JSLHR

Research Article

Differentiating School-Aged Children With and Without Language Impairment Using Tense and Grammaticality Measures From a Narrative Task

Ling-Yu Guo[a] and Phyllis Schneider[b]

Purpose: To determine the diagnostic accuracy of the finite verb morphology composite (FVMC), number of errors per C-unit (Errors/CU), and percent grammatical C-units (PGCUs) in differentiating school-aged children with language impairment (LI) and those with typical language development (TL).

Method: Participants were 61 six-year-olds (50 TL, 11 LI) and 67 eight-year-olds (50 TL, 17 LI). Narrative samples were collected using a story-generation format. FVMC, Errors/CU, and PGCUs were computed from the samples.

Results: All of the three measures showed acceptable to good diagnostic accuracy at age 6, but only PGCUs showed acceptable diagnostic accuracy at age 8 when sensitivity, specificity, and likelihood ratios were considered.

Conclusion: FVMC, Errors/CU, and PGCUs can all be used in combination with other tools to identify school-aged children with LI. However, FVMC and Errors/CU may be an appropriate diagnostic tool up to age 6. PGCUs, in contrast, may be a sensitive tool for identifying children with LI at least up to age 8 years.

A major clinical issue in the field of child language disorders is how to correctly identify children with language impairment (LI). Although norm-referenced, standardized language tests are widely used for identifying children with LI, the tests typically evaluate what children know but not what they do with this knowledge in real-life situations (Costanza-Smith, 2010). Thus, a child might pass standardized language tests but still show difficulty using language in daily activities (Ebert & Scott, 2014; Nippold et al., 2014). In addition, in a review of 43 standardized tests, Spaulding, Plante, and Farinella (2006) found that only five tests had acceptable (i.e., 80%–89% accurate) or good (90% accurate or higher) identification accuracy as was reported in the test manual. Moreover, regardless of the tests, if an arbitrary cutoff score (e.g., 1.25 *SD* below the mean) is applied for diagnosis, a child who is truly affected by LI has an approximately equal chance of being correctly or incorrectly identified (Spaulding et al., 2006). Given the limitations of standardized tests, it is important for clinicians to consider assessment tools beyond standardized tests in order to reduce the possibility of underidentifying children with LI. One way to improve the identification of children with LI is to supplement norm-referenced, standardized tests with language sample analysis (Ebert & Scott, 2014; Paul & Norbury, 2012). Previous studies have found that measures from language sample analysis may even be more accurate in identifying children with LI than standardized tests in some cases (Aram, Morris, & Hall, 1993; Dunn, Flax, Sliwinski, & Aram, 1996; Eisenberg & Guo, 2013).

However, not all language sample measures can be used for differentiating children with and without LI. To be able to supplement standardized tests for the improvement of identification accuracy, a language sample measure must demonstrate at least acceptable diagnostic accuracy because a measure with unacceptably low diagnostic accuracy does not provide useful supplementary information for diagnosis (Plante & Vance, 1994; Spaulding et al., 2006). Among language sample measures, those that specifically evaluate children's performance on clinical markers (e.g., tense deficits) could be particularly effective in identifying children with LI (Rice & Wexler, 1996; Tager-Flusberg & Cooper, 1999). One such measure is the *finite verb morphology composite*

[a] University at Buffalo, NY
[b] University of Alberta, Edmonton

Correspondence to Ling-Yu Guo: lingyugu@buffalo.edu

Editor: Rhea Paul
Associate Editor: Shelley Gray

Received February 12, 2015
Revision received April 14, 2015
Accepted September 10, 2015
DOI: 10.1044/2015_JSLHR-L-15-0066

Disclosure: The authors have declared that no competing interests existed at the time of publication.

(FVMC; Bedore & Leonard, 1998). FVMC is a measure of English-speaking children's overall accuracy of using tense and agreement morphemes (hereafter, "tense morphemes") in spoken discourse. FVMC focuses on tense morphemes because it has been documented that English-speaking children with LI have particular difficulty learning tense morphemes (Leonard, 2014). Preschool children with LI are more likely to omit tense morphemes than children without LI in spoken discourse (Rice & Wexler, 1996). The difficulty in producing and comprehending tense morphemes may persist into school ages (Windsor, Scott, & Street, 2000) or even adulthood (Poll, Betz, & Miller, 2010). Together, the existing studies suggest that the tense deficit is a potential clinical marker of LI and that it can be used to identify children with LI (Tager-Flusberg & Cooper, 1999). Across studies, FVMC yields acceptable to good diagnostic accuracy in differentiating preschool children with and without LI (Bedore & Leonard, 1998; Guo & Eisenberg, 2014). However, the extent to which FVMC can accurately identify school-aged children with LI remains unclear because of the discrepant findings on the diagnostic accuracy of FVMC between studies (Moyle, Karasinski, Ellis Weismer, & Gorman, 2011; Souto, Leonard, & Deevy, 2014).

Another measure that could potentially improve the identification of school-aged children with LI is the production of grammatical utterances (Dunn et al., 1996; Eisenberg & Guo, 2013). Although tense marking errors are a potential clinical maker of LI, children with LI also show other grammatical difficulties, such as pronoun errors (Moore, 2001), argument structure errors (Ebbels, Van Der Lely, & Dockrell, 2007), and errors with grammatical morphemes other than pronouns and tense markers (e.g., plural *–s*, articles *the, a*; Leonard, 2014; Thordardottir, 2008). Previous studies have shown that school-aged children with LI tend to make more lexical and grammatical errors (Heilmann, Miller, & Nockerts, 2010; Reilly, Losh, Bellugi, & Wulfeck, 2004; Scott & Windsor, 2000) and produce grammatically correct utterances at a lower rate (Fey, Catts, Proctor-Williams, Tomblin, & Zhang, 2004) than peers with typical language development (TL) in spoken discourse. However, the extent to which the grammaticality measures can accurately identify school-aged children with LI has yet to be determined.

The purpose of the present study was to evaluate the diagnostic accuracy of FVMC and two grammaticality measures (i.e., number of grammatical errors per utterance and percent grammatical utterances) in identifying school-aged children with and without LI who were 6 or 8 years old using a narrative task. Including two age groups allowed us to determine whether diagnostic accuracy of the tense and grammaticality measures would change as a function of age. This is because children's knowledge of tense morphemes and other grammatical elements develops over time, and thus the diagnosis accuracy of the tense and grammaticality measures may change over time as well. In what follows, we first review the studies on identifying school-aged children with LI using FVMC or grammaticality measures and then lay out the scope of the present study.

Identifying School-Aged Children With LI Using FVMC

Moyle et al. (2011) examined the extent to which FVMC can differentiate school-aged children with and without LI between the ages of 5;5 (years;months) and 9;8 using language samples. The inclusionary criteria for the condition of LI included intervention status and performance on standardized language tests. To be identified as having LI, children must have been receiving speech-language services at the time of participation. They also had to score lower than 1 *SD* below the mean (i.e., −1 *SD*) on the Peabody Picture Vocabulary Test–Revised (Dunn & Dunn, 1981) and/or the Test for Auditory Comprehension of Language (Carrow-Woolfolk, 1985). The language samples that were used to compute FVMC involved the examiner interviewing children by asking them to describe a fictional (e.g., a favorite movie) or personal (e.g., a recent birthday party) experience and to explain how to do particular things (e.g., cook a favorite meal, play chess). FVMC (Bedore & Leonard, 1998), which examined the overall percentage of correct use in obligatory contexts of third person singular present *–s*, past tense *–ed*, copula *be,* and auxiliary *be* in the language samples, was computed for each child. FVMC was lower in children with LI (*M* = 94%, *SD* = 6%) than in children with typical language (*M* = 98%, *SD* = 4%). However, FVMC had a sensitivity level of 50% and a specificity level of 86% (the cutoff score was not reported). That is, FVMC underidentified 50% of school-aged children with LI in the sampled age range. On the basis of the findings, Moyle et al. (2011) argued that the low sensitivity of FVMC occurred because "by the time children with LI are examined at the school-age level, they have acquired the verb morphemes" (p. 556). They further concluded that FVMC alone was not a useful clinical tool in identifying school-aged children with LI.

Gladfelter and Leonard (2013) examined the diagnostic accuracy of FVMC for school-aged children between the ages of 5;0 and 5;6 using conversational language samples that involved the child and the examiner playing with age-appropriate toys. The inclusionary criterion for the condition of LI was that children must score below −2.20 *SD* on the Structured Photographic Expressive Language Test–Second Edition (Werner & Kresheck, 1983). FVMC was lower in children with LI (*M* = 68%, *SD* = 19%) than in children with TL (*M* = 97%, *SD* = 5%). In addition, a cutoff FVMC of 85% yielded good sensitivity (92%) and specificity (93%) levels for differentiating 5-year-olds with and without LI. Using procedures similar to those of Gladfelter and Leonard (2013), Souto et al. (2014) further examined the diagnostic accuracy of FVMC for school-aged children between the ages of 5;0 and 5;10. FVMC was lower in children with LI (*M* = 70%, *SD* = 19%) than in children with TL (*M* = 98%, *SD* = 4%). In addition, FVMC showed good sensitivity (91%) and specificity (93%), but the cutoff score was not reported.

In summary, whereas Moyle et al. (2011) indicated that FVMC had unacceptably low sensitivity and tended to underidentify school-aged children with LI, Gladfelter and Leonard (2013) and Souto et al. (2014) both found that FVMC demonstrated good diagnostic accuracy, at least for school-aged children who were 5 years old. The discrepancy may have resulted from the age range of children in these studies. Whereas Moyle et al. (2011) included children between the ages of 5;5 and 9;8, Gladfelter and Leonard (2013) and Souto et al. (2014) limited the participants to 5-year-olds only. It is possible that tense measures may yield at least acceptable diagnostic accuracy in differentiating younger school-aged children with and without LI. Indeed, Moyle et al. (2011) attempted to address the issue of children's age by limiting the analysis to the 20 youngest children with LI in the sample. The age of these children ranged from 5;5 to 7;4, and the sensitivity remained low (45%). Thus, the low sensitivity of FVMC in Moyle et al. (2011) may have resulted from the wide age range of children rather than the adequacy of FVMC. To address the age range issue in Moyle et al. (2011) and to extend the study of Souto et al. (2014), the present study examined the diagnostic accuracy of FVMC in school-aged children with and without LI who were 6 or 8 years old using a narrative generation task. To be specific, the diagnostic accuracy of FVMC was evaluated separately for each age group. Evaluating the diagnostic accuracy of measures by age has been recommended by other researchers (Nelson, Plante, & Anderson, 2014). Oetting and Hadley (2009) indicated that it is relatively difficult to differentiate children with and without LI using tense measures after age 8 on the basis of the studies by Conti-Ramsden, Botting, and Faragher (2001) and Rice and Wexler (2001). It is possible that the diagnostic accuracy of FVMC may be higher in younger (e.g., 6-year-old) than in older (e.g., 8-year-old) school-aged children. We included both age groups in order to test this possibility, which, in turn, may inform clinicians about the appropriate age range for using FVMC in the diagnostic process.

Identifying School-Aged Children With LI Using Grammaticality Measures

Two methods have been used to quantify children's production of grammatical utterances: (a) number of grammatical errors per utterance, and (b) percent grammatical utterances. Scott and Windsor (2000) examined the number of grammatical errors per terminable unit (T-unit; Hunt, 1965) in the spoken narrative of school-aged children (9;10–12;11) with LI and those with TL. To be identified as having LI, children had to have a diagnosis of language learning disability at the time of participation. They also had to score below −1 *SD* on the expressive language subtest of the Test of Language Development–Intermediate (Hammill & Newcomer, 1988). Children were asked to watch one video about a story and then retell the story. The spoken narratives were segmented into T-units. Any errors that made a T-unit ungrammatical were tallied. The number of grammatical errors per T-unit was significantly greater in children with LI ($M = 0.13$, $SD = 0.09$) than in their age-matched ($M = 0.03$, $SD = 0.03$) or their language-matched ($M = 0.04$, $SD = 0.04$) peers with TL. Similar findings were also observed in other studies (e.g., Colozzo, Gillam, Wood, Schnell, & Johnston, 2011; Ebert & Scott, 2014; Heilmann et al., 2010; Reilly et al., 2004).

Fey et al. (2004) examined percent grammatical C-units (PGCUs) in the spoken and written narratives of second- and fourth-grade children with and without LI. To be identified as having LI, children had to score below −1.14 *SD* on the language composite derived from three standardized language tests. Children were asked to tell a story on the basis of a three-picture sequence and write a story on the basis of another sequence. The spoken and written narratives were segmented into communication units (C-units; Loban, 1976). PGCUs were computed by dividing the total number of C-units that did not contain any errors by the total number of C-units. Regardless of modality (i.e., spoken or written), the means of PGCUs were significantly lower in the LI group (second grade: $M = 78\%$, $SD = 13\%$; fourth grade: $M = 75\%$, $SD = 14\%$) than in the TL group (second grade: $M = 86\%$, $SD = 11\%$; fourth grade: $M = 84\%$, $SD = 13\%$) at the second grade and at the fourth grade. Altogether, previous studies (Ebert & Scott, 2014; Fey et al., 2004; Scott & Windsor, 2000) suggested that measures of the production of grammatical utterances can potentially be used to differentiate school-aged children with and without LI given the group differences in these studies.

Souto et al. (2014) examined the diagnostic accuracy of percent grammatical utterances in identifying 5-year-old children with and without LI. Following the procedure of Developmental Sentence Scoring (DSS; Lee, 1974), Souto et al. (2014) included 50 utterances that had at least a subject and a verb in the language sample for analysis. On the basis of the scoring rules of DSS, each utterance was counted as grammatical only if it did not contain any grammatical errors or semantic irregularities (e.g., *He's writing a picture*). Percent grammatical utterances was computed by dividing the total number of grammatical utterances by 50. Percent grammatical utterances was significantly lower in children with LI ($M = 70\%$, $SD = 9\%$) than in children with TL ($M = 93\%$, $SD = 3\%$). More importantly, percent grammatical utterances showed good sensitivity (100%) and specificity (100%) levels; the cutoff score was not reported. As a replication, Souto et al. (2014) further examined the diagnostic accuracy of percent grammatical utterances in a different sample of 5-year-olds. Percent grammatical utterances yielded good sensitivity (100%) and specificity (100%) levels in the replication sample as well; again, the cutoff score was not reported. In a similar vein, Eisenberg and Guo (2013) found that, with the cutoff score of 58%, percent grammatical utterances yielded good sensitivity (100%) and acceptable specificity (88%) levels for 3-year-olds in picture description samples.

Despite the fact that percent grammatical utterances showed good sensitivity and specificity for differentiating 5-year-olds with and without LI in Souto et al. (2014), the clinical application of the results is limited because no cutoff scores were reported. Without the cutoff scores of percent grammatical utterances, clinicians will not be able to use the results of Souto et al. (2014) to determine with confidence whether a 5-year-old has LI. In addition, to the best of our knowledge, no studies have examined the diagnostic accuracy of percent grammatical utterances beyond age 5. Furthermore, because number of grammatical errors per utterance specifically counts each error that children make, it is presumably a more fine-grained analysis, and may yield higher diagnostic accuracy, than percent grammatical utterances. However, no studies have evaluated how well number of grammatical errors per utterance can identify school-aged children with and without LI.

The Present Study

The purpose of the present study was to evaluate the diagnostic accuracy of FVMC, number of grammatical errors per utterance, and percent grammatical utterances for differentiating school-aged children with and without LI who were 6 or 8 years old using a narrative generation task. The diagnostic accuracy and the associated cutoff scores were reported for each measure by age. We included both age groups in order to determine the extent to which the diagnostic accuracy of the target measures would change as a function of children's ages, which may, in turn, provide a guide for clinicians in using those measures appropriately. The cutoff scores would further help clinicians in making clinical decisions for children at different ages. In addition, we chose a narrative generation task, instead of conversation, as the elicitation context because narratives are cognitively more demanding than conversations (Colozzo et al., 2011; Thordardottir, 2008). It is more likely to observe breakdowns in the use of tense morphemes and other grammatical elements in narratives than in conversations (Thordardottir, 2008).

In the present study, we asked two specific questions. First, would FVMC, number of errors per utterance, and percent grammatical utterances show at least acceptable diagnostic accuracy in 6- and 8-year-old children? On the basis of previous studies (Fey et al., 2004; Rice & Wexler, 1996; Scott & Windsor, 2000; Souto et al., 2014), we predicted that all three measures would yield at least acceptable sensitivity and specificity in differentiating children with and without LI who were 6 or 8 years old. Second, would the diagnostic accuracy of FVMC, number of errors per utterance, and percent grammatical utterances change from 6 to 8 years of age? On the basis of previous studies (Moyle et al., 2011; Oetting & Hadley, 2009; Rice & Wexler, 1996), we predicted that the diagnostic accuracy of FVMC would be higher for 6-year-olds than for 8-year-olds. The diagnostic accuracy of the two grammaticality measures would be similar across the age groups.

Method

Participants

Participants were 61 six-year-olds (30 girls, 31 boys) and 67 eight-year-olds (32 girls, 35 boys) recruited from the area of Edmonton, Canada. Within the 6-year-old group, 50 children had TL and 11 had LI; within the 8-year-old group, 50 children had TL and 17 had LI. The chronological ages (see Table 1) were not significantly different between the TL and the LI groups at either age, Fs < 2.41, p > .13, $\eta_p^2 < .04$. The distribution of gender was also not significantly different between the TL and the LI groups at either age, $\chi^2 < 0.40$, ps > .53; or between the 6- and 8-year-old groups regardless of language status, $\chi^2 = 0.03$, $p = .87$. All children were from English-speaking families and spoke English at home from birth; in some cases, another language may have been spoken in the home, as we specified only that English must be a first language in our inclusion criteria. These children were part of a local normative sample for the Edmonton Narrative Norms Instrument (ENNI) that included children from 4 to 9 years of age (Schneider, Dubé, & Hayward, 2005).

To identify potential participants with TL for the normative sample, we chose children from 34 public elementary schools in Edmonton, Alberta. The teachers in those schools who had students in the age range in the present study were asked to refer two children in the upper level of achievement, two children from the middle level, and two children in the lower level (one boy and one girl at each level). This decision was made to ensure that the normative sample consisted of children with TL who had varying language abilities. In all cases, the children did not have speech or language difficulties or any other diagnostic label such as attention deficit disorder (ADD), learning disability, or autism on the basis of the teachers' reports. Signed consents were obtained from the parents of all of the children who participated.

Children with LI in the normative sample were recruited from three sites: a public school serving children with communication disorders; a rehabilitation hospital that has several programs for children with LIs; and Capital Health Authority, which serves preschool and school-aged children throughout the city of Edmonton. Those sites were asked to refer children with a rating of 2 to 5 on a severity rating scale designed and used by Capital Health that rates each child's LI from 1 (*mild*) to 5 (*severe*). Children could be referred even if they had fine or gross motor delays, ADD with or without hyperactivity (ADD/ADHD) with medication, a diagnosed learning disability, or mild or moderate speech disorders. Sites were asked not to refer children who had received diagnoses of mental retardation, ADD/ADHD without medication, autism, hearing impairment, severe visual impairment that would result in inability to see pictures even with correction, or severe speech impairments that would preclude accurate orthographic transcription of their stories. Information regarding nonverbal/performance IQ was not collected; the speech-language pathologists referring children for the study were asked to refer children for whom they had no concerns regarding cognitive abilities. The examiners who tested the participants for the present study did not have concerns about cognitive abilities of the participants either.

Table 1. Mean (*SD*) of demographic measures of children by age and language status.

Group	Gender	Age in months	SES	CELF-3
6-year-olds				
TL (*n* = 50)	25 F, 25 M	78.94 (3.99)	48.31 (14.75)	108.40 (14.41)
LI (*n* = 11)	5 F, 6 M	79.55 (3.17)	40.26 (13.97)	78.55 (8.18)
8-year-olds				
TL (*n* = 50)	25 F, 25 M	102.92 (3.34)	45.04 (11.55)	107.20 (12.77)
LI (*n* = 17)	7 F, 10 M	104.35 (3.12)	42.42 (7.40)	76.29 (11.94)

Note. SES = socioeconomic status; CELF-3 = Clinical Evaluation of Language Fundamentals–Third Edition; TL = children with typical language; F = female; M = male; LI = children with language impairment. For children with TL, the composite standard scores of the subtests of Concepts and Directions and Recalling Sentences were reported. To derive the composite standard scores, we first obtained the scaled scores (*M* = 10, *SD* = 3) for both subtests using the norm in the manual of CELF-3 (Semel et al., 1995). The scaled scores were converted into *Z* scores (*M* = 0, *SD* = 1), which were then averaged and converted into standard scores (*M* = 100, *SD* = 15). For children with LI, the standard scores of Total Language Composites were reported using the manual of CELF-3.

As a further confirmation of language status, the children in the present study were administered the Clinical Test of Language Fundamentals–Third Edition (CELF–3; Semel, Wiig, & Secord, 1995). For children in the present study who were referred as having TL, the subtests of Concepts and Directions and Recalling Sentences were used to screen their language abilities. To pass the screening, children had to score at −1 *SD* or better on both of the subtests (Semel et al., 1995). All children who were referred as having TL met this criterion. Table 1 presents the composite standard scores of the subtests of Concepts and Directions and Recalling Sentences for the children with TL in the present study. For children in the present study who were referred as having LI, the full CELF-3 was administered. All of the children who were classified as having an LI in the present study scored below −1 *SD* (i.e., a standard score of 85) on at least one of the composite scores (i.e., Receptive, Expressive, and Total Language Composites) of the CELF-3. The cutoff standard score of 85 was based on the recommendation of the CELF-3 manual and was consistent with those in previous studies on children with LI (Coady, Evans, & Kluender, 2010; Munson, Kurtz, & Windsor, 2005). The percentage of 6-year-olds with LI who scored below the cutoff was 73% (eight of 11) for the Receptive Composite, 82% (nine of eleven) for the Expressive Composite, and 82% (nine of eleven) for the Total Language Composite. The percentage of 8-year-olds with LI who scored below the cutoff was 41% (seven of 17) for the Receptive Composite, 100% (17/17) for the Expressive Composite, and 76% (13/17) for the Total Language Composite. Thus, all children with LI in the present study were receiving language intervention at the time of data collection and scored below −1 *SD* on the CELF-3. Those inclusionary criteria were consistent with previous studies (Ebert & Scott, 2014; Heilmann et al., 2010; Scott & Windsor, 2000). The Total Language Composites of children with LI in the present study are presented in Table 1.

As part of the assessment, the socioeconomic status (SES) of the children was estimated from parents' occupations using the Blishen Scales (Blishen, Carroll, & Moore, 1987). On the basis of Canadian census information, this index reflects equally weighted components of education and income level by occupation. For instance, newspaper carriers and vendors are assigned a score of 17.81, and dentists are assigned a score of 101.74 on the scale. Table 1 also presents the mean SES score by age and language status. The SES scores were not significantly different between the TL and LI groups at either age (*p*s > .12) or between the 6- and 8-year-old groups regardless of language status (*p* = .29). The information regarding whether children with LI also had ADD/ADHD (with medication), motor delay, or speech disorders was not available at the time of data collection because access to children's clinical records was not obtained.

Ethnic composition of children in the present study corresponded closely to the range of ethnic diversity in the city of Edmonton according to Statistics Canada data (Statistics Canada, n.d.): Approximately 72% of the participants were of European origin, and 28% were of non-European origin.

Materials

Six original picture sequences with animal characters were created for the ENNI (Schneider et al., 2005) to elicit narratives from children. All of the picture sequences were black and white line drawings drawn by a professional cartoonist based on the scripts created by the ENNI authors. The picture sequences depicted stories that varied in three levels of complexity (two picture sequences for each level). To reflect the complexity of the stories, the picture sequences systematically varied in length (i.e., five, eight, and 13 pictures), amount of story information, and number and gender of characters (i.e., two, three, and four characters). The six picture sequences were equally divided into two sets (i.e., Set A and Set B) such that each set had one picture sequence from each complexity level (i.e., three picture sequences per set). These picture sets may be viewed and downloaded from the ENNI website.

Procedures

Three full-time female research assistants with a bachelor's degree in education or psychology were employed to collect the data. Each child was seen individually by one examiner in the child's school. Before the task, the child was instructed that he could see all the pages first and then tell the story to the examiner. The instructions emphasized that the examiner would not be able to see the pictures so the child would have to tell a really good story in order for the examiner to understand it.

The pictures for each story were placed in page protectors in a binder. Each story was in its own binder. The examiner was required to hold the binder in such a way that she could not see the pictures as the child told the story, which meant that the child needed to be explicit so the examiner would be able to understand the story; the child could not legitimately use pointing in lieu of language when telling the story. When administering each story, the examiner first went through all the pages so that the child could preview the story, after which the examiner turned the pages again as the child told the story. The examiner turned the page after the child appeared to be finished telling the story for a particular picture.

The child was first given a training story consisting of a single episode story in five pictures in order to familiarize the child with the procedure and to allow the examiner to give explicit prompts if the child had difficulty with the task. The training story was not included for analysis. After the training story was administered, the two story sets were given. Administration of the story sets was counterbalanced, with half of the children telling stories from Set A first and the other half telling stories from Set B first. The examiner was restricted to less explicit assistance for story Sets A and B (e.g., general encouragement, repetition of the child's previous utterances) than for the training story. Stories were audio-recorded for transcription and analysis.

Data Transcription, Coding, and Computation

The narrative samples were transcribed and coded by trained research assistants on the basis of the conventions of Systematic Analysis of Language Transcripts (Miller & Chapman, 2000). Children's narratives were segmented into C-units. A *C-unit* is typically an independent clause plus all of its dependent clauses (Loban, 1976). Nonclausal utterances that expressed complete thoughts (e.g., *A giraffe and an elephant*) were also counted as C-units. Only intelligible, complete, and spontaneous C-units that described the stories were included for computing the descriptive measures (e.g., mean length of C-units). To be included for the analysis for FVMC or grammaticality measures, a C-unit also had to have at least a verb (e.g., *He's mad at her*, *The elephant dropped the ball*, *Fell into the pool*), except in the case of C-units with omitted copula *be* (e.g., *She sad*; Eisenberg & Guo, 2013). It should be noted that whereas other studies (e.g., Lee, 1974) did not include C-units without subjects or C-units with omitted copula *be* for grammaticality analysis, the present study did. This decision was made on the basis of a prior study with younger children (Eisenberg & Guo, 2013), which found that including those C-units in the analysis yielded higher diagnostic accuracy in identifying 3-year-olds with and without LI than excluding those C-units.

FVMC

FVMC computes the percentage correct of third person singular present *–s* (3SG *–s*), regular past tense *–ed*, and copula and auxiliary *be* (i.e., *am, are, is, was, were*) in obligatory contexts. An *obligatory context* is operationally defined as an instance in which a particular tense marker is required for the C-unit to be grammatical. For instance, the C-unit *They playing outside* has an obligatory context of auxiliary *be*, but the child omits it. C-units without subjects (e.g., *Pull the doctor on the street*) are not coded for FVMC because they do not provide obligatory contexts for tense usage. We also excluded overgeneralization of 3SG *–s* (e.g., *The rabbit haves balloons*) and regular past tense *–ed* (e.g., *The dog breaked the castle*) from computation because verbs in these contexts, by definition, do not require the use of 3SG *–s* or past tense *–ed*. For each obligatory context, the target tense morpheme was coded as (a) correctly used, (b) omitted, or (c) incorrectly used (e.g., *The dog and the rabbit is playing*). FVMC was computed by dividing the total number of correct uses of target tense morphemes in obligatory contexts by the total number of obligatory contexts for these morphemes in the narrative (see the Supplemental Material 1 for examples).

Grammaticality Measures

We evaluated children's production of grammatical C-units with two measures: number of errors per C-unit (Errors/CU; Scott & Windsor, 2000; Colozzo et al., 2011) and PGCUs (Fey et al., 2004). To this end, we first identified the errors that children made in the narratives on the basis of the coding scheme of Eisenberg and Guo (2013).

1. *Tense marking errors* were operationally defined as omissions or incorrect usage of tense markers, including 3SG *–s*, regular past tense *–ed*, copula *be*, auxiliary *be*, auxiliary *do*, irregular past tense, and irregular third person verb forms (e.g., '*has*' as in *He has a dog*). Verbs produced without an inflection, modal, or auxiliary (e.g., *The elephant jump into the pool*) were transcribed as *bare verbs*, and inappropriate uses of bare verbs were counted as tense marking errors. However, unmarked verbs with first-person, second-person, or plural subjects were not coded as errors (e.g., *They play ball by the pool*) unless the context clearly required a tense marker (e.g., *They go to the beach yesterday*). It should be noted that the types of tense markers and errors that were covered in the grammaticality measures were broader than those in FVMC. For instance, whereas overgeneralization of past tense *–ed* was counted as a tense marking error in the grammaticality measures, it was not included in the computation of FVMC.
2. *Pronoun errors* were operationally defined as substitution errors for subject pronouns (e.g., *him walks over*), object pronouns, reflexive pronouns,

possessive pronouns, possessive determiners, and omissions or incorrect uses of relative pronouns (e.g., *This is the guy *who helped me get the ball*; "*" indicates omission). Gender errors of pronouns were determined based on inconsistency between the child's C-units within the same story and were not based on whether the gender of pronouns matched the picture characters. For instance, if a child refers to a pictured character as *the girl* and uses the pronoun *he* later to refer to the same character, it would be counted as a gender error.

3. *Grammatical morpheme errors* were operationally defined as omission or incorrect uses of grammatical morphemes other than pronouns and tense markers, such as determiners (e.g., *a, the, that*), plural *–s*, prepositions, and present and past participles. Errors of present and past participles were coded only when there were obligatory contexts (e.g., *He is run now*).
4. *Argument structure errors* were operationally defined as omissions of required constituents (i.e., argument) before or after verbs (e.g., *Made a sandcastle; The elephant hugged*). Decisions about the required arguments for verbs were made on the basis of the *Longman Dictionary of Contemporary English* (2014). Any omissions that could be considered as pragmatically allowable elision were not counted as argument structure errors.
5. *Other errors* were operationally defined as any other syntactic errors (e.g., *The rabbit did not know what was the dog doing*) or semantic irregularities (e.g., *The balloon is gonna pop and tick*) that could not be classified into any error categories. We counted semantic irregularities as errors for two reasons. First, syntax is not independent of meaning. Rather, semantics contributes to the well-formedness of sentences (Halliday, 1994; Saeed, 2009). Second, this decision is consistent with other assessments, such as DSS (Lee, 1974) and the Sentence Formulation subtest of the Clinical Evaluation of Language Fundamentals–Fourth Edition (Semel, Wiig, & Secord, 2003), both of which score semantic irregularities as errors.

After the errors were identified, we tallied the total number of errors and the total number of ungrammatical C-units for each child. A C-unit was counted as ungrammatical if it had one or more of the errors mentioned above. Number of errors per C-unit (Errors/CU) was computed by dividing total number of errors by total number of C-units that were included for analysis; percent grammatical C-units (PGCUs) were computed by subtracting the total number of ungrammatical C-units from the total number of C-units and then dividing by the total number of C-units. Supplemental Material 1 presents the examples for computing Errors/CU and PGCUs, and Supplemental Material 2 presents the frequency and percentage of each error type by age and language status.

Reliability of Transcription and Coding

To check the transcription reliability, the narratives were first transcribed by research assistants majoring in speech-language pathology. The transcripts were then checked against the recordings by the second author. Another research assistant majoring in speech-language pathology independently transcribed the narratives of two children with LI and six children with TL for each age group (i.e., approximately 12.5% for each group by age; $n = 16$). The word-by-word consistency was 96%.

To check the coding reliability for FVMC, Errors/CU, and PGCUs, we adapted a consensus procedure from Shriberg, Kwiatkowski, and Hoffman (1984). Two graduate assistants first coded the morphemes for FVMC and the errors for grammatical measures for all children. The first author then checked the coded transcripts for all children. Discrepancies were discussed between the first and the second authors. Overall, 79 (2%) of 4,392 C-units in the 6-year-old group and 61 (1%) of 5,192 C-units in the 8-year-old group were discussed. All of the discrepancies were resolved.

Statistical Analysis

We used one-way analysis of variance to examine group differences in the descriptive measures (e.g., number of obligatory contexts for FVMC) and the target measures (i.e., FVMC, Errors/CU, and PGCUs). The significance level was set at .05. In addition, because FVMC and the PGCUs were in percentage, they were arcsine-transformed in the analyses of variance.

To evaluate the diagnostic accuracy of FVMC, Errors/CU, and PGCUs, we computed the sensitivity, specificity, and likelihood ratios for these measures (Dollaghan, 2007). *Sensitivity* refers to the extent to which a measure can accurately identify children with LI. It was computed as the percentage of children with LI who were also identified by FVMC, Errors/CU, or PGCUs. In contrast, *specificity* refers to the extent to which a measure can accurately identify children with TL. It was computed as the percentage of children with TL who were also identified by FVMC, Errors/CU, or PGCUs. According to Plante and Vance (1994), sensitivity and specificity between 80% and 89% are considered acceptable, and sensitivity and specificity at or greater than 90% are considered good/preferred.

Likelihood ratios were computed from the levels of sensitivity and specificity (Dollaghan, 2007). The positive likelihood ratio (LR+) was calculated as the ratio of true LI to false LI (i.e., sensitivity/[1 − specificity]). A higher LR+ value for a positive test result refers to a higher likelihood that the positive result comes from a child with LI than from a child with TL. In contrast, the negative likelihood ratio (LR−) was calculated as the ratio of false TL to true TL (i.e., [1 − sensitivity]/specificity). A lower LR− value for a negative result refers to a lower likelihood that the negative result comes from a child with LI than from a child with TL. According to Dollaghan (2007) and

Geyman, Deyo, and Ramsey (2000), an LR+ value ≥ 10.00 or an LR− value ≤ 0.10 is considered as good/preferred, and an LR+ value between 5.00 and 9.99 or an LR− value between 0.11 and 0.20 is considered acceptable.

To compute sensitivity, specificity, and likelihood ratios, cutoff scores for a positive result were first determined by using the receiver operating characteristic (ROC) curve (Sackett, Haynes, Guyatt, & Tugwell, 1991) in the software SigmaPlot 12.0 (Systat Software, Inc., 2011). The ROC curve analysis automatically calculates pairs of sensitivity and specificity rates for a range of cutoff scores. Following Sacket et al. (1991), we chose the score that maximized the diagnostic accuracy, where sensitivity plus specificity divided by 2 is largest, as the cutoff. This procedure avoided the arbitrariness of setting the cutoff with prescriptive standards (e.g., −1.25 *SD*).

Results

Preliminary Analyses of Narrative Samples

Table 2 presents the descriptive measures from the narratives, and Table 3 presents the number of obligatory contexts of FVMC and the number of C-units that were included for computing the grammaticality measures across language groups. Children with TL produced more obligatory contexts for FVMC than those with LI for the 8-year-old group, $F(1, 65) = 5.90$, $p = .02$, $\eta_p^2 = .08$, but not for the 6-year-old group, $F(1, 59) = 0.51$, $p = .48$, $\eta_p^2 = .01$. It should be noted that even though there was a group difference for 8-year-olds, children with LI produced at least 21 obligatory contexts for FVMC. Thus, children with LI did produce sufficient obligatory contexts of FVMC in the present study. In addition, children with LI did not differ from those with TL in the number of C-units that were included for the computation of grammaticality measures at either age, *F*s < 1.12, $p > .29$, $\eta_p^2 < .02$. All children in both groups across ages produced at least 46 C-units for grammaticality measures. Note that the number of C-units that were included for computing the grammaticality measures was lower than the total number of C-units in the narratives (see Table 2) because some C-units were not included for the grammaticality analysis.

Table 2 also presents the target measures by age and language status. Children with TL produced higher FVMC, fewer Errors/CU, and higher PGCUs than those with LI at either age, *F*s > 25.80, $p < .001$, $\eta_p^2 > .28$. We also examined the age differences in the target measures for children with TL and for children with LI. Children with TL did not show age differences in FVMC, $F(1, 98) = 2.44$, $p = .12$, $\eta_p^2 = .02$, possibly because those children's FVMC scores were approaching ceiling at both ages (97% for 6-year-olds and 99% for 8-year-olds). In contrast, 8-year-olds with TL produced fewer Errors/CU and higher PGCUs than 6-year-olds with TL, *F*s > 13.72, $p < .001$, $\eta_p^2 > .12$. In a similar way, children with LI did not show an age difference in FVMC, possibly as a result of the small number of participants and the variability within each age group, $F(1, 26) = 2.63$, $p = .12$, $\eta_p^2 = .09$. Eight-year-olds with LI, however, produced fewer Errors/CU and higher PGCUs than 6-year-olds with LI, *F*s > 4.72, $p < .04$, $\eta_p^2 > .15$.

Indices of Diagnostic Accuracy of Target Measures

Table 4 presents the cutoff values and the indices of diagnostic accuracy for FVMC, Errors/CU, and PGCUs. It should be noted again that the cutoff values were derived empirically from the ROC curve analyses to maximize diagnostic accuracy for each target measure. With the cutoff values, FVMC, Errors/CU, and PGCUs had acceptable to good levels of sensitivity at age 6, in which 82%–91% of children with LI were correctly identified. At age 8,

Table 2. Mean (*SD*) and range of descriptive measures and target measures by age and language status.

	Descriptive measures			Target measures		
Group	# of C-units	MLCUm	NDW	FVMC (%)	Errors/CU	PGCU (%)
6-year-olds						
TL						
M (*SD*)	71.64 (19.25)	7.59 (0.97)	142.90 (29.74)	97 (3)	0.10 (0.08)	91 (7)
Range	50–129	5.10–10.00	94–225	88–100	0.00–0.35	65–100
LI						
M (*SD*)	73.63 (24.46)	7.04 (1.05)	123.45 (26.55)	79 (18)	0.46 (0.30)	64 (19)
Range	52–129	5.50–8.50	89–179	45–98	0.13–1.02	33–88
8-year-olds						
TL						
M (*SD*)	78.70 (21.71)	8.70 (1.07)	172.94 (42.07)	99 (2)	0.05 (0.04)	95 (3)
Range	49–146	6.73–11.08	113–279	91–100	0.00–0.16	84–100
LI						
M (*SD*)	73.94 (21.41)	7.20 (0.81)	141.47 (30.95)	88 (18)	0.25 (0.19)	78 (15)
Range	46–124	5.83–8.62	99–202	26–100	0.07–0.86	32–94

Note. # of C-units = total number of C-units in the narratives; MCLUm = mean length of C-units in morphemes; NDW = number of different words; FVMC = finite verb morphology composite; Errors/CU = number of errors per C-unit; PGCU = percent grammatical C-units; TL = children with typical language; LI = children with language impairment.

Table 3. Mean number of obligatory contexts for the finite verb morphology composites and mean number of C-units for the grammaticality measures by age and language status.

	# OC for FVMC			# C-units for grammaticality measures		
Group	*M*	*SD*	Range	*M*	*SD*	Range
6-year-olds						
TL	45.64	19.05	19–106	70.62	18.56	49–128
LI	41.36	12.01	27–60	71.18	24.16	50–122
8-year-olds						
TL	57.90	22.07	20–117	78.44	21.40	49–143
LI	43.18	20.07	21–82	72.06	21.70	46–122

Note. # OC for FVMC = number of obligatory contexts for the finite verb morphology composites; # C-units for grammaticality measures = number of C-units that were included for computing number of errors per C-unit and percent grammatical C-units; TL = typically developing children; LI = language impairment.

Errors/CU and PGCUs also had acceptable to good levels of sensitivity, in which 88%–94% of children with LI were correctly identified. FVMC, in contrast, demonstrated a sensitivity level of 76%, meaning that 24% of 8-year-old children with LI in the present study were underidentified or misclassified by FVMC. That is, although the sensitivity of Errors/CU and PGCUs was at or above the acceptable level at both ages, the sensitivity of FVMC was at the acceptable level only at age 6.

FVMC, Errors/CU, and PGCUs had acceptable to good levels of specificity at age 6, in which 82%–90% of children with TL were correctly identified. At age 8, all of these measures yielded acceptable levels of specificity, in which 80%–84% of children with TL were correctly identified. That is, the specificity of all three measures was at or above the acceptable level at both ages.

The LR+ and LR− values of FVMC, Errors/CU, and PGCUs were all at the acceptable level at age 6. The LR+ values ranged from 5.05 to 8.18, meaning that 6-year-olds with LI were 5.05 to 8.18 times more likely to obtain a fail score (i.e., below the cutoff) for the target measures than those with TL. The LR− values ranged from 0.11 to 0.20, meaning that 6-year-olds with LI were only 0.11 to 0.20 times as likely to obtain a pass score for the target measures than those with TL. At age 8, only the PGCUs reached acceptable levels for both the LR+ (5.52) and the LR− (0.14) values. Errors/CU had an unacceptable LR+ value (4.70), although it had a good LR− value (0.07). FVMC did not reach the acceptable level for either index (LR+: 3.82; LR−: 0.29).

Discussion

The present study evaluated the diagnostic accuracy of FVMC, Errors/CU, and PGCUs for differentiating school-aged children with and without LI who were 6 or 8 years old using a narrative generation task. At age 6, all three target measures demonstrated acceptable to good levels of sensitivity and specificity. The LR+/LR− values

Table 4. Indices of diagnostic accuracy by age and measure.

Age/measure	Sensitivity[a]	Specificity	Overall accuracy	LR+[b]	LR–[b]
6-year-olds					
FVMC (cutoff = 93.50%)	82%* (9/11)	90%** (45/50)	89% (54/61)	8.18*	0.20*
Errors/CU (cutoff = 0.14)	91%** (10/11)	82%* (42/50)	85% (52/61)	5.05*	0.11*
PGCU (cutoff = 83.00%)	82%* (9/11)	90%** (45/50)	89% (54/61)	8.18*	0.20*
8-year-olds					
FVMC (cutoff = 97.50%)	76% (13/17)	80%* (40/50)	79% (53/67)	3.82	0.29
Errors/CU (cutoff = 0.09)	94%** (16/17)	80%* (40/50)	84% (56/67)	4.70	0.07**
PGCU (cutoff = 91.50%)	88%* (15/17)	84%* (42/50)	85% (57/67)	5.52*	0.14*

Note. LR+ = positive likelihood ratio; LR– = negative likelihood ratio. FVMC = finite verb morphology composites; Errors/CU = number of errors per C-unit; PGCU = percent grammatical C-units.

[a] For the columns of sensitivity/specificity, a single asterisk indicates that sensitivity/specificity of a given measure reaches the acceptable level of accuracy—that is, 80% accuracy (Plante & Vance, 1994). Double asterisks indicate that sensitivity/specificity of a given measure reaches a good or preferred level of accuracy—that is, 90% accuracy. The numbers within the parentheses indicate the number of children that are correctly classified—for example, nine out of 11 six-year-olds with LI were correctly classified by FVMC. [b] For the columns of LR+/LR–, a single asterisk indicates that the LR+/LR– of a given measure reaches the acceptable level—that is, an LR+ value between 5.00 and 9.99 or an LR– value between 0.11 and 0.20 (Dollaghan, 2007; Geyman et al., 2000). Double asterisks indicate that the LR+/LR– of a given measure reaches the good level—that is, an LR+ value ≥ 10.00 or an LR– value ≤ 0.10.

were at the acceptable levels for all three measures. At age 8, FVMC showed sensitivity below the acceptable level. The LR+/LR− values for FVMC were both below the acceptable level as well. In contrast, Errors/CU and PGCUs showed acceptable to good levels of sensitivity and specificity. However, the LR+ value for Errors/CU was below the acceptable level. We explore these findings below.

Diagnostic Accuracy of FVMC

In this study, we found that the diagnostic accuracy of FVMC, a tense measure, was unacceptably low for 8-year-olds. FVMC tended to underidentify 8-year-olds with LI. This finding was consistent with Moyle et al. (2011), who concluded that tense measures computed from language samples are not useful tools for diagnosing school-aged children with LI. On the other hand, we also found that FVMC yielded acceptable diagnostic accuracy for 6-year-olds, which was consistent with Souto et al. (2014). Together, the findings from Souto et al. (2014) and the present study show that FVMC remains an appropriate tool for differentiating school-aged children with and without LI who are 5 or 6 years old, at least in language samples that are collected by means of similar procedures.

If tense deficits are a potential clinical marker of children with LI (Rice & Wexler, 1996), why did FVMC show diagnostic accuracy below the acceptable level for 8-year-olds in the present study? One possibility is that, despite no significant age differences in FVMC for children with LI, proportionally more 8-year-olds with LI produced FVMC at the customary level of mastery (i.e., 90% accurate) than 6-year-olds with LI. For children with LI, four (36%) of 11 children in the 6-year-old group produced FVMC at the customary level of mastery, whereas 12 (71%) of 17 children in the 8-year-old group did so in the present study. This difference seems to suggest that children with LI still showed improvement in tense usage from age 6 to age 8 in the present study. Some children with LI may even overcome difficulty producing tense morphemes by age 8. This improvement, in turn, may lead to greater overlapping in FVMC between the TL and LI groups, and consequently lower sensitivity of FVMC, at age 8 than at age 6. However, we are not claiming that the majority of children with LI would grow out of tense deficits by age 8. Instead, we suggest that FVMC computed from the current sampling context may be a sensitive tool for identifying children with LI up to age 6.

Diagnostic Accuracy of Grammaticality Measures

In the present study, we also found that both Errors/CU and PGCUs showed acceptable to preferred levels of sensitivity and specificity for both 6- and 8-year-olds. This finding was consistent with Souto et al. (2014), who found that percent grammatical utterances showed good sensitivity and specificity for school-aged children who were 5 years old. To the best of our knowledge, the diagnostic accuracy of Errors/CU has never been determined, although Errors/CU has been used to characterize grammatical development in children with LI (Colozzo et al., 2011; Reilly et al., 2004; Scott & Windsor, 2000). The present study might be the first to provide empirical evidence to support the use of Errors/CU in identifying school-aged children with LI. However, despite the fact that Errors/CU counts each error in the narratives and is presumably a more fine-grained analysis than PGCUs, Errors/CU did not show higher diagnostic accuracy than PGCUs in the present study. Contrary to our expectation, the LR+ value for Errors/CU at age 8 was even below the acceptable level, whereas the LR+ value for PGCUs was at the acceptable level. Taken together, the present findings suggest that Errors/CU and PGCUs both are appropriate tools for differentiating children with and without LI who are 6 years old but only PGCUs are appropriate for identifying 8-year-olds with LI when sensitivity, specificity, and likelihood ratios are considered as a whole.

Among the three measures, PGCUs showed a clear advantage over FVMC in the diagnostic accuracy for 8-year-olds. Why would this be? Although tense deficits are a potential clinical marker of children with LI, these children also demonstrate other grammatical deficits (Leonard, 2014). Because PGCUs take tense morphemes as well as other morphological and syntactic structures (e.g., plural *–s*, conjunctions, relative pronouns, argument structures, word orders) into consideration, PGCUs further reflect the morphological and syntactic deficits that are not evaluated by FVMC. It is possible that PGCUs would be more sensitive to LI than FVMC, especially when children's performance on FVMC has approached or reached the customary level of mastery in particular language sampling contexts.

Limitations

One limitation of the present study is that we had a small number of children for both ages. In particular, we had a small number of children with LI in an attempt to avoid overrepresenting children with LI in the norm for the ENNI (Schneider, Hayward, & Dubé, 2006). One problem of including a small number of children with LI is that the differences in sensitivity and specificity across ages and measures can be easily overinterpreted. For instance, we had 17 children with LI in the 8-year-old group. Misclassification of one child with LI could lead to approximately a 6% difference in the level of sensitivity. Consider the sensitivity for FVMC (76%) and for PGCUs (88%). The 12% difference occurred because FVMC misclassified two more children with LI than PGCUs. It is possible that the differences between these measures could be smaller or the sensitivity of FVMC for 8-year-olds could reach the acceptable level when more children are included.

In addition, we did not measure children's nonverbal intelligence, although there were no concerns regarding their nonverbal cognitive skills from the teachers and clinicians who referred children to the present study or from the examiners. The information regarding whether children with LI also had ADD/ADHD, speech disorders, or motor delay was also not available because access to children's clinical records was not obtained. Thus, the current findings may not be generalized to children with LI who clearly do not have ADD/ADHD, speech disorders, or motor delay. However, one could argue that the current findings could still be generalized to the clinical population that resembles the profile of children in the present study.

Furthermore, the present study focused on the language sample measures that evaluated the accuracy of sentence elements (e.g., tense morphemes, conjunctions) and did not examine those that evaluated the complexity of sentence structures (e.g., subordination index; Ebert & Scott, 2014). One concern is that we may have overlooked the complexity measures that could yield good diagnostic accuracy given that school-aged children with LI tend to perform more poorly than those with TL on complexity measures (e.g., Colozzo et al., 2011; Ebert & Scott, 2014; Souto et al., 2014). As Nelson et al. (2014) suggested, measures that show group differences between children with and without LI do not mean that those measures will also show good diagnostic accuracy. Indeed, Souto et al. (2014) showed that complexity measures did not show higher diagnostic accuracy than the accuracy measures for young school-aged children. However, given that sentence complexity continues to develop throughout school ages (Nippold et al., 2014), future studies that compare the diagnostic accuracy of complexity measures and accuracy measures for older school-aged children would be a worthwhile pursuit.

Last, we did not evaluate the diagnostic accuracy of combined language sample measures (e.g., Heilmann et al., 2010; Moyle et al., 2011). This decision was made for two reasons. First, combined language sample measures do not necessarily yield higher diagnostic accuracy than a single measure (e.g., PGCUs). For instance, Heilmann et al. (2010) examined the diagnostic accuracy of a combination of 10 language sample measures in differentiating children between the ages of 6;0 and 9;11. The combination of 10 measures yielded a sensitivity of 80%, a specificity of 85%, an LR+ value of 5.33, and an LR− value of 0.24, which was similar to the findings for PGCUs in the present study. Second, to the best of our knowledge, when the diagnostic accuracy of combined measures is examined, the statistical model (i.e., discriminant function analysis) does not automatically generate the cutoff score of each measure in the combination for clinical decision making. Even if the model can generate the cutoff score for each measure, there is currently no guideline for the clinician with regard to how to use the cutoff scores from multiple measures for diagnosis (e.g., should a child be diagnosed as having LI only when he scores below the cutoff for all of the 10 measures?). In addition, using more measures may introduce more errors in making diagnostic decisions. This is because when more measures are included in the diagnostic process, the likelihood that an individual would fall into the clinical range on at least one measure also increases. Thus, although using a combination of language sample measures for diagnosis may be promising, the clinical application of this approach remains unclear until further research is conducted.

Clinical Implications

A number of studies have consistently indicated that FVMC is an appropriate tool for identifying children with LI between 3;0 and 5;10 (Bedore & Leonard, 1998; Eisenberg & Guo, 2013; Guo & Eisenberg, 2014; Souto et al., 2014). The present study extended previous studies and found that FVMC remained useful for identifying school-aged children with LI who were 6 years old. Together, previous studies and the present study suggest that the clinician may compute FVMC from language samples to supplement norm-referenced, standardized tests for differentiating children with and without LI between 3;0 and 6;11. If a clinician uses the same narrative task and coding procedure from the present study for identifying 6-year-olds with LI who had similar characteristics to children in the current study (e.g., coming from an English-speaking family, may or may not have mild/moderate speech disorders), a cutoff FVMC of 93.5% is recommended.

At this point of time, there is no trustworthy evidence indicating that FVMC computed from language samples can be used for children beyond age 6. However, it does not necessarily mean that the clinician should not evaluate FVMC for children who are older than 6 years. Because tense deficits are a hallmark of children with LI, some children with LI who are older than 6 years may still show difficulties using tense morphemes in spoken discourse. If the production of tense morphemes in spoken discourse is a treatment goal for those children, FVMC may be computed to monitor treatment progress.

Similar to FVMC, previous studies (Eisenberg & Guo, 2013; Souto et al., 2014) have demonstrated that PGCUs are a useful tool for identifying children with LI between 3;0 and 5;10. The present study further shows that PGCUs remain appropriate for identifying children with LI who are 6 or 8 years old, although this finding is based on a small number of children. Taken together, these studies suggest that the clinician may include PGCUs as a measure to supplement norm-referenced, standardized tests for identifying children with LI who are between 3;0 and 8;11. For children who are 6;0 or older, a narrative generation task may be needed for the computation of PGCU. If a clinician uses the same narrative task and coding procedure from the present study for identifying children with LI who share similar characteristics with those in the current study, a cutoff PGCU of 83.0% can be used for determining the language status of 6-year-olds and 91.5% for 8-year-olds.

In addition, although Errors/CU did not show a clear advantage over PGCUs in general, the clinician can still identify the errors that children make in the language samples and then conduct in-depth analyses for the errors on the basis of the classification system in the present study. If any patterns emerge in the error analysis, the patterns can be chosen as therapy goals (Paul & Norbury, 2012).

Last, despite the usefulness of FVMC, Errors/CU, and PGCUs, they should not be the only assessment tools in the evaluation process given that the diagnostic accuracy is not high enough. For instance, the sensitivity of PGCUs was 82% for 6-year-olds and 88% for 8-year-olds. It means that PGCU underidentifies 18% of 6-year-olds with LI and 12% of 8-year-olds with LI. Thus, FVMC, Errors/CU, and PGCUs should be used in combination with norm-referenced standardized tests that demonstrate superior diagnostic accuracy to those measures (e.g., the Structured Photographic Expressive Language Test–Third Edition; Dawson, Stout, & Eyer, 2003) in the diagnostic process to generate convergent evidence for children.

Conclusion

Including language sample measures for diagnosis not only adds ecological validity to the assessment process, but also augments the identification accuracy of norm-referenced, standardized tests (Costanza-Smith, 2010). However, empirical evidence for the use of language sample measures to identify school-aged children with LI remains sparse. The current study addresses this clinical issue by providing empirical evidence regarding the diagnostic accuracy of FVMC, Errors/CU, and PGCUs. We conclude that FVMC and Errors/CU could be appropriate diagnostic tools up to age 6 in language samples that are collected with procedures similar to those of the present study. The grammaticality measure, PGCUs, could be a sensitive tool in identifying children with LI at least up to age 8. It should be noted that we are not claiming that clinicians should only consider FVMC, Errors/CU, and PGCUs in the diagnosis process. Instead, clinicians should also consider other language sample measures, such as subordination index (Ebert & Scott, 2014) and story grammar scoring (Schneider et al., 2006), depending on the purpose of assessment and the available evidence in the literature.

Acknowledgment

The development of the Edmonton Narrative Norms Instrument was supported by a grant from the Children's Health Foundation of Northern Alberta, awarded to Phyllis Schneider. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Children's Health Foundation of Northern Alberta. We are grateful to the children who participated and the teachers and clinicians who referred children to the present study. We also thank Amy Briggs and Sanjana Nair for coding the data and Dr. Linda Spencer for her suggestion on this article.

References

Aram, D. M., Morris, R., & Hall, N. E. (1993). Clinical and research congruence in identifying children with specific language impairment. *Journal of Speech and Hearing Research, 36,* 580–591. doi:10.1044/jshr.3603.580

Bedore, L. M., & Leonard, L. B. (1998). Specific language impairment and grammatical morphology: A discriminant function analysis. *Journal of Speech, Language, and Hearing Research, 41,* 1185–1192. doi:10.1044/jslhr.4105.1185

Blishen, B. R., Carroll, W. K., & Moore, C. (1987). The 1981 socioeconomic index for occupations in Canada. *Canadian Review of Sociology and Anthropology, 24,* 465–488.

Carrow-Woolfolk, E. (1985). *Test for Auditory Comprehension of Language—Revised*. Austin, TX: Pro-Ed.

Coady, J. A., Evans, J. L., & Kluender, K. R. (2010). The role of phonotactic frequency in sentence repetition by children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 53,* 1401–1415. doi:10.1044/1092-4388

Colozzo, P., Gillam, R. B., Wood, M., Schnell, R., & Johnston, J. R. (2011). Content and form in the narratives of children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 54,* 1609–1627. doi:10.1044/1092-4388 (2011/10-0247)

Conti-Ramsden, G., Botting, N., & Faragher, B. (2001). Psycholinguistic markers for specific language impairment (SLI). *Journal of Child Psychology and Psychiatry, 42,* 741–748.

Costanza-Smith, A. (2010). The clinical utility of language samples. *SIG 1: Perspectives on Language Learning and Education, 17,* 9–15. doi:10.1044/lle17.1.9

Dawson, J. I., Stout, C. E., & Eyer, J. A. (2003). *The Structured Photographic Expressive Language Test–Third Edition*. Dekalb, IL: Janelle Publications.

Dollaghan, C. (2007). *The handbook for evidence-based practice in communication disorders*. Baltimore, MD: Brookes.

Dunn, L., & Dunn, L. (1981). *Peabody Picture Vocabulary Test–Revised*. Circle Pines, MN: AGS.

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

Ebbels, S. H., Van Der Lely, H. K., & Dockrell, J. E. (2007). Intervention for verb argument structure in children with persistent SLI: A randomized control trial. *Journal of Speech, Language, and Hearing Research, 50,* 1330–1349. doi:10.1044/1092-4388(2007/093)

Ebert, K. D., & Scott, C. M. (2014). Relationships between narrative language samples and norm-referenced test scores in language assessments of school-age children. *Language, Speech, and Hearing Services in Schools, 45,* 337–350.

Eisenberg, S. L., & Guo, L.-Y. (2013). Differentiating children with and without language impairment based on grammaticality. *Language, Speech, and Hearing Services in Schools, 44,* 20–31. doi:10.1044/0161-1461(2012/11-0089)

Fey, M. E., Catts, H. W., Proctor-Williams, K., Tomblin, J. B., & Zhang, X. (2004). Oral and written story composition skills of children with language impairment. *Journal of Speech, Language, and Hearing Research, 47,* 1301–1318. doi:10.1044/1092-4388(2004/098)

Geyman, J., Deyo, R., & Ramsey, S. (2000). *Evidence-based clinical practice—Concepts & approaches*. Boston, MA: Butterworth-Heinemann.

Gladfelter, A., & Leonard, L. B. (2013). Alternative tense and agreement morpheme measures for assessing grammatical deficits during the preschool period. *Journal of Speech, Language, and Hearing Research, 56,* 542–552. doi:10.1044/1092-4388 (2012/12-0100)

Guo, L.-Y., & Eisenberg, S. (2014). The diagnostic accuracy of two tense measures for identifying 3-year-olds with language impairment. *American Journal of Speech-Language Pathology, 23,* 203–212. doi:10.1044/2013_AJSLP-13-0007

Halliday, M. A. K. (1994). *Introduction to functional grammar* (2nd ed.). London, UK: Edward Arnold.

Hammill, D., & Newcomer, P. (1988). *Test of Language Development–Intermediate* (2nd ed.). Austin, TX: Pro-Ed.

Heilmann, J. J., Miller, J. F., & Nockerts, A. (2010). Using language sample databases. *Language, Speech, and Hearing Services in Schools, 41,* 84–95. doi:10.1044/0161-1461(2009/08-0075)

Hunt, K. W. (1965). *Grammatical structures written at three grade levels* (NCTE Research Report No. 3). Washington, DC: Office of Education.

Lee, L. (1974). *Developmental sentence analysis: A grammatical assessment procedure for speech and language clinicians*. Evanston, IL: Northwestern University Press.

Leonard, L. B. (2014). *Children with specific language impairment* (2nd ed.). Cambridge, MA: MIT Press.

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

Longman Dictionary of Contemporary English (6th ed.). (2014). Harlow, UK: Pearson Longman.

Miller, J., & Chapman, A. (2000). *Systematic Analysis of Language Transcripts* (Version 6.1) [Computer software]. Madison, WI: SALT Software.

Moore, M. E. (2001). Third person pronoun errors by children with and without language impairment. *Journal of Communication Disorders, 34,* 207–228.

Moyle, M., Karasinski, C., Ellis Weismer, S., & Gorman, B. (2011). Grammatical morphology in school-age children with and without language impairment: A discriminant function analysis. *Language, Speech, and Hearing Services in Schools, 42,* 550–560. doi:10.1044/0161-1461(2011/10-0029)

Munson, B., Kurtz, B. A., & Windsor, J. (2005). The influence of vocabulary size, phonotactic probability, and wordlikeness on nonword repetitions of children with and without specific language impairment. *Journal of Speech, Language, and Hearing Research, 48,* 1033–1047. doi:10.1044/1092-4388(2005/072)

Nelson, N., Plante, E., & Anderson, M. (2014). *Assessing oral and written language among school-age students: Issues of validity*. Seminar presented at the 2014 Annual Convention of the American Speech-Language-Hearing Association, Miami, FL.

Nippold, M. A., Frantz-Kaspar, M. W., Cramond, P. M., Kirk, C., Hayward-Mayhew, C., & MacKinnon, M. (2014). Conversational and narrative speaking in adolescents: Examining the use of complex syntax. *Journal of Speech, Language, and Hearing Research, 57,* 876–886. doi:10.1044/1092-4388(2013/13-0097)

Oetting, J. B., & Hadley, P. A. (2009). Morphosyntax in child language disorders. In R. Schwartz (Ed.), *Handbook of child language disorders* (pp. 341–364). New York, NY: Psychology Press.

Paul, R., & Norbury, C. F. (2012). *Language disorders from infancy through adolescence: Listening, speaking, reading, writing, and communicating* (4th ed.). St. Louis, MO: Elsevier.

Plante, E., & Vance, R. (1994). Selection of preschool language tests: A data-based approach. *Language, Speech, and Hearing Services in Schools, 25,* 15–24.

Poll, G. H., Betz, S. K., & Miller, C. A. (2010). Identification of clinical markers of specific language impairment in adults. *Journal of Speech, Language, and Hearing Research, 53,* 414–429. doi:10.1044/1092-4388(2009/08-0016)

Reilly, J., Losh, M., Bellugi, U., & Wulfeck, B. (2004). “Frog, where are you?” Narratives in children with specific language impairment, early focal brain injury, and Williams syndrome. *Brain and Language, 88,* 229–247.

Rice, M., & Wexler, K. (1996). Toward tense as a clinical marker of specific language impairment in English-speaking children. *Journal of Speech, Language, and Hearing Research, 39,* 1239–1257. doi:10.1044/jshr.3906.1239

Rice, M., & Wexler, K. (2001). *Test for Early Grammatical Impairment*. San Antonio, TX: The Psychological Corporation.

Sackett, D., Haynes, R. B., Guyatt, G. H., & Tugwell, P. (1991). *Clinical epidemiology: A basic science for clinical medicine* (2nd ed.). Boston, MA: Little Brown.

Saeed, J. (2009). *Semantics* (3rd ed.). Malden, MA: Wiley-Blackwell.

Schneider, P., Dubé, R. V., & Hayward, D. (2005). *The Edmonton Narrative Norms Instrument*. Available from http://www.rehabmed.ualberta.ca/spa/enni/

Schneider, P., Hayward, D., & Dubé, R. V. (2006). Storytelling from pictures using the Edmonton narrative norms instrument. *Journal of Speech Language Pathology and Audiology, 30,* 224–238.

Scott, C. M., & Windsor, J. (2000). General language performance measures in spoken and written narrative and expository discourse of school-age children with language learning disabilities. *Journal of Speech, Language, and Hearing Research, 43,* 324–339.

Semel, E., Wiig, E. H., & Secord, W. A. (1995). *Clinical Evaluation of Language Fundamentals–Third Edition*. San Antonio, TX: Psychological Corporation.

Semel, E., Wiig, E. H., & Secord, W. A. (2003). *Clinical Evaluation of Language Fundamentals–Fourth Edition*. San Antonio, TX: Pearson.

Shriberg, L. D., Kwiatkowski, J., & Hoffman, K. (1984). A procedure for phonetic transcription by consensus. *Journal of Speech and Hearing Research, 27,* 456–465.

Souto, S. M., Leonard, L. B., & Deevy, P. (2014). Identifying risk for specific language impairment with narrow and global measures of grammar. *Clinical Linguistics & Phonetics, 28,* 741–756. doi:10.3109/02699206.2014.893372

Spaulding, T. J., Plante, E., & Farinella, K. A. (2006). Eligibility criteria for language impairment: Is the low end of normal always appropriate? *Language, Speech, and Hearing Services in Schools, 37,* 61–72. doi:10.1044/0161-1461(2006/007)

Statistics Canada. (n.d.). *Canada dimensions: The people*. Retrieved from http://www.statcan.ca

Systat Software, Inc. (2011). SigmaPlot 12.0. Point Richmond, CA: Author.

Tager-Flusberg, H., & Cooper, J. (1999). Present and future possibilities for defining a phenotype for specific language impairment. *Journal of Speech, Language, and Hearing Research, 42,* 414–429. doi:10.1044/jslhr.4205.1275

Thordardottir, E. (2008). Language-specific effects of task demands on the manifestation of specific language impairment: A comparison of English and Icelandic. *Journal of Speech, Language, and Hearing Research, 51,* 922–937. doi:10.1044/1092-4388(2008/068)

Werner, E. O., & Kresheck, J. D. (1983). *Structured Photographic Expressive Language Test–Second Edition*. DeKalb, IL: Janelle.

Windsor, J., Scott, C. M., & Street, C. K. (2000). Verb and noun morphology in the spoken and written language of children with language learning disabilities. *Journal of Speech, Language, and Hearing Research, 43,* 1322–1336. doi:10.1044/jslhr.4306.1322