Figure 1 displays results for Classification 2—performance grouping for the FASD participants: low performance versus average performance on the RS-TLC. Figure 1 includes asymptotic significance levels, AUC values, and 95% confidence intervals for each measure (i.e., test) that achieved an asymptotic significance better than .02 for each classification. The shaded regions of Figure 1 indicate the range used to determine that a particular test's confidence interval indicated reasonable accuracy. The four measures that reached the criteria for a reasonable measure are indicated by a double asterisk. Those measures not included on Figure 1 were not statistically different from a random test.

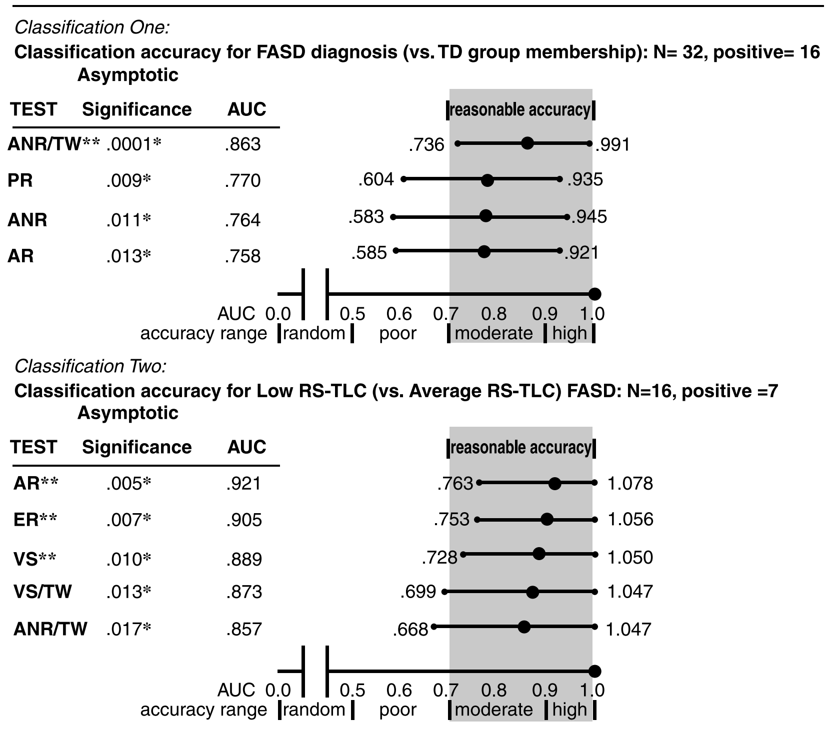
Table 4 displays information on the sensitivity, specificity, and efficiency at the best cut point for the most promising measures. These measures were chosen based on visual inspection of ROC curve shape (see Kraemer, 1992). The top of Table 4 includes data for Classification 1—TD versus FASD, while the lower portion displays data for Classification 2—RS-TLC performance group.

Accuracy for Classification 1—FASD Versus TD Group Membership

As can be seen in the top portion of Figure 1, the rate of ambiguous nominal reference (ANR/TW) obtained an AUC of .863, asymptotic significance = .0001. This AUC indicates classification accuracy solidly within the moderate-to-high accuracy range with a 95% confidence interval from .736 to .991.

As revealed in Table 4, the rate of ambiguous nominal reference (ANR/TW) achieves strong sensitivity, specificity, and efficiency for Classification 1. At the best possible cut point, >.0165 (i.e., greater than 1.65% of total words being ambiguous nominal references equals a positive test for FASD status), ANR/TW achieved a sensitivity of 87.5%, a specificity of 75%, and overall efficiency of 81.25%. At this particular cut point, it correctly

Figure 1. Area under the receiver operating characteristic (AUC) curve with 95% confidence intervals for measures asymptotically different than a random test. Analysis compares classification by Semantic Elaboration Coding System measures indicated and that by reference standard. For Classification 1, reference standard is diagnosis of an FASD by interdisciplinary team. For Classification 2, reference standard is based on average performance (within 1 *SD* of the mean) or low performance (< −2 *SD* below the mean) on the Re-Creating Sentences subtest of the Test of Language Competence (RS-TLC). FASD = fetal alcohol spectrum disorder; ANR = ambiguous nominal reference; TW = total words; PR = pronoun reference; AR = ambiguity rate; ER = elaboration rate; VS = verb satellite. *Asymptotic significance better than .02. **Reasonable test based on a lower bound of AUC confidence interval above 0.7. All values generated with SPSS.



classified 14 of 16 children from the FASD group and 12 of 16 from the TD group.

Post hoc analysis of ANR coding revealed that nominal references were most likely to be rated ambiguous when definite nominal forms were used to introduce or reintroduce concepts (>92% of 185 ambiguous nominal references). In a concordance search of all 32 stories, only 14 (<8%) nominals out of 185 that were rated ambiguous contained the indefinite articles *a* or *an*. Of these 14, 8 were in stories told by children with a diagnosis of FASD and 6 were in stories told by children in the TD group.

As seen in the top portion of Figure 1, the raw frequency of two codes, PR and ANR, achieved a significant point estimate of AUC in the moderately accurate range. However, both measures have AUC confidence intervals that fall below the criteria of .7 set for a reasonable measure. Because the calculation of the ROC curve is based on fewer values, the width of a confidence interval for AUC increases with smaller sample sizes and greater numbers of participants with tied scores. Both ANR and PR codes resulted in large numbers of tied scores relative to sample size. Twenty-nine participants shared their ANR frequency with at least 1 other participant, while 20 shared their PR frequency with at least 1 other participant. As a result, these measures had wide 95% confidence intervals for AUC that ranged below the criteria (see Figure 1). Notice in Table 4, however, that at its best cut-point, ANR was highly sensitive, specific, and efficient for Classification 1, correctly classifying 13 children from each group. While PR correctly classified an equal number from the FASD group, its relatively poor specificity makes it a less promising measure.

Accuracy for Classification 2—RS-TLC Performance for Participants With an FASD

AR, ER, and VS met study criteria for reasonable measures with high classification accuracy. As can be seen in the lower portion of Figure 1, a high AUC was

Table 4. Sensitivity, specificity, and efficiency of most promising measures at their best cut-point for both Classification 1 and Classification 2.

Measure	Cut-Point	Sensitivity (%)	Specificity (%)	Efficiency (%)
Classification 1: FASD vs. TD group membership				
ANR/TW	>.0165	87.50	75	81.25
PR	≤18.0	81.25	62.50	71.88
ANR	>5.0	81.25	81.25	81.25
Classification 2: RS-TLC performance group (FASD only)				
AR	>.0588	85.7	100	93.75
ER	≤0.4816	71.4	100	87.50
VS	≤89.0	100	66.7	81.25

Note. Best cut-point chosen as most efficient test weighing sensitivity over specificity for equally efficient tests. High sensitivity and a negative test help rule out diagnosis, which is useful for screening. High specificity and a positive test help to confirm a diagnosis. Higher efficiency indicates better overall diagnostic performance. RS-TLC = Re-Creating Sentences subtest of the Test of Language Competence.

achieved for the AR. The obtained value of .921, asymptotic significance = .005, places this result in the highly accurate range, with a 95% confidence interval ranging from .763 to 1.00. The lower portion of Table 4 shows that at the best cut-point, >.0588 (i.e., more than 5.88% of words being ambiguous indicates low RS-TLC performance group), this measure obtained a sensitivity of 85.7%, a specificity of 100%, and an efficiency of 93.75%. At this cut-point, the measure accurately classified 6 of 7 children in the low RS-TLC performance group and all 9 of those in the average RS-TLC performance group.

The AUC of the ER, as shown in the lower portion of Figure 1, is .905, asymptotic significance = .007. This AUC falls in the highly accurate range, with a 95% confidence interval ranging from .753 to 1.00. As seen in lower portion of Table 4, at the best cut-point, ≤.4816 (i.e., 48.16% or less of words being elaborators indicating low RS-TLC performance group), ER obtained a sensitivity of 71.4%, a specificity of 100%, and an efficiency of 87.5%. At this cut-point, the measure accurately classified 5 of 7 children in the low RS-TLC performance group and all 9 of those in the average RS-TLC performance group.

The total number of VS met study criteria for Classification 2. As can be seen in the lower portion of Figure 1, this measure obtained an AUC of .889, asymptotic significance = .01, placing it in the top of the moderately accurate range with a 95% confidence interval ranging from .728 to 1.00. The lower portion of Table 4 shows that at the best cut-point, ≤89 (89 or fewer VS indicating low RS-TLC performance group), this measure obtained a sensitivity of 100%, a specificity of 66.7%, and an efficiency of 81.25%. At this cut-point, the measure accurately classified 7 of 7 children in the low RS-TLC performance group and 6 of 9 children in the average RS-TLC performance group. This measure's high sensitivity but relatively modest specificity limits its diagnostic utility to that of a potential screening measure.

Combination of Classifiers

Based on their complementary distribution of classification errors, a combination of ER and AR was tested in a post hoc analysis to determine if combining these measures would improve classification accuracy for Classification 2. The two summary rates were combined using an *or* rule. A positive classification as belonging to the low RS-TLC performance group was obtained if a child had an ER less than or equal to 48.16% or an AR greater than 5.88%. This combination of measures perfectly predicted low RS-TLC performance group (vs. average RS-TLC performance group) with 100% sensitivity, specificity, and efficiency.

Discussion

This study investigated the classification accuracy of 26 measures generated using a new narrative analysis tool. The investigators explored ROC curves to identify the classification capabilities of the Semantic Elaboration Coding System as compared with two reference standards: (a) an interdisciplinary team diagnosis of an FASD and (b) performance grouping based on scores from a standardized expressive language measure, the RS-TLC (Wiig & Secord, 1989). The results are discussed for each topic in turn beginning with classification based on the performance of children with FASD on the RS-TLC.

Concurrent Validity of the Semantic Elaboration Coding System

The degree of elaboration a child uses in narrative discourse is a developmental skill (see Curenton & Justice, 2004; Eisenberg & Gillam, 2005; Loban, 1976; Scott, 1988) that is predictive of language impairment (see Greenhalgh & Strong, 2001; Johnston & Kamhi, 1984). Our results support these findings. The ER was a reasonable and highly accurate predictor of poor performance on the RS-TLC for school-aged children with FASD. At its best cut-point, ER was able to match classification based on the RS-TLC for 14 of 16 children (87.5%) into either low RS-TLC performance or average RS-TLC performance groups. In addition, we found that the summary AR was also a reasonable and highly accurate classifier of language performance group. At its best cut-point, AR matched the classification based on the RS-TLC for 15 of 16 children (93.75%). Post hoc analysis indicated that a logical combination of these two measures using an *or* rule correctly classified all 16 children with respect to their performance on the RS-TLC.

These results support the notion that narrative analysis should play an important role in diagnostic decision making. Classification using information from a narrative analysis perfectly matched that based on performance on a standardized measure of expressive language. Testing of the Semantic Elaboration Coding System against a wider range of standardized language measures to establish a better understanding of its concurrent validity appears warranted.

Classification Accuracy for FASD Diagnosis

A second goal of this feasibility research was to determine whether the Semantic Elaboration Coding System would provide the information needed for a more accurate classification than the standardized measure to support the idea that narrative analysis provides a more ecologically valid description of a child's ability to produce meaningful language in socially integrated discourse. This is clearly shown with the results of our between-groups classification.

One ambiguity measure satisfied strict accuracy criteria in matching the reference standard classification of study participants into FASD and TD groups. The rate of ANR (calculated as a function of narrative length) accurately identified participants previously identified by an interdisciplinary team as having an FASD regardless of the participant's performance on a standardized expressive language task. At its best value, the rate of ANR matched the interdisciplinary team classification for 14 of 16 (87.5%) children with an FASD, and identified 12 of 16 (75%) children with typical development. Moreover, this measure achieved a substantially reliable kappa statistic ($\kappa = .764$).

This finding is arguably the most important from the current study. The data demonstrate a procedure for reliably quantifying meaningful performance differences in the use of reference strategies that has the potential to be diagnostically informative for a population of children who have resisted easy classification using standardized language tests. In spite of the FASD group's wide range of cognitive and linguistic abilities, a single narrative measure keyed to the rate at which a child used inappropriate nominal reference strategies was able to match an interdisciplinary team diagnosis of an FASD for all but 2 children.

Ambiguous Nominal Reference by Children in the FASD Group

A closer look at the post hoc analysis of ambiguous nominal reference sheds light on the behaviors children in the FASD group were using in their narratives that was captured by the ANR code. For a reference to be rated ambiguous in the Semantic Elaboration Coding System, an equivocal word choice must occur. In other words, the storyteller has selected a word that fails in its mission to either introduce or unambiguously reference a concept. There are three basic conditions under which a nominal reference may be considered ambiguous in the coding system.

In Condition 1, an existing concept (e.g., THE BOY) is treated as if it were new by using an indefinite nominal form (e.g., *a boy*) to maintain or reintroduce it. Condition 1 involves pragmatically inappropriate use of an indefinite nominal. Post hoc analysis found this type of error to be rare (<8%) and evenly distributed between the TD and FASD groups. In Condition 2, ambiguity results during reintroduction or maintenance of a concept when the storyteller treats a nominal form new to the discourse as referentially equivalent to a form used previously in the discourse when, in fact, the two are not referentially equivalent. In Condition 3, ambiguity results when a new concept is treated as if it were familiar or already existing in the discourse by using a definite nominal to introduce it into the discourse (e.g., *the boy* used to introduce THE BOY or *the barking* used to introduce an event, BARKING). Conditions 2 and 3 both involve pragmatically inappropriate use of a definite nominal form. It was this inappropriate use of definite nominal forms for introduction, maintenance, or reintroduction of concepts that most commonly (>92%) led to a nominal reference being judged as ambiguous. Condition 2 is considered in detail first.

When children use a particular reference word, they are, in essence, making a categorical decision. That is, they are deciding that a particular entity (e.g., the glass container holding the boy's frog in Mayer's, 1969, story) is a member of a category that can be named with a particular nominal form, for example, *jar*. Certainly, *jar* is not the only possible name that could be used to reference the concept GLASS CONTAINER HOLDING FROG. It could reasonably be referred to as a *glass container* or a *glass frog holder,* for example. If the form *jar* is chosen, however, it will be the form used by listeners as the basis for their understanding of the concept GLASS CONTAINER HOLDING FROG.

Decontextualized narratives obligate the storyteller to find linguistic strategies for keeping entities and events distinct. Heim (as cited by van Hoek, 1997) likened each initial reference within a story to a file card that is introduced into a file catalogue (the narrative discourse). In this scheme, subsequent references to a particular event or entity access and potentially update the information on the appropriate file card (cf. Fauconnier, 2004; Levine & Klin, 2001). The pragmatics of English allow specific linguistic strategies to introduce new file cards into decontextualized discourse (indefinite forms) that are distinct from those for accessing (i.e., maintaining or reintroducing) file cards already present in the discourse (definite nominals and pronouns; see Croft, 2001; Klin et al., 2004; Langacker, 1991; Maratsos, 1976; van Hoek, 1997; Wong & Johnston, 2004).

Applying this analogy, if *jar* is used to introduce the entity concept GLASS CONTAINER, it will be the nominal form listed first on the listener's file card for that concept. It is not the only lexical form available on the file card, but it will be the one that most effectively refers to the concept. When storytellers who have introduced the concept into the story with the nominal complex *a jar* decide to maintain or reintroduce that same concept using a different nominal form, let's say *the bottle,* they are acting ostensively as if their listener will recognize the overlapping nature of JAR and BOTTLE (both are glass containers) and will assign reference appropriately. However, context plays a significant role in when and how the quite distinct categories JAR and BOTTLE are equivalent and when they are not.

In a shared visual context in which only one item fitting in the category GLASS CONTAINER is seen, visual information aids the listener in disambiguating a reference to that glass container whether the term *jar* or *bottle* is used. In a decontextualized discourse, however, that visual information is unavailable to the listener (despite being available to the storyteller). With only their developing conception of the entities involved in the story to support inferences about the referent, listeners may not be able to quickly and easily determine if *the bottle* being referred to is the same GLASS CONTAINER as *the jar* introduced into the story earlier.

When this occurs, a storyteller who has visual support has not recognized the potential increase in processing demands that a switch in reference forms causes for the listener who does not have access to that visual support (Wong & Johnston, 2004). This increased processing effort may or may not lead to an equivalent understanding of the concepts in the story as listeners attempt to find the most efficient way to resolve the ambiguity. Because this switching of reference forms would not create an equivalent difficulty for a listener who shared the visual context, we refer to this reference strategy as *picture-bound reference* (following Shapiro & Hudson, 1991).

A picture-bound reference strategy is even more apparent in Condition 3, in which new concepts are introduced as if they already existed in the discourse. When there is a shared visual context, picture-bound reference is a reasonable and pragmatically appropriate strategy for introducing concepts because visual information will support listeners as they quickly and easily disambiguate the reference. There is not an obligation to verbally introduce concepts into the discourse if they can be introduced visually.

In decontextualized narrative discourse, the visual information is not available to the listener. So, storytellers who use picture-bound reference strategies to introduce concepts are not recognizing the increased processing demands they are placing on their listeners and are risking that their listeners will not develop an equivalent understanding of the concepts in the story. By identifying this picture-bound reference strategy, the Semantic Elaboration Coding System was able to substantially match interdisciplinary team classification of children into the group with an FASD independent of the child's performance on a standardized expressive language task.

Pronouns and Picture-Bound Referencing

As a definite form, pronouns have the potential to be markers of picture-bound reference. Our results, however, do not show ambiguous pronoun reference to be a reasonable measure for accurately classifying children. It may be that pronouns represent a more complex reference form providing for more fine-grained manipulation of listener attention than full nominal phrases (Gundel, Hedberg, & Zacharski, 2001). As relatively complex forms, it is more likely that pronominal forms will be used in error by children in this age group (cf. Schelletter & Leinonen, 2003; van Der Lely, 1997; Wigglesworth, 1997; Wong & Johnston, 2004). This may be seen in the relatively poor specificity of PR in Table 4. Results indicate that despite the fact that the children in the FASD group tended to use relatively few unambiguous pronoun references, the same can be said for many of their TD peers, leading to a relatively high rate of false-positive classifications. The poor classification accuracy of APR may also reflect the fact that, unlike PR ($\kappa = .767$), APR was relatively difficult for judges to agree upon ($\kappa = .540$). Lower precision in a measure results in a higher chance of a Type II error and potentially masks the utility of the underlying construct. Given the uncertainties, this study highlights the need to consider nominal and pronominal lexical forms separately in narrative research but does not diminish the need for continued study of children's development and use of pronominal forms of reference in discourse.

Conclusion

Whether they exhibited average or low performance on a standardized language measure, children in this study who had an existing interdisciplinary team diagnosis of a disorder on the fetal alcohol spectrum (FASD) were more likely than were their typically developing peers to use a picture-bound reference strategy during storytelling. This strategy could be identified with reasonable accuracy using the rate of ANR calculated as a function of narrative length as defined in the Semantic Elaboration Coding System. If this result can be replicated in a well-designed validation study, the rate of ANR has the potential to provide useful diagnostic information to clinicians trying to identify clinical populations independent of their performance on standardized measures of expressive language.

As measured by the Semantic Elaboration Coding System, both pragmatically inappropriate reference strategies and semantic elaboration demonstrated the potential to play a role in the diagnosis of FASD by identifying languages behaviors that may be quite prevalent in this population (i.e., picture-bound referencing and unelaborated concepts in decontextualized narratives). It is unlikely that these impairments are specific to FASD (see Bates, 2004). Given that both semantic and pragmatic language ability are frequently compromised in children with complex clinical profiles of diverse etiology, these results also point to the possibility that narrative analysis using the Semantic Elaboration Coding System may have utility in other contexts. In particular, the rate of ANR in a narrative, which is easy to compute and has excellent reliability, may be a potential tool for reliably identifying pragmatic deficits in children who perform well on standardized language tasks despite poor performance during the socially integrated discourse of everyday communication.

Limitations of the Current Study

The TD participants in this study did not undergo the same interdisciplinary assessment as did the children with an FASD. TD participants were chosen because records indicated unremarkable behavior and adequate school achievement—a profile that would not, in a clinical setting, trigger such an assessment. This is not the same as undergoing the comprehensive assessment but provides a reasonable basis for contrasting the two groups in the context of a feasibility study. The lack of objective measures confirming that these children were indeed "typically developing" and not subject to a prenatal alcohol exposure increases the risk of Type II errors (potentially masking reasonable measures). The lack of objective language and cognitive measures on these children increases the chances of Type I errors in the unlikely event that as a group, their ability was significantly above average along the parameters of interest (potentially enhancing the apparent differences between the groups). Both of these limitations enhance the need for the results of this initial feasibility research to be confirmed with validation research.

Also, although strict criteria were used to screen out potentially useless test measures in this feasibility study, numerous measures were examined, so our results may be overfitted to the study population. Consequently, any measure and, particularly, any specific cutoff value reported here will need to be confirmed and validated in subsequent research. Results of this feasibility research can point to potentially useful diagnostic or screening measures, but until these measures have been shown to perform similarly in a well-designed validation study, their utility remains potential but unproven.

Acknowledgments

This research was supported in part by Health Resources and Services Administration: Maternal and Child Health Bureau Grant 6T73MC00041-11-01 and the Centers for Disease Control and Prevention Grant U84/CCU020163-03. Special thanks to Heather Collins and everyone at the University of Washington Child Language Laboratory and the Fetal Alcohol Diagnostic and Prevention Network.

References

Abkarian, G. G. (1992). Communication effects of prenatal alcohol exposure. *Journal of Communication Disorders, 25,* 221–240.

Acklin, M. W., McDowell, C. J., Verschell, M. S., & Chan, D. (2000). Interobserver agreement, intraobserver reliability, and the Rorschach Comprehensive System. *Journal of Personality Assessment, 74,* 15–74.

Adams, C. V., Lloyd, J., Aldred, C., & Baxendale, J. (2006). Exploring the effects of communication intervention for developmental pragmatic language impairments: A signal-generation study. *International Journal of Language and Communication Disorders, 41,* 41–65.

Astley, S. J. (2004). *Diagnostic guide for fetal alcohol spectrum disorders: The Four-Digit Diagnostic Code* (3rd ed.). Seattle: FAS Diagnostic and Prevention Network, University of Washington. (Electronic version available at http://fasdpn.org)

Astley, S. J., & Clarren, S. K. (1999). *Diagnostic guide for fetal alcohol syndrome and related conditions: The 4-Digit Diagnostic Code* (2nd ed.). Seattle, WA: University Publication Services.

Astley, S. J., & Clarren, S. K. (2000). Diagnosing the full spectrum of fetal alcohol exposed individuals: Introducing the 4-Digit Diagnostic Code. *Alcohol and Alcoholism, 35,* 400–410.

Bakeman, R., & Gottman, J. M. (1997). *Observing interaction* (2nd ed.). Cambridge, England: Cambridge University Press.

Bates, E. A. (2004). Explaining and interpreting deficits in language development across clinical groups: Where do we go from here? *Brain and Language, 88,* 248–253.

Becker, M., Warr-Leeper, G. A., & Leeper, H. A., Jr. (1990). Fetal alcohol syndrome: A description of oral motor, articulatory, short-term memory, grammatical, and semantic abilities. *Journal of Communication Disorders, 23,* 97–124.

Bishop, D. V., & Edmundson, A. (1987). Language-impaired 4-year-olds: Distinguishing transient from persistent impairment. *Journal of Speech and Hearing Disorders, 52,* 156–173.

Botting, N., & Adams, C. V. (2005). Semantic and inferencing abilities in children with communication disorders. *International Journal of Language and Communication Disorders, 40,* 49–66.

Carmichael-Olson, H., & Astley, S. J. (2005). [Intervening with children/adolescents with FAS/ARND]. Unpublished raw data.

Carmichael-Olson, H., Morse, B. A., & Huffine, C. (1998). Development and psychopathology: Fetal alcohol syndrome and related conditions. *Seminars in Clinical Neuropsychiatry, 3,* 262–284.

Carney, L. J., & Chermak, G. D. (1991). Performance of American Indian children with fetal alcohol syndrome on the test of language development. *Journal of Communication Disorders, 24,* 123–134.

Chudley, A. E., Conry, J., Cook, J. L., Loock, C., Rosales, T., & LeBlanc, N. (2005). Fetal alcohol spectrum disorder: Canadian guidelines for diagnosis. *Canadian Medical Association Journal, 175*(Suppl. 5), S1–S21.

Church, M. W., Eldis, F., Blakley, B. W., & Bawle, E. V. (1997). Hearing, language, speech, vestibular, and dentofacial disorders in fetal alcohol syndrome. *Alcoholism: Clinical and Experimental Research, 21,* 227–237.

Church, M. W., & Kaltenbach, J. A. (1997). Hearing, speech, language, and vestibular disorders in the fetal alcohol syndrome: A literature review. *Alcoholism: Clinical and Experimental Research, 21,* 495–512.

Cicchetti, D. V., & Sparrow, S. S. (1981). Developing criteria for establishing interrater reliabiity of specific items: Applications to assessment of adaptive behavior. *American Journal of Mental Deficiency, 86,* 127–137.

Coggins, T. E. (1995). [Narratives produced by typically developing school-aged children]. Unpublished raw data.

Coggins, T. E., Friet, T., & Morgan, T. (1998). Analysing narrative productions in older school-age children and adolescents with fetal alcohol syndrome: An experimental tool for clinical applications. *Clinical Linguistics & Phonetics, 12,* 221–236.

Coggins, T. E., Olswang, L. B., Carmichael-Olsen, H., & Timler, G. (2003). On becoming socially competent communicators: The challenge for children with fetal alcohol exposure. *International Review of Research in Mental Retardation, 27,* 121–150.

Condouris, K., Meyer, E., & Tager-Flusberg, H. (2003). The relationship between standardized measures of language and measures of spontaneous speech in children with autism. *American Journal of Speech-Language Pathology, 12,* 349–358.

Cornish, F. (1999). *Anaphora, discourse, and understanding: Evidence from English and French*. New York: Oxford University Press.

Croft, W. (2001). *Radical construction grammar: Syntactic theory in typological perspective*. Oxford, England: Oxford University Press.

Croft, W., & Cruse, D. A. (2004). *Cognitive linguistics*. Cambridge, England: Cambridge University Press.

Culatta, B., Page, J. L., & Ellis, J. (1983). Story retelling as a communicative performance screening tool. *Language, Speech, and Hearing Services in Schools, 14,* 67–74.

Curenton, S. M., & Justice, L. M. (2004). African American and Caucasian preschoolers' use of decontextualized language: Literate language features in oral narratives. *Language, Speech, and Hearing Services in Schools, 35,* 240–253.

Dwolatzky, T., Whitehead, V., Doniger, G., Simon, E., Schweiger, A., Jaffe, D., et al. (2003). Validity of a novel computerized cognitive battery for mild cognitive impairment. *BMC Geriatrics, 3*(1), 4.

Eisenberg, S., & Gillam, R. B. (2005, June). *Complex noun phrase elaboration in children's stories*. Poster session presented at the Symposium on Research in Child Language Disorders, Madison, WI.

Fauconnier, G. (2004). Mental spaces, language modalities, conceptual integration. In S. Davis & B. S. Gillon (Eds.), *Semantics: A reader*. Oxford, England: Oxford University Press.

Feagans, L., & Appelbaum, M. (1986). Validation of language subtypes in learning disabled children. *Journal of Educational Psychology, 78,* 358–364.

Feagans, L., & Short, E. J. (1984). Developmental differences in the comprehension and production of narratives by reading-disabled and normally achieving children. *Child Development, 55,* 1727–1736.

Gentry, B., Griffith, L., Dancer, J., Davis, P., Eaton, B., & Schulz, E. (1998). Prenatal alcohol exposure and communication, behavior, and nonverbal intelligence of 3 school-age children. *Perceptual and Motor Skills, 86*(3, Pt. 1), 1089–1090.

Gundel, J. K., Hedberg, N., & Zacharski, R. (2001). Definite descriptions and cognitive status in English: Why accommodation is unnecessary. *English Language and Linguistics, 5*(02), 273–295.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Hamilton, M. (1981). *Linguistic abilities of children with fetal alcohol syndrome*. Unpublished doctoral dissertation, University of Washington.

Hammer, C. S., Yont, K. M., & Tomblin, J. B. (2005). Language sampling for kindergarten children with and without SLI: Mean length of utterance, IPSYN, and NDW. *Journal of Communication Disorders, 38,* 197–213.

Heilmann, J., Weismer, S. E., Evans, J., & Hollar, C. (2005). Utility of the MacArthur–Bates Communicative Development Inventory in identifying language abilities of late-talking and typically developing toddlers. *American Journal of Speech-Language Pathology, 14,* 40–51.

Hesketh, A. (2004). Grammatical performance of children with language disorder on structured elicitation and narrative tasks. *Clinical Linguistics & Phonetics, 18,* 161–182.

Janzen, L. A., Nanson, J. L., & Block, G. W. (1995). Neuropsychological evaluation of preschoolers with fetal alcohol syndrome. *Neurotoxicology and Teratology, 17,* 273–279.

Juncos-Rabadan, O., Pereiro, A. X., & Rodriguez, M. S. (2005). Narrative speech in aging: Quantity, information content, and cohesion. *Brain and Language, 95,* 423–434.

Kaufman, A. S., & Kaufman, N. L. (1990). *The Kaufman Brief Intelligence Test*. Circle Pines, MN: AGS.

Klin, C. M., Weingartner, K. M., Guzman, A. E., & Levine, W. H. (2004). Readers' sensitivity to linguistic cues in narratives: How salience influences anaphor resolution. *Memory & Cognition, 32,* 511–522.

Kraemer, H. C. (1988). Assessment of 2X2 associations: Generalization of signal-detection methodology. *The American Statistician, 42*(1), 37–49.

Kraemer, H. C. (1992). *Evaluating medical tests: Objective and quantitative guidelines*. London: Sage Publications.

Kraemer, H. C. (n.d.). *Quality of measurement*. Retrieved December 20, 2004, from http://mirecc.stanford.edu/video_streaming.htm#Methods

Kraemer, H. C., Noda, A., & O'Hara, R. (2004). Categorical versus dimensional approaches to diagnosis: Methodological challenges. *Journal of Psychiatric Research, 38,* 17–25.

Kraemer, H. C., Periyakoil, V. S., & Noda, A. (2004). Tutorial in biostatistics: Kappa coefficients in medical research. In R. B. D'Agostino (Ed.), *Tutorials in biostatistics: Statistical methods in clinical studies* (Vol. 1, pp. 85–105). Hoboken, NJ: Wiley.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159–174.

Langacker, R. W. (1991). *Concept, image and symbol: The cognitive basis of grammar.* Berlin: Mouton de Gruyter.

Levine, W. H., & Klin, C. M. (2001). Tracking of spatial information in narratives. *Memory & Cognition, 29,* 327–335.

Lewis, D. (2004). Scorekeeping in a language game. In S. Davis & B. S. Gillon (Eds.), *Semantics: A reader* (pp. 803–816). Oxford, England: Oxford University Press.

Liles, B. Z., Duffy, R. J., Merritt, D. D., & Purcell, S. L. (1995). Measurement of narrative discourse ability in children with language disorders. *Journal of Speech and Hearing Research, 38,* 415–425.

Loban, W. (1976). *Language development: Kindergarten through grade twelve* (No. 18). Urbana, IL: National Council of Teachers of English.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Mahwah, NJ: Erlbaum.

Maratsos, M. P. (1976). *The use of definite and indefinite reference in young children: An experimental study of semantic acquisition*. New York: Cambridge University Press.

Mattson, S. N., & Riley, E. P. (1998). A review of the neurobehavioral deficits in children with fetal alcohol syndrome or prenatal exposure to alcohol. *Alcoholism: Clinical and Experimental Research, 22,* 279–294.

Mayer, M. (1969). *Frog where are you?* New York: Dial Press.

McFall, R. M., & Treat, T. A. (1999). Quantifying the information value of clinical assessments with signal detection theory. *Annual Review of Psychology, 50,* 215–241.

Miller, J. F. (2004). Systematic Analysis of Language Transcripts (Version 8.02, Research edition) [Computer software]. Madison, WI: University of Wisconsin—Madison.

Norbury, C. F., & Bishop, D. V. (2003). Narrative skills of children with communication impairments. *International Journal of Communication Disorders, 38,* 287–313.

Owens, R. (1999). *Language disorders: A functional approach to assessment and intervention* (3rd ed.). Boston: Allyn & Bacon.

Schelletter, C., & Leinonen, E. (2003). Normal and language-impaired children's use of reference: Syntactic versus pragmatic processing. *Clinical Linguistics & Phonetics, 17*(4–5), 335–343.

Shapiro, L. R., & Hudson, J. A. (1991). Tell me a make-believe story: Coherence and cohesion in young children's picture-elicited narratives. *Developmental Psychology, 27,* 960–974.

SPSS. (1998). SPSS for Windows (Version 9.0) [Computer software]. Chicago: Author.

Stephens, M. I. (1988). Pragmatics. In M. A. Nippold (Ed.), *Later language development: Ages nine through nineteen* (pp. 247–261). Boston: College-Hill.

Streissguth, A. P. (1997). *Fetal alcohol syndrome: A guide for families and communities*. Baltimore: Brookes.

Streissguth, A. P., Barr, H. M., Kogan, J., & Bookstein, F. L. (1996). *Understanding the occurrence of secondary disabilities in clients with fetal alcohol syndrome (FAS) and fetal alcohol effects (FAE): Final report*. Seattle: University of Washington.

Swets, J. A. (1988, June 3). Measuring the accuracy of diagnostic systems. *Science, 240,* 1285–1293.

Talmy, L. (2000a). *Toward a cognitive semantics: Typology and process in concept structuring* (Vol. II, pp. 417–482). Cambridge, MA: MIT Press.

Talmy, L. (2000b). *Toward a cognitive semantics: Volumes I and II*. Cambridge, MA: MIT Press.

Thorne, J. C. (2004). *The semantic elaboration coding system*. Unpublished training manual, University of Washington.

Thorne, J. C., & Coggins, T. E. (2004, November). *Semantic elaboration in narratives of children with FASD: The Semantic Elaboration Coding System*. Poster presented at the annual convention of the American Speech-Language-Hearing Association, Philadelphia, PA.

Thorne, J. C., & Coggins, T. E. (2005, June). *Exploring the utility of narrative analysis to aid in the diagnosis of fetal alcohol spectrum disorders (FASD): The Semantic Elaboration Coding System (SECS)*. Poster presented at the Symposium on Research in Child Language Disorders, Madison, WI.

Timler, G. R., Olswang, L. B., & Coggins, T. E. (2005). "Do I know what I need to do?" A social communication intervention for children with complex clinical profiles. *Language, Speech, and Hearing Services in Schools, 36,* 73–85.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

van Der Lely, H. K. J. (1997). Narrative discourse in grammatical specific language impaired children: A modular language deficit? *Journal of Child Language, 24,* 221–256.

van Hoek, K. (1997). *Anaphora and conceptual structure*. Chicago: The University of Chicago Press.

Wagner, C. R., Nettelbladt, U., Sahlen, B., & Nilholm, C. (2000). Conversation versus narration in pre-school children with language impairment. *International Journal of Language and Communication Disorders, 35,* 83–93.

Weinberg, N. Z. (1997). Cognitive and behavioral deficits associated with parental alcohol use. *Journal of the American Academy of Child and Adolescent Psychiatry, 36,* 1177–1186.

Wigglesworth, G. (1997). Children's individual approaches to the organization of narrative. *Journal of Child Language, 24,* 279–309.

Wiig, E. H., & Secord, W. A. (1989). *Test of Language Competence—Expanded edition*. San Antonio, TX: Psychological Corporation.

Wong, A. M.-Y. (2001). *Who are you talking about? The development of reference in Cantonese-speaking children with and without specific language impairment*. Unpublished doctoral dissertation, University of British Columbia.

Wong, A. M.-Y., Au, C. W. S., & Stokes, S. F. (2004). Three measures of language production for Cantonese-speaking school-age children in a story-retelling task. *Journal of Speech, Language, and Hearing Research, 47,* 1164–1178.

Wong, A. M.-Y., & Johnston, J. R. (2004). The development of discourse referencing in Cantonese-speaking children. *Journal of Child Language, 31,* 633–660.

Received July 27, 2005

Revision received December 23, 2005

Accepted July 13, 2006

DOI: 10.1044/1092-4388(2007/032)

Contact author: John C. Thorne, Department of Speech and Hearing Sciences, University of Washington, 1417 N.E. 42nd Street, Seattle, WA 98105-6246. E-mail: jct6@u.washington.edu.