AJSLP

Research Article

Percent Grammatical Utterances Between 4 and 9 Years of Age for the Edmonton Narrative Norms Instrument: Reference Data and Psychometric Properties

Ling-Yu Guo,[a,b] Sarita Eisenberg,[c] Phyllis Schneider,[d] and Linda Spencer[e]

Purpose: The purpose of this article was to provide the reference data and evaluate psychometric properties for the percent grammatical utterances (PGU; Eisenberg & Guo, 2013) in children between 4 and 9 years of age from the database of the Edmonton Narrative Norms Instrument (ENNI; Schneider, Dubé, & Hayward, 2005).
Method: Participants were 377 children who were between 4 and 9 years of age, including 300 children with typical language (TL) and 77 children with language impairment (LI). Narrative samples were collected using the ENNI protocol (i.e., a story generation task). PGU was computed from the samples. Split-half reliability, concurrent criterion validity, and diagnostic accuracy for PGU were further evaluated.
Results: PGU increased significantly in children between 4 and 9 years of age in both the TL and LI groups. In addition, the correlation coefficients for the split-half reliability and concurrent criterion validity of PGU were all large ($rs \geq .557$, $ps < .001$). The diagnostic accuracy of PGU was also good or acceptable from ages 4 to 9 years.
Conclusions: With the attested psychometric properties, PGU computed from the ENNI could be used as an assessment tool for identifying children with LI between 4 and 9 years of age. The reference data of PGU could also be used for monitoring treatment progress.
Supplemental Material: https://doi.org/10.23641/asha.9630590

Grammatical deficits are a hallmark of English-speaking children with language impairment (LI; Leonard, 2014). As compared to children with typical language (TL), children with LI tend to show reduced productivity and complexity in using morphological and syntactic structures in spoken discourse (e.g., Eisenberg, 2003; Hewitt, Hammer, Yont, & Tomblin, 2005). Children with LI also demonstrate lower accuracy in producing grammatical structures than children with TL (e.g., Souto, Leonard, & Deevy, 2014). The grammatical errors produced by children with LI may include, but are not limited to, tense marking errors (Leonard, Haebig, Deevy, & Brown, 2017; Rice, Wexler, & Hershberger, 1998), personal and relative pronoun errors (Moore, 2001; Schuele & Tolbert, 2001), argument structure errors (Ebbels, van der Lely, & Dockrell, 2007; Grela & Leonard, 1997), and errors with grammatical morphemes other than pronouns and tense markers (e.g., infinitive *to,* dative preposition *to,* particles such as *put on the shirt*; Arndt & Schuele, 2012; Grela, Rashiti, & Soares, 2004; Watkins & Rice, 1991). The difficulty in accurately using grammatical structures by children with LI is observed during the preschool years and may persist into the school-age years (Guo & Schneider, 2016; Lee, 1974).

Because even just one grammatical error would make a sentence ungrammatical, evaluating the extent to which children are able to produce grammatical sentences may reflect their ability in using grammatical structures. Lee (1974) first created a grammaticality measure—sentence point—to count the number of grammatical sentences a child produced in 50 different sentences extracted from

[a]Department of Audiology and Speech-Language Pathology, Asia University, Taichung City, Taiwan
[b]Department of Communicative Disorders and Sciences, University at Buffalo, NY
[c]Department of Communication Sciences and Disorders, Montclair State University, Bloomfield, NJ
[d]Department of Communication Sciences and Disorders, University of Alberta, Edmonton, Canada
[e]Master of Science in Speech-Language Pathology Program, Rocky Mountain University of Health Professions, Provo, UT
Correspondence to Ling-Yu Guo: lingyugu@buffalo.edu

Editor-in-Chief: Julie Barkmeier-Kraemer
Editor: Stacy Betz
Received September 23, 2018
Revision received February 18, 2019
Accepted May 11, 2019
https://doi.org/10.1044/2019_AJSLP-18-0228

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

language samples. More recently, Eisenberg and colleagues (Eisenberg & Guo, 2013; Eisenberg, Guo, & Germezia, 2012) adapted the sentence point score and developed "percent grammatical utterances (PGU)" as a broad grammaticality measure for identifying children with LI. An important difference between the two measures is the utterance inclusion criteria. Lee only included utterances that contained both a subject and a verb for calculating the sentence point score. This means that utterances with subject argument omissions (e.g., *Played by the pool*) or copula omissions (e.g., *She mad*), two common errors by children with LI (Grela & Leonard, 1997; Rice et al., 1998), are excluded from the sentence point analysis. In contrast, PGU includes utterances that obligate both a subject a and verb but does not require that both be produced.

Across studies, measures of grammatical accuracy have been shown to differentiate children with and without LI during their preschool and early school years (Eisenberg & Guo, 2013; Fey, Catts, Proctor-Williams, Tomblin, & Zhang, 2004; Guo & Schneider, 2016; Scott & Windsor, 2000; Souto et al., 2014; Westerveld & Gillon, 2010). However, existing studies that have examined grammatical accuracy in preschool and school-age children have utilized different sampling protocols (e.g., narration, free play, picture description task) as well as different methods for calculating grammatical accuracy. This makes it difficult to compare results across studies and apply the data to clinical decision making for identifying children with LI. By reanalyzing the archival data of the Edmonton Narrative Norms Instrument (ENNI; Schneider, Dubé, & Hayward, 2005), the present study aimed to provide reference data for one measure of grammatical accuracy, PGU, in children between 4 and 9 years of age who were administered a consistent language sampling protocol (i.e., a story generation task). To validate the use of PGU computed from the ENNI protocol, we also evaluated the psychometric properties of PGU (i.e., split-half reliability, concurrent criterion validity, and diagnostic accuracy). In what follows, we first review the studies that investigated the development of grammatical accuracy in English-speaking children with and without LI. We then lay out the scope of the present study.

Grammatical Accuracy in Children With and Without LI

Eisenberg and Guo (2013) compared two measures of grammatical accuracy—PGU and percent sentence point (PSP)—in 3-year-olds with and without LI using a picture description task. In this task, children were asked to describe 15 pictures by responding to four elicitation questions for each picture. Children's responses were segmented into communication units (C-units), defined as an independent clause plus any number of dependent clauses. For PGU, all C-units that obligated both a subject and a verb were included in the analysis. For PSP, only utterances that included both of these constituents were included, following Lee's (1974) utterance inclusion criteria. For both PGU and PSP analyses, each of the included C-units was coded as grammatical or ungrammatical. Both PGU and PSP scores were computed by dividing the number of grammatical C-units by the total number of C-units that were included for that analysis and multiplying the resultant quotient by 100%. The mean PGU score was 72% (*SD* = 12%) for 3-year-olds with TL and 38% (*SD* = 12%) for those with LI. The mean PSP score was 75% (*SD* = 12%) for 3-year-olds with TL and 45% (*SD* = 12%) for those with LI. Both PGU and PSP scores were significantly higher in the TL group than in the LI group.

Souto et al. (2014) further examined the sentence point score in 4- and 5-year-olds with and without LI using conversational language samples that involved the child playing with toys with an examiner. Based on the sentence inclusion rules in Lee (1974), 50 consecutive, different, and complete sentences (i.e., sentences with both a subject and a verb) were selected and evaluated for grammaticality. The authors did not indicate the unit of segmentation; however, because they followed Lee's guidelines, it is likely that they used phonological unit segmentation, which allows up to two independent clauses per utterance. Each grammatical sentence was awarded a sentence point. Mean sentence points were computed by dividing the total number of sentence points by 50. To compare across studies, we converted the mean sentence point scores into a percentage score (i.e., PSP). The mean PSP score was 91% (*SD* = 8%) for 4-year-old children with TL and 60% (*SD* = 14%) for those with LI. Moreover, the mean PSP score was 93% (*SD* = 3%) for 5-year-old children with TL and 70% (*SD* = 9%) for those with LI. At both ages, the PSP score was significantly higher in the TL group than in the LI group.

Extending prior studies, Guo and Schneider (2016) examined PGU in 6- and 8-year-olds with and without LI using a story generation task. In this task, children were asked to tell stories based on six picture sequences. Children's narratives were segmented into C-units. The utterance inclusion criterion was modified from Eisenberg and Guo (2013) to exclude fragments (e.g., *A giraffe and an elephant*). That is, each C-unit had to have at least a verb in order to be included in the PGU analysis. The mean PGU score was 91% (*SD* = 7%) for 6-year-old children with TL and 64% (*SD* = 19%) for those with LI. In addition, the mean PGU score was 95% (*SD* = 3%) for 8-year-old children with TL and 78% (*SD* = 15%) for those with LI. At both ages, the PGU score was significantly higher in the TL group than in the LI group.

In contrast to Guo and Schneider (2016), Fey et al. (2004) used a story generation task to collect not only spoken narratives but also written narratives to examine grammaticality scores from second and fourth graders with and without LI. Children's narratives were segmented into C-units. No information was provided about the utterance inclusion criteria, so it is unknown whether C-units without a subject and/or verb were included. Collapsing spoken and written narratives together, Fey et al. reported a mean grammaticality score of 86% (*SD* = 11%) for second-grade children with TL and 78% (*SD* = 13%) for those with LI. The mean grammaticality score was 84% (*SD* = 13%)

for fourth-grade children with TL and 75% (*SD* = 14%) for those with LI. Children with TL thus showed lower grammaticality in Fey et al. than those in Guo and Schneider even though children with TL in Fey et al. would have been a year older. In addition, Fey et al. found no increase in grammatical accuracy between second and fourth graders and a smaller difference in grammatical accuracy scores between the TL and LI groups as compared to Guo and Schneider.

Different from other studies (e.g., Fey et al., 2004; Guo & Schneider, 2016), Westerveld and Gillon (2010) investigated grammatical accuracy only in school-age children with TL. Children who were 5, 6, or 7 years old completed a story retelling task. The retold narratives were segmented into C-units, but there was no information about the utterance inclusion criteria. Grammatical accuracy was 86.2% (*SD* = 14.9%) at the age of 5 years, 87.9% (*SD* = 11.0%) at the age of 6 years, and 91.8% (*SD* = 9.8%) at the age of 7 years. Thus, grammatical accuracy at the age of 6 years was slightly lower than that reported by Guo and Schneider (2016).

Across the studies, grammatical accuracy increased between the ages of 3 and 8 years, and children with LI had lower accuracy than children with TL. However, extant studies used different language sampling protocols to generate reference data (e.g., mean, standard deviation) for grammatical accuracy measures from children with TL between the preschool and early school–age years. Eisenberg and Guo (2013) used a picture description task for 3-year-old children; Souto et al. (2014) used free play for 4- and 5-year-old children; Westerveld and Gillon (2010) used a story retelling task for 5-, 6-, and 7-year-old children; and both Fey et al. (2004) and Guo and Schneider (2016) used a story generation task for older children. In addition, the utterance inclusion criteria, and hence the scoring rules, for calculating grammatical accuracy have also varied across studies. For example, Souto et al. included only utterances that had at least a subject and a verb, whereas Guo and Schneider included utterances that had at least a verb in the analysis. Thus, a sentence like "*Dropped the airplane by accident*" would be excluded in Souto et al. but would be included and counted as ungrammatical (i.e., omission of sentence subject) based on the rules from Guo and Schneider.

Without separate reference data for different age levels based on the same language sampling protocol and utterance inclusion criteria, it is difficult for clinicians to determine whether a child's grammatical accuracy is within the typical range or to evaluate whether a child makes significant improvement on grammatical accuracy over time. This issue is critical because practicing clinicians report they do not regularly include language sample analysis in the assessment process due to an absence of reference data (Pavelko, Owens, Ireland, & Hahs-Vaughn, 2016). Although Lee (1974) provided group data for grammatical accuracy (i.e., sentence point scores) for children between 2;0 and 6;11 (years;months) in 12-month intervals, only group means, but not standard deviations, for the sentence point scores were reported. The data, therefore, cannot be readily used for clinical purposes. Thus, although a number of studies have reported grammatical accuracy data for different age groups, there remains a critical need to obtain reference data for grammatical accuracy from children with TL at different ages between preschool and early school–age years using the same language sampling protocol and utterance inclusion criteria.

The Present Study

Although previous studies have collectively shown that measures of grammatical accuracy could be a clinically useful tool for differentiating children with and without LI, the lack of reference data and the discrepancies in language sampling protocol and utterance inclusion criteria have undermined the use of grammatical accuracy measures in the assessment process. To address these issues, the present study aimed to provide the reference data for one measure of grammatical accuracy, PGU, in children between 4 and 9 years of age from the database of the ENNI (Schneider et al., 2005). The ENNI was originally designed to evaluate children's narrative skills between the ages of 4 and 9 years using picture sequences. We chose the archival data from the ENNI database for the PGU analysis because it was based on a consistent story generation protocol across the designated age range. The assessment protocol and reference data for a number of macrostructural and microstructural measures (e.g., story grammar, mean length of utterances, and number of different words) are publicly available on the ENNI website. However, the existing reference data do not include PGU. This provides an opportunity for the present study to perform the PGU analysis on the archival data, which are accessible on the website of the Child Language Data Exchange System (MacWhinney, 2000).

Using narrative, rather than conversation, as the sampling genre has two advantages. First, although conversational discourse during play is widely used in clinical work with young children, this sampling context is not appropriate for school-age children. Instead, interview-based conversations seem to be more appropriate than play-based conversations for the school-age population (see Hadley, 1998). However, prior studies have shown that different sampling contexts (e.g., play, interview) could lead to significant differences in language sample measures (e.g., Southwood & Russell, 2004) and, hence, confound the age-related changes for these measures. In the ENNI database, all of the children were asked to tell stories based on six sets of pictures, which allowed us to keep the language sampling protocol the same for children between the preschool and early school ages. Second, narrative is cognitively more demanding than conversation (Johnston, 2006). Grammatical weaknesses, such as tense marking errors, are more likely to appear in narratives than in conversations (Thordardottir, 2008). It could be relatively easier for clinicians to identify children with LI using narrative than using conversation.

The present study had two specific aims. First, we calculated not only the means and standard deviations of PGU but also 90% and 95% confidence intervals of PGU for children with TL between the ages of 4 and 9 years in

12-month intervals as the reference data. Computing the means and standard deviations for PGU by age and determining whether PGU scores increased with age would establish the initial validity for PGU (i.e., construct validity; Aiken & Groth-Marnat, 2006). Data about age-related changes in PGU could also inform theories about grammatical development in older children. In addition, few studies, if any, have reported confidence intervals for language sample measures. The provision of confidence intervals for PGU would allow clinicians to estimate the range of a child's "true" score for PGU, which would help clinicians interpret the results in the diagnosis process (McCauley & Swisher, 1984). When treatment focuses on grammatical structures and PGU is used as an outcome measure, the availability of confidence intervals would also allow clinicians to determine whether a child's performance on PGU significantly improves over time.

Second, to further validate the use of PGU computed from the ENNI protocol, we also evaluated the psychometric properties for PGU at each age level, including split-half reliability, concurrent criterion validity, and diagnostic accuracy. Split-half reliability of a measure evaluates the extent to which the results of one half are consistent with those of the other half. To this end, we computed the correlation of PGU from the first set of stories (i.e., Set A) and PGU from the second set of stories (i.e., Set B) in the ENNI protocol. Concurrent criterion validity of a measure evaluates the extent to which this measure correlates with other measures designed to assess the same skill areas. To this end, we computed the correlation between PGU and a measure of expressive grammar—the Recalling Sentences in Contexts subtest of the Clinical Evaluation of Language Fundamentals–Preschool (CELF-P; Wiig, Secord, & Semel, 1992) or the Recalling Sentences subtest of the Clinical Evaluation of Language Fundamentals–Third Edition (CELF-3; Semel, Wiig, & Secord, 1995). In addition, we computed sensitivity, specificity, and likelihood ratios to evaluate the diagnostic accuracy for PGU computed from the ENNI protocol.

It should be noted that the means, standard deviations, and diagnostic accuracy for PGU in 6- and 8-year-olds have been reported in Guo and Schneider (2016), whereas the rest of the data were unique to the present study and have never been reported elsewhere. Thus, the present study contributed to evidence-based assessment by expanding the reference data for PGU in Guo and Schneider to a wider age range, by providing confidence intervals for PGU, and by further evaluating the psychometric properties of PGU.

In the present study, we asked three research questions. First, does PGU increase significantly in children with and without LI who are between 4 and 9 years of age on the narrative generation task of ENNI? Second, are there significant differences in PGU between children with and without LI who are between 4 and 9 years of age? Third, does PGU show appropriate psychometric properties (i.e., split-half reliability, concurrent criterion validity, and diagnostic accuracy) in children between 4 and 9 years of age? On the basis of prior studies (Eisenberg & Guo, 2013, 2015; Guo & Schneider, 2016; Souto et al., 2014; Westerveld & Gillon, 2010), we predicted that PGU would increase significantly in children with and without LI who were between 4 and 9 years of age. Children with TL would produce significantly higher PGU than children with LI at each age level. In addition, PGU should show appropriate psychometric properties in children between 4 and 9 years of age.

Method

Participants

The data for this study are from the normative sample in the ENNI (Schneider et al., 2005). Participants in the normative sample were 377 children (300 TL, 77 LI) between 4 and 9 years of age recruited from Edmonton, Canada. Ethics approval was obtained from the University of Alberta ethics board. There were 50 children with TL for each age level; the number of children with LI varied (see Table 1). The chronological ages were not significantly different between the TL and LI groups at any age level, Fs ≤ 2.41, ps $\geq .13$, ds ≤ 0.13. The distribution of gender was controlled in the TL group but not in the LI group so as to reflect the actual distribution in the LI group; the distribution was not significantly different between the TL and LI groups at any of the age levels, $\chi^2 \leq 3.03$, ps $\geq .08$. 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 it was only specified that English must be the first language in the inclusion criteria when participants were recruited for the ENNI database.

Children with TL in the normative sample were recruited from 13 day cares and preschools and 34 elementary schools in the Edmonton area, all of which were randomly selected (Schneider et al., 2005). Teachers in the participating schools who had students in the target age range 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, with one boy and one girl at each level. This was to ensure that the normative sample would consist of children with TL who had varying language skills. All children in the TL group were typically developing (TD) per teachers' reports and did not have any known speech or language difficulties or any other disorders such as learning disability, autism spectrum disorders, or attention-deficit/hyperactivity disorder.

As part of the study protocol, two subtests of the CELF-P (Wiig et al., 1992) or the CELF-3 (Semel et al., 1995) were administered to children referred as having TL, depending on the child's age. Children younger than 6;0 were tested using the Linguistic Concepts and Recalling Sentences in Context subtests from the CELF-P. Children aged 6;0 and older were tested using the Concepts and Directions and Recalling Sentences subtests from the CELF-3. The purpose of administering these subtests was to provide a description of language skills for children with TL. Children's performance on these subtests, however, did not affect inclusion or exclusion from analysis (Schneider & Hayward, 2010).

Table 1. Mean (standard deviation) of demographic measures of children by language status and age.

Language status and age	*N*	Gender	Age in months	SES	Concepts and Directions[a]	Recalling Sentences (in Contexts)[b]	CELF-P/ CELF-3[c]
Typical language							
4-year-olds	50	25 G, 25 B	55.20 (2.88)	47.38 (13.58)	10.82 (3.32)	9.96 (2.38)	—
5-year-olds	50	25 G, 25 B	66.12 (3.24)	46.64 (12.12)	10.74 (2.63)	9.96 (2.79)	—
6-year-olds	50	25 G, 25 B	78.94 (3.99)	48.31 (14.75)	11.58 (3.03)	11.76 (3.32)	—
7-year-olds	50	25 G, 25 B	90.48 (3.36)	45.13 (13.65)	12.24 (3.26)	11.66 (2.79)	—
8-year-olds	50	25 G, 25 B	102.92 (3.34)	45.04 (11.55)	12.16 (2.92)	10.84 (2.74)	—
9-year-olds	50	25 G, 25 B	113.88 (3.36)	48.79 (12.04)	11.84 (2.80)	11.14 (2.60)	—
Language impairment							
4-year-olds	12	3 G, 9 B	55.92 (2.76)	47.17 (10.80)	4.33 (2.6)	5.42 (1.17)	76.83 (8.62)
5-year-olds	14	6 G, 8 B	64.92 (3.12)	46.52 (12.00)	5.00 (2.88)	4.43 (1.28)	76.21 (11.35)
6-year-olds	11	5 G, 6 B	79.55 (3.17)	40.26 (13.97)	5.73 (1.79)	5.27 (2.20)	78.55 (8.18)
7-year-olds	13	3 G, 10 B	90.72 (2.76)	42.42 (13.30)	6.38 (2.36)	4.31 (1.49)	74.00 (11.05)
8-year-olds	17	7 G, 10 B	104.35 (3.12)	42.42 (7.40)	7.47 (2.37)	5.00 (1.80)	76.29 (11.94)
9-year-olds	10	5 G, 5 B	114.00 (2.52)	48.71 (9.66)	8.10 (2.55)	5.40 (1.96)	73.50 (11.27)

Note. TL = children with typical language; LI = children with language impairment; G = girl; B = boy; SES = socioeconomic status as measured by the Blishen scale (Blishen et al., 1987); CELF-P = Clinical Evaluation of Language Fundamentals–Preschool (Wiig et al., 1992); CELF-3 = Clinical Evaluation of Language Fundamentals–Third Edition (Semel et al., 1995); em dashes = data not obtained.

[a]The standard score (M = 10, SD = 3) was reported for the Concepts and Directions subtest in the CELF-P or CELF-3. [b]For 4- and 5-year-olds, the standard score (M = 10, SD = 3) was reported for the Recalling Sentences in Context subtest in the CELF-P. For 6-year-olds or older, the standard score (M = 10, SD = 3) was reported for the Recalling Sentences subtest in the CELF-3. [c]The standard scores of Total Language Composites (M = 100, SD = 15) were reported for the LI group using the manuals of the CELF-P or CELF-3. No Total Language Composites were reported for children with TL because they were administered only the Concepts and Directions subtest and the Recalling Sentences (in Contexts) subtest.

Across all ages, 19 children with TL (one to six per age group) had standard scores below 7 (i.e., −1 *SD*) on the Linguistic Concepts (CELF-P) or Concepts and Directions (CELF-3) subtest, 11 children (one to three per age group) had scores below 7 on the Recalling Sentences in Context (CELF-P) or Recalling Sentences (CELF-3) subtest, and six children (zero to two per age group) had scores lower than 7 on both subtests. These children were still included in the TL group for several reasons. First, the teachers did not have any concerns for these children on their speech/language skills (Bishop, Snowling, Thompson, Greenhalgh, & the CATALISE Consortium, 2016; Paul, 2007). Second, it is possible for a child with typical language skills to score below −1 *SD* on individual subtests (Semel et al., 1995; Wiig et al., 1992). Third, only the two subtest scores were available for most of the children in the TD group (see Schneider, Hayward, & Dubé, 2006, for an explanation), which would not be adequate for identifying LI without supporting information. This was because the CELF-P/CELF-3 manuals emphasized that, while the Composite Standard Scores (i.e., Receptive Composite, Expressive Composite, and Total Language Composite) could be used for diagnosing the presence/absence of LI, the standard scores for individual subtests cannot be used for this purpose. Finally, eliminating the children from the TD group who had the lowest CELF scores would potentially bias the sample in the direction of greater differences between the groups on the ENNI. Table 1 presents the standard scores (M = 10, SD = 3) of the two subtests in the CELF-P or CELF-3 for children with TL in the normative sample of the ENNI by age level.

The subsample of children with LI was recruited from three sites: a public school serving children with communication disorders, a rehabilitation hospital that had several programs for children with LI, and Capital Health Authority, which served preschool and school-age children throughout the city of Edmonton. Each site was asked to refer children with a rating of 2–5 on a severity rating scale designed by Capital Health that rates a child's LI from 1 (mild) to 5 (severe). Children could be referred even if they had a concomitant diagnosis of learning disability, mild-to-moderate speech sound disorder, fine or gross motor delay, or attention-deficit disorder with or without hyperactivity (ADD/ADHD) with medication. Participating sites were also asked not to refer children who had a diagnosis of autism, intellectual disability, hearing impairment, severe speech sound disorder, ADD/ADHD without medication, or severe visual impairment that would result in inability to see pictures even with correction. However, the information regarding whether children with LI in the normative sample also had concomitant speech disorders, motor delay, or ADD/ADHD (with medication) was not available at the time of data collection because access to children's clinical records was not obtained. Information regarding nonverbal IQ was not collected; neither professionals who referred children nor the examiners who tested the participants for the study had concerns about cognitive abilities for any of the children with LI or with TL.

To further confirm the language status of children with LI, the full CELF-P or CELF-3 was administered, depending on the child's age (i.e., younger than 6;0 or not). All of the children in the LI group 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-P or CELF-3. The cutoff standard score of 85 was based on the recommendation of the

CELF manuals and was consistent with the cutoff used in previous studies of children with LI (e.g., Munson, Kurtz, & Windsor, 2005). Across ages, the percentage of children with LI who scored below the cutoff was 66% (51/77) for the Receptive Composite, 95% (73/77) for the Expressive Composite, and 84% (65/77) 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 at least one of the composite scores on the CELF-P or CELF-3. Table 1 presents the Total Language Composite ($M = 100$, $SD = 15$) for children with LI in the normative sample by age.

Demographic information, including ethnicity and socioeconomic status (SES), was also collected. Ethnic composition of children in the normative sample 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. The SES of the children was estimated from parents' occupations using the Blishen scale (Blishen, Carroll, & Moore, 1987). Based on Canadian census information, this index reflects equally weighted components of education and income level by occupation. For instance, cashiers are assigned a score of 28.31 and architects are assigned a score of 68.12 on the scale. Table 1 also presents the mean SES score by language status and age. The SES scores were not significantly different between the TL and LI groups at any of the age levels ($ps \geq .12$, $ds \leq 0.42$) or between the age levels regardless of children's language status ($p = .39$, $d = 0.25$).

Materials

The ENNI (Schneider et al., 2005) used a story generation task to elicit narratives from children. The task involved children telling stories about six original picture sequences with animal characters. The picture sequences were all black-and-white line drawings produced 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), number and gender of characters (i.e., two, three, and four characters), and amount of story information. The two sets of the ENNI (i.e., Set A and Set B) each consist of three stories with one picture sequence from each complexity level. These picture sets may be viewed and downloaded from the ENNI website.

Procedure

Each participant was seen individually by a trained examiner in the child's preschool, day care, or school. Three female examiners with a bachelor's degree in education or psychology were employed to evaluate children and administer the story generation tasks. The task began by instructing the child that he or she could see all the pages first and then tell a story to the examiner. The instructions also stressed 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 a separate 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. This meant that the child needed to be explicit in order for the examiner to understand. For example, the child could not legitimately use a pointing gesture to replace language when referring to a specific character or object in the picture. When a given story was administered, the child first previewed all pictures pertaining to the story and then started to tell the story. The examiner turned the pages 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 difficulties with the task. The training story was not included in the analysis. After the training story, the two story sets were given. All children completed story generation for both picture sets (i.e., six sequences in total). Administration of the story sets was counterbalanced across children. 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 based on the conventions of Systematic Analysis of Language Transcripts (Miller & Chapman, 2000). Responses for the six picture sequences were combined together into one transcript for each child and were further segmented into C-units. A C-unit is an independent clause plus all of its dependent clauses (Loban, 1976). Nonclausal utterances that expressed complete thoughts (e.g., *The girl and her mommy*) 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 PGU, a C-unit also had to have at least a verb (e.g., *There was an elephant; He got the ball; Wanted to go for a swim*), except in the case of C-units with omitted copula *BE* (e.g., *The elephant mad*; Eisenberg et al., 2012).

To compute PGU and to determine the patterns of grammatical errors that children made, we first identified the errors that children made in the narratives using the coding scheme of Eisenberg et al. (2012).

1. Tense marking errors were operationally defined as omissions or incorrect usage of tense markers, including third-person singular present *–s,* regular past tense *–ed,* copula *BE,* auxiliary *BE,* auxiliary *DO,* auxiliary *have,* irregular past tense (e.g., *She flew the airplane*),

and irregular third-person verb forms (e.g., *He has a ball*). Verbs produced without an inflection, modal, or auxiliary (e.g., *The rabbit talk to the doctor*) 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 with a ball by the pool*) unless the context clearly required a tense marker (e.g., *They fall in the swimming pool just now*).

2. Pronoun errors were operationally defined as substitution errors for subject pronouns (e.g., *Him built a castle*), object pronouns, reflexive pronouns, possessive pronouns, and possessive determiners. 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 character as *a boy* and uses the pronoun *she* later to refer to the same character, it would be coded 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*), 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 walk by the pool*).
4. Argument structure errors were operationally defined as omissions of required constituents (i.e., argument) before or after verbs (e.g., *Got stuck in the sand; He put the sand*). Decisions about the required arguments for verbs were made based on 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., *She didn't know what was that*) or semantic irregularities (e.g., *She asked the doctor we want a balloon*) 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 (Saeed, 2009). Second, this decision is consistent with other assessments, such as Developmental Sentence Scoring (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.

When a C-unit contained one or more error codes, it was marked as ungrammatical. PGU was computed for each child by subtracting the total number of ungrammatical C-units from the total number of C-units that were included for analysis and then dividing by the total number of C-units that were included for analysis. The resultant quotient was multiplied by 100% to obtain a percentage. It should be noted that, although 28% of children were of non-European origin, ethnicity was not associated with any known linguistic variations that would impact the scoring of grammaticality for those who spoke English as the first language in Canada. Thus, we did not adjust the scoring of grammaticality for children's ethnicity or dialect. Supplemental Material S1 presents an example for computing PGU; Supplemental Material S2 presents the rate of each error type by language status and age.

Reliability of Transcription and Coding

To check 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 third author of the present study (i.e., 100% check) before transcription reliability was assessed. Another research assistant majoring in speech-language pathology, who had not transcribed the stories, independently transcribed one story from 24% of the participants for transcription reliability purposes. The C-unit segmentation consistency was 96%, and the word-by-word consistency was 97%.

To check the coding reliability for PGU, we adapted a consensus procedure from Shriberg, Kwiatkowski, and Hoffman (1984). Four other graduate assistants majoring in speech-language pathology, who did not transcribe the narratives, first coded the errors for PGU for all children. The first author then checked the coded transcripts for all children. Discrepancies were discussed between the first and third/fourth authors. Across the age groups, 391 (1.41%) out of 27,715 C-units were discussed. All of the discrepancies were resolved.

Statistical Analysis

We used a two-way analysis of variance (ANOVA) to evaluate the effects of language status and age on PGU. We used the *d* value to quantify the effect size or magnitude of the differences in PGU. Following Cohen (1988), we interpreted the effect size as small ($.2 \leq d < .5$), medium ($.5 \leq d < .8$), or large ($d \geq .8$) whenever appropriate. Because PGU was calculated as a percentage, it was arcsine-transformed in the ANOVAs.

To compute the 90% and 95% confidence intervals for PGU for children with TL at each age level, we followed the steps specified in McCauley and Swisher (1984). We first obtained the split-half reliability of PGU in children with TL by computing the correlation of PGU scores in Story Set A and Story Set B for each age level, which was consistent with the current clinical practice (Dawson et al., 2005; Wiig, Secord, & Semel, 2004). The correlation coefficients (*r*) were then used to calculate the standard error of measurement in the equation (i.e., $SEM = SD^*\sqrt{1-r}$, where *SD* is the standard deviation of PGU for a given age level). A 90% confidence interval is computed by

multiplying the standard error of measurement value by −1.645 (lower limit) or +1.645 (upper limit). A 95% confidence interval is computed in a similar way, except that the standard error of measurement is multiplied by −1.96 (lower limit) or +1.96 (upper limit).

To evaluate the split-half reliability of PGU, we computed the correlation of PGU scores in Story Set A and Story Set B from all of the children for each age level. That is, children with and without LI were combined into one group for this analysis so that we had sufficient variability in PGU scores for each age range (Kleinbaum, Kupper, Muller, & Nizam, 1998). Note that this is different from how split-half reliability was determined for calculating confidence intervals. Children with LI were not included in the computation of split-half reliability for confidence intervals because reference data were computed based only on children with TL. To determine the concurrent criterion validity of PGU, we computed the correlation of overall PGU scores and the raw scores of the Recalling Sentences in Context subtest of the CELF-P (4- and 5-year-olds) or the Recalling Sentences subtest of the CELF-3 (6- to 9-year-olds). Note that, although Table 1 presents the standard scores of the subtests for ease of interpretation, raw scores, rather than standard scores, of the subtests were used for the correlation analysis. We chose these subtests as the criterion measures because sentence recall/repetition has been considered a task that taps expressive grammar (Pawłowska, 2014; Semel et al., 1995; Wiig et al., 1992). Following Cohen (1988), we interpreted the correlation coefficients as small ($.1 \leq r < .3$), medium ($.3 \leq r < .5$), or large ($r \geq .5$) whenever appropriate.

To evaluate the diagnostic accuracy of PGU, we computed the sensitivity, specificity, and likelihood ratios (Dollaghan, 2007). Likelihood ratios, in general, are less affected by sample size than sensitivity and specificity. 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 as LI by PGU. 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 as TL by PGU. Based on Plante and Vance (1994), sensitivity and specificity levels between 80% and 89% were considered acceptable, and sensitivity and specificity levels at or greater than 90% were 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, 1991) in the software SigmaPlot 12.0 (Systat Software, 2011). The ROC curve analysis automatically calculates pairs of sensitivity and specificity levels for a range of cutoff scores. Following Sackett (1991), we chose the score that maximized the diagnostic accuracy, where sensitivity plus specificity divided by 2 is largest.

Results

Descriptive Analyses of Narrative Samples

Table 2 presents the descriptive measures from the narratives and the number of C-units that were included for computing the score of PGU. Regardless of age, children with TL did not differ significantly from those with LI in the total number of C-units that were used to talk about the stories, $F(1, 375) = 0.08$, $p = .79$, $d = 0.06$. However, children with TL produced significantly longer mean length of C-units and more different words than those with LI when talking about the stories (Fs ≥ 30.78, ps < .001, ds ≥ 0.57). For the number of C-units that were included for computing PGU scores, there was no significant difference between children with and without LI, $F(1, 375) = 1.62$, $p = .20$, $d = 0.13$. Across ages, children with TL produced at least nine C-units and those with LI produced at least 11 C-units for computing PGU scores (see data for 4-year-olds in Table 2). Note that the number of C-units that were included for computing PGU scores was lower than the total number of C-units in the narratives (see Table 2) because some C-units in the narratives were not included for the PGU analysis due to the lack of verbs. Supplemental Material S3 further provides statistical comparisons (i.e., F values, p values, and effect sizes) regarding differences between the TL and LI groups in total number of C-units, mean length of C-units in morphemes, number of different words, and number of C-units that were included for the PGU analysis by age level.

Age-Related Changes for PGU in Children With and Without LI

Table 3 presents children's PGU scores by language status and age. A 2 (language status: TL vs. LI) × 6 (age level: 4–9 years) ANOVA revealed that there was a main effect of language status on PGU scores, $F(1, 365) = 281.86$, $p < .001$, $d = 1.76$. There was also a main effect of age level on PGU scores, $F(5, 365) = 35.91$, $p < .001$, $d = 1.40$. The main effects were further qualified by a significant interaction effect, $F(5, 365) = 3.31$, $p < .006$, $d = 0.42$. To answer our research questions, we performed the follow-up analyses in two ways.

Table 2. Mean (standard deviation) and range of descriptive measures by language status and age.

Language status and age	Total no. C-units	$MLCU_m$	NDW	No. CU for PGU
Typical language				
4-year-olds	68.06 (19.49) 39–151	6.97 (1.08) 4.44–8.96	130.32 (35.87) 37–252	66.04 (20.59) 9–148
5-year-olds	72.04 (21.23) 48–136	7.73 (0.97) 5.51–10.22	141.40 (31.17) 77–219	71.16 (20.44) 48–134
6-year-olds	71.64 (19.25) 50–129	7.59 (0.97) 5.10–10.00	142.90 (29.74) 94–225	70.62 (18.56) 49–128
7-year-olds	73.88 (18.46) 45–136	8.49 (1.15) 5.96–10.96	154.84 (28.78) 90–228	73.64 (18.25) 45–133
8-year-olds	78.70 (21.71) 49–146	8.70 (1.07) 6.73–11.08	172.94 (42.07) 113–279	78.44 (21.40) 49–143
9-year-olds	77.78 (24.08) 51–160	9.00 (1.07) 6.80–11.18	172.06 (36.55) 112–315	77.56 (24.11) 51–160
Language impairment				
4-year-olds	58.00 (17.93) 33–106	5.18 (1.37) 3.12–6.89	84.25 (26.67) 43–140	51.75 (21.30) 11–100
5-year-olds	70.64 (22.29) 40–129	5.94 (1.35) 2.70–8.05	113.36 (30.94) 47–168	64.79 (23.88) 21–126
6-year-olds	73.63 (24.46) 52–129	7.04 (1.05) 5.50–8.50	123.45 (26.55) 89–179	71.18 (24.16) 50–122
7-year-olds	80.31 (44.07) 45–181	6.90 (1.30) 4.74–9.12	135.62 (51.35) 84–267	76.77 (40.02) 45–181
8-year-olds	73.94 (21.41) 46–124	7.20 (0.81) 5.83–8.62	141.47 (30.95) 99–202	72.06 (21.70) 46–122
9-year-olds	81.40 (38.48) 51–174	7.91 (0.91) 6.67–9.32	164.20 (45.26) 120–262	79.60 (35.25) 50–163

Note. Total no. C-units = total number of communication units (C-units) in the narratives; $MLCU_m$ = mean length of C-units in morphemes; NDW = number of different words; No. CU for PGU = number of C-units that were included for percent grammatical utterances analysis.

First, we evaluated the age differences in PGU scores within the TL and LI groups separately. One-way ANOVAs showed that there was a main effect of age levels for both the TL group, $F(5, 294) = 26.35$, $p < .001$, $d = 1.34$, and the LI group, $F(5, 71) = 12.10$, $p < .001$, $d = 1.85$. Post hoc Tukey tests indicated that, for children with TL, 6- to 9-year-olds produced higher PGU scores than 4- and 5-year-olds. Eight-year-olds also produced higher PGU scores than 6-year-olds. There were no other significant age differences within the TL group. At the group level, children with TL reached the customary level of mastery (i.e., 90% accurate) for PGU around the age of 6 years. For children with LI, 6- to 9-year-olds produced higher PGU scores than 4-year-olds. Eight- and 9-year-olds also produced higher PGU scores than 5-year-olds. There were no other significant age differences in PGU scores within the LI group. At the group level, children with LI did not reach the customary level of mastery for PGU even at the age of 9 years.

Table 3. Descriptive statistics of percent grammatical utterances (PGU) scores (in percentage) by language status and age.

Language status and age	*M*	*SD*	Range	Effect size (*d*)[a]	*SEM*	90% CI	95% CI
Typical language							
4-year-olds	78.82	14.87	17.65–98.21	1.98	7.16	±11.74	±14.04
5-year-olds	83.81	15.03	3.85–98.31	1.66	6.20	±10.17	±12.15
6-year-olds	90.64	6.60	65.00–100	1.88	5.30	±8.70	±10.40
7-year-olds	91.86	5.89	73.33–100	1.89	3.95	±6.48	±7.74
8-year-olds	94.94	3.48	84.00–100	2.02	3.33	±5.46	±6.52
9-year-olds	93.80	4.25	82.46–100	1.29	3.28	±5.38	±6.43
Language impairment							
4-year-olds	33.32	23.34	7.14–79.66	—	—	—	—
5-year-olds	53.34	20.32	0–80.65	—	—	—	—
6-year-olds	63.82	19.21	33.00–88.00	—	—	—	—
7-year-olds	68.26	20.60	12.96–90.74	—	—	—	—
8-year-olds	77.76	15.32	32.00–94.00	—	—	—	—
9-year-olds	83.61	7.69	66.07–95.08	—	—	—	—

Note. SEM = standard error of measurement; CI = confidence interval; em dashes = data not obtained.

[a]Effect size for the difference in PGU between children with and without language impairment.

Second, we examined the differences in PGU scores between children with and without LI by age level. One-way ANOVAs indicated that children with TL produced higher PGU scores than those with LI at each age level (Fs ≥ 24.15, ps < .001; significant with Bonferroni correction) and the effect sizes were all large (ds ≥ 1.29; see Table 3).

Table 3 also presents the 90% and 95% confidence intervals for PGU scores in children with TL by age. In general, the 90% and 95% confidence intervals decreased with age partly because of the decreasing individual variability in PGU scores (i.e., standard deviations).

Psychometric Properties of PGU

Table 4 presents the correlation coefficients for the split-half reliability and concurrent criterion validity by age. The correlation coefficient for the split-half reliability of PGU was large for each age level (rs ≥ .631, ps < .001), meaning that PGU scores computed from Set A were consistent with those computed from Set B between the ages of 4 and 9 years. Similarly, the correlation coefficient for the concurrent criterion validity of PGU was large for each age level (rs ≥ .557, ps < .001), meaning that children's performance on PGU was consistent with their performance on the Recalling Sentences in Context subtest of the CELF-P (ages 4 and 5 years) or the Recalling Sentences subtest of the CELF-3 (ages 6 to 9 years).

Table 5 presents the indices of diagnostic accuracy for PGU by age. It should be noted that, given the small sample size of children with LI, the results regarding the diagnostic accuracy of PGU here were considered preliminary and should be interpreted with caution. Sensitivity and specificity were both acceptable to good (range: 82%–100%) for PGU between the ages of 4 and 9 years. For example, PGU demonstrated a sensitivity level of 92% at the age of 7 years, meaning that 92% (12/13) of 7-year-old children with LI in the present study were correctly identified as LI by PGU. PGU also demonstrated a specificity level of 88%, meaning that 88% (44/50) of 7-year-old children with TL in the present study were correctly identified as TL by PGU.

The likelihood ratios showed a similar trend for the diagnostic accuracy of PGU: The positive and negative likelihood ratios (LR+ and LR−) were acceptable to good for PGU between the ages of 4 and 9 years. For example, the LR+ value was 7.69 at the age of 7 years, meaning that 7-year-old children obtaining a fail score (i.e., below the cutoff) for PGU in the present study were 7.69 times as likely to have LI. The LR− value was 0.09 at the age of 7 years, meaning that 7-year-old children obtaining a pass score (i.e., at or above the cutoff) for PGU in the present study were only 0.09 times as likely to have LI.

Table 4. Split-half reliability and concurrent criterion validity in correlation coefficients for percent grammatical utterances by age.

Age	Split-half reliability	Concurrent criterion validity
4-year-olds	0.897	0.768
5-year-olds	0.860	0.633
6-year-olds	0.794	0.557
7-year-olds	0.854	0.659
8-year-olds	0.818	0.637
9-year-olds	0.631	0.596

Note. All correlation coefficients are significant at the .001 level (one-tailed).

Discussion

The present study provided reference data and evaluated the psychometric properties for PGU in children between 4 and 9 years of age from the database of the ENNI. PGU showed age-related changes and demonstrated appropriate psychometric properties in children between 4 and 9 years of age. We explore these findings below.

PGU Increased Between 4 and 9 Years of Age in Children With and Without LI

In the present study, we found that PGU scores increased significantly between the ages of 4 and 9 years in children with TL. On average, children with TL reached the customary level of mastery for PGU (i.e., 90% accurate) around the age of 6 years, and their performance on PGU became stable after the age of 7 years. The current findings were comparable to those in Westerveld and Gillon (2010). In their study, children's grammatical accuracy scores on a story retelling task were 86.2% at the age of 5 years, 87.9% at the age of 6 years, and 91.8% at the age of 7 years, all of which were close to the PGU scores in children at the corresponding age level in the present study. Our results were also consistent with those in Fey et al. (2004), which found that grammatical accuracy scores did not differ significantly between second and fourth graders. Together, prior studies (Lee, 1974) and current results indicate that the ability to produce grammatical sentences starts early and develops over time in a gradual manner.

The present findings could be interpreted from the framework of Bock and Levelt's (1994) language production model. When producing grammatical sentences, speakers need to activate appropriate lexical items and integrate them into syntactic frames. Thus, during the process of language acquisition, children must learn an extensive body of morphological and syntactic structures, such as third-person singular *–s,* infinitive *–to,* prepositional phrase structure, and word order, in order to generate grammatical sentences (Lee, 1974; Schuele, 2013; Tomasello, 2003). Acquisition of grammatical structures does not occur in an all-or-none fashion. Instead, it could vary in degree. Some children may have relatively stronger representation of grammatical structures (e.g., tense markers, passive sentences) than other children due to individual differences in language learning capacity (e.g., pattern-finding skills; Tomasello, 2003). Within the same child, some grammatical structures could have relatively stronger representations than others due to the nature of those particular grammatical structures (e.g., complexity, input frequency; Ibbotson, 2013).

Table 5. Indices of diagnostic accuracy for percent grammatical utterances (PGU) by age using the cutoff score from the receiver operating characteristic curve analysis.

Age	Cutoff	Sensitivity[a]	Specificity	Overall accuracy	LR+[b]	LR–
4-year-olds	54.04%	83%* (10/12)	96%** (48/50)	94% (58/62)	20.83**	0.17*
5-year-olds	79.10%	100%** (14/14)	82%* (41/50)	86% (55/64)	5.56*	< 0.01**
6-year-olds	83.00%	82%* (9/11)	90%** (45/50)	89% (54/61)	8.18*	0.20*
7-year-olds	85.40%	92%** (12/13)	88%* (44/50)	89% (56/63)	7.69*	0.09**
8-year-olds	91.50%	88%* (15/17)	84%* (42/50)	85% (57/67)	5.52*	0.14*
9-year-olds	88.42%	90%** (9/10)	90%** (45/50)	90% (54/60)	9.00*	0.11*

[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 who are correctly classified; for example, nine out of eleven 6-year-olds with language impairment were correctly classified by the PGU. [b]LR+ = positive likelihood ratio; LR– = negative likelihood ratio. 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 at or above 10.00 or an LR– value at or below 0.10.

Grammatical structures with stronger representations would be activated relatively more easily and less prone to errors than those with weaker representations (Charest & Johnston, 2011; MacDonald & Christiansen, 2002). When children start to activate grammatical structures for producing sentences at a younger age, they may be more likely to make errors because their representations of grammatical structures would be weaker, leading to lower grammatical accuracy in language production. The strength of representations for grammatical structures, however, may increase over developmental time due to increased exposures. Grammatical structures could then be activated relatively more easily with fewer errors, leading to higher grammatical accuracy.

Although children's performance on PGU became stable by the age of 7 years in the present study, this does not necessarily mean that grammatical development stops at the age of 7 years. After all, representations for existing grammatical structures would become even stronger, and new grammatical structures (e.g., structures with subjunctive mood) would be learned by older children. In fact, a longitudinal study by Tomblin and Nippold (2014) has shown that grammatical skills (e.g., syntactic complexity) continue improving into the adolescent years as revealed in standardized tests and spoken discourse. Thus, our findings only suggest that children with TL are able to produce grammatical sentences at the customary level of mastery at least in early elementary school years during a narrative generation task. Beyond the early elementary years, further improvement in grammatical skills might not be reflected in overall grammatical accuracy due to a ceiling effect.

As in children with TL, PGU scores also increased significantly between the ages of 4 and 9 years in children with LI. However, children with LI, as a group, did not reach the customary level of mastery for PGU even at the age of 9 years in the narrative generation task. In addition, PGU scores were significantly lower in the LI group than in the TL group between the ages of 4 and 9 years, which was consistent with previous studies (Fey et al., 2004; Scott & Windsor, 2000; Souto et al., 2014). Inspecting the errors (Supplemental Material S2), we found that children with LI produced a higher rate for each error type than those with TL across the age levels, which was also consistent with prior studies (Ebbels et al., 2007; Moore, 2001; Rice et al., 1998; Schuele & Tolbert, 2001). At present, we do not have evidence showing whether or when children with LI, as a group, would reach the customary level of mastery on PGU. Nor do we have evidence showing whether children with LI would catch up with children with TL on the performance of PGU. Future studies that examine the performance of PGU beyond the age of 9 years are needed to answer these questions.

Why do children with LI tend to have difficulty learning and using grammatical structures? Extant theories have attributed grammatical difficulties in children with LI to deficits in innate grammatical representations, general processing capacity, or specific learning mechanisms (see Leonard, 2014, for a comprehensive review). Evaluating the theoretical accounts is beyond the scope of the present study. However, we would like to point out that any viable theory must be able to explain why children with LI have difficulty learning and using grammatical structures at different levels of complexity in general and are relatively more vulnerable on certain aspects of grammar (e.g., tense markers; Leonard et al., 2017; Rice et al., 1998) in particular. Such theories would not only improve our understanding of the nature of grammatical deficits in children with LI but also inform how we should approach grammatical intervention.

PGU Showed Appropriate Psychometric Properties Between the Ages of 4 and 9 Years

In the present study, we evaluated three psychometric properties for PGU computed from the ENNI protocol. First, PGU demonstrated appropriate split-half reliability, which was evidenced in the result that PGU scores from Story Set A were significantly correlated with those from Story Set B between the ages of 4 and 9 years with large correlation coefficients. This finding was consistent with

our prior finding for PGU in 3-year-olds participating in a picture description task (Eisenberg & Guo, 2015). Second, PGU showed appropriate concurrent criterion validity, which was evidenced in the finding that PGU scores were significantly correlated with raw scores for the sentence recall subtest of the CELF-P/CELF-3, a task tapping expressive grammar, at each age level with large correlation coefficients. This suggests that PGU is a valid measure that evaluates children's expressive grammatical skills between the ages of 4 and 9 years. Third, PGU demonstrated acceptable-to-good diagnostic accuracy between the ages of 4 and 9 years. The results were compatible with those in 3-year-olds (Eisenberg & Guo, 2013) and in 4- and 5-year-olds (Souto et al., 2014). The present study also extended our previous study (Guo & Schneider, 2016) by showing that PGU demonstrated acceptable-to-good diagnostic accuracy at the ages of 7 and 9 years. Finally, although the present study was a cross-sectional study and the age-related changes could not be directly used to infer the development of PGU over time, the findings that PGU scores increased with age within the TL and LI groups could be considered as an additional piece of evidence for the validity of PGU (i.e., construct validity; Aiken & Groth-Marnat, 2006). Together, the results from the age-related changes and the psychometric properties collectively suggest that PGU computed from the ENNI protocol is reliable and valid and could be used for differentiating children with and without LI between 4 and 9 years of age.

It should be noted that the correlation coefficients for split-half reliability and concurrent criterion validity, although large in their magnitude, fluctuated between ages. One possibility for this finding is that we had a small sample size per age level and the fluctuation of correlation coefficients for these psychometric properties may have resulted from sampling errors. However, we would also like to point out that it may not be uncommon to observe variability in split-half reliability or concurrent criterion validity between ages for a given test or measure. For example, the split-half reliability for the Structured Photographic Expressive Language Test–Preschool 2 (Dawson et al., 2005) ranged from 0.80 to 0.89 between the ages of 3 and 5 years. Similarly, the split-half reliability for the Sentence Structure subtest in the Clinical Evaluation of Language Fundamentals Preschool–Second Edition (Wiig et al., 2004) ranged from 0.69 to 0.84 between the ages of 3 and 6 years. Interestingly, neither of these tests reported concurrent criterion validity by age level. For example, the Structured Photographic Expressive Language Test–Preschool 2 reported a correlation coefficient of .86 with the criterion measure, with children between 3 and 5 years of age combined in the computation. Thus, the present study was ahead of many other empirical studies in reporting concurrent criterion validity by age level despite the variability between ages.

Limitations and Future Directions

By reanalyzing the archival data from the ENNI (Schneider et al., 2005), we faced some limitations that must be considered. First, the present study was a two-gate, instead of one-gate, design because we preselected children with and without LI for investigating the diagnostic accuracy of PGU. One potential drawback of a two-gate design is that the diagnostic accuracy of PGU may have been inflated (Dollaghan & Horner, 2011). A related issue is that we had small sample sizes for children with LI at each age level. Although there is no requirement for the minimum number of cases per group in order to conduct the ROC curve analysis, it has been recommended that each group have at least 50 cases (Metz, 1978) so that one case accounts for no more than 2% of diagnostic accuracy (e.g., sensitivity, specificity). While the number of children with TL in the present study met this recommended guideline, the number of children with LI did not. One problem for having a small number of children with LI is that the differences in the sensitivity levels between ages could be overinterpreted. For example, misclassifying one LI would lead to a 10% decrease in sensitivity at the age of 9 years but only about a 6% decrease in sensitivity at the age of 8 years. Thus, given the two-gate design and the small sample sizes for children with LI, the findings regarding the diagnostic accuracy of PGU here were considered preliminary and should be interpreted with caution. Future studies that include more children with LI at each age level for investigating the diagnostic accuracy of PGU are needed.

In addition, information about the proportion of children who were exposed to a language other than English was not documented in the ENNI normative sample. Information regarding children's nonverbal intelligence was also not available because it was not collected for the ENNI normative sample. However, neither the teachers/clinicians who referred children to the ENNI normative sample nor the examiners had any concerns on children's cognitive development at the time of data collection. Similarly, we did not have the information about whether children with LI in the normative sample also had concomitant speech disorders, motor delay, or ADD/ADHD. One concern is that the current findings may not be generalizable for those who clearly did not have any of the concomitant disorders. However, in a recent discussion (Bishop, Snowling, Thompson, Greenhalgh, & the CATALISE-2 Consortium, 2017), there was an emerging agreement that concomitant disorders, such as speech sound disorders and motor delay, did not preclude the diagnosis of LI in children. In addition, using the data from tense marking, nonword repetition, and sentence recall, Redmond (2016) argued that the presence of ADD/ADHD did not exacerbate children's LI. Thus, the current findings may still be applicable to children with LI who have similar profiles to those in the present study (e.g., may or may not have speech sound disorders) when the same language sampling (i.e., story generation task from the ENNI) and analysis protocol is used.

Third, we did not compare PGU across different tasks for the same group of children. Thus, we were not able to determine whether similar findings on the reference data would be obtained when a different language sampling

protocol (e.g., picture description task, story retell task) was used. A direct comparison of PGU using different language sampling protocols would be a worthwhile pursuit. Given this limitation, it is important that clinicians follow the ENNI protocol in order to use the reference data from the present study.

Clinical Implications and Applications

In evaluating children's language skills, clinicians must use assessment tools that are technically sound (Individuals with Disabilities Education Act, 2004). Given the attested psychometric properties of PGU, we recommend that PGU computed from the ENNI be used as one tool for children between 4 and 9 years of age for determining eligibility for speech-language services. For example, some states mandate the use of *z* scores (i.e., number of standard deviations from the mean) to determine whether a child is eligible for services. In this case, the reference data (mean, standard deviation) from children with TL in the present study could be used to assist in clinical decision making. To illustrate, consider a 7-year-old child from a state using −1.25 *SD* as the cutoff to determine eligibility for speech-language services. Suppose that the child is administered the ENNI protocol and produces a PGU score of 68.15%. Using the reference data, the clinician can compute the child's *z* score, which is −4.03 ([68.15%–91.86%]/5.89%) and is well below the cutoff of −1.25. The result could then be considered as one piece of evidence suggesting that the child has an LI and is eligible for therapy. To facilitate the computation process, Supplemental Material S4 provides a calculation template for clinicians to convert PGU raw scores into *z* scores.

However, clinicians need to be aware that, although using a prescriptive standard (e.g., −1 *SD,* −1.25 *SD,* or −1.5 *SD*) to determine the eligibility for services is the current guideline for many workplaces, Spaulding, Plante, and Farinella (2006) have pointed out that it may not be the best strategy. This is because those prescriptive standards tend to be arbitrary without empirical evidence and may increase the likelihood of misdiagnosis. To address this issue, we further present the diagnostic accuracy for PGU using −1 *SD,* −1.25 *SD,* −1.5 *SD,* or −2 *SD* as the cutoff scores in Supplemental Material S5. It is evident that −1 *SD* is the most appropriate prescriptive standard for PGU because the associated indices for diagnostic accuracy are all at the acceptable level or higher.

If the production of grammatical structures is selected as the treatment goal, PGU could also be computed as a global measure for monitoring treatment progress. When PGU is used for this purpose, the upper limits of the 90% and 95% confidence intervals of PGU scores from the present study could be used to determine whether the child makes significant progress in producing grammatical structures (McCauley & Swisher, 1984). Consider the same 7-year-old who produces a PGU score of 68.15% in the ENNI protocol before treatment. Using the 95% confidence interval of PGU scores from the present study (i.e., ±7.74), the clinician could estimate that the child's "true" PGU score falls between 60.41% (lower limit) and 75.89% (upper limit; see Supplemental Material S4 for a calculation template for all ages). After treatment, the child produces a PGU score of 78.58%. Because the posttreatment PGU score is higher than the upper limit of the pretreatment PGU score, the child is considered to show significant progress in the performance of PGU (Gillam et al., 2008), although a PGU score of 78.58% is still 2.25 *SD*s below the mean, indicating that there remains a need for the clinician to work on grammatical structures with this child.

How do clinicians select grammatical structures for therapy goals? Although we did not focus the analysis on the grammatical errors, clinicians could use the coding system from the present study (i.e., error types) to conduct in-depth analyses of grammatical errors that children make in the narratives. If any patterns of grammatical errors emerge in the narratives (e.g., subject pronoun errors, omission of tense markers), these patterns could be readily chosen as therapy goals.

Concluding Thoughts

In an opinion paper, Schuele (2013) suggested that clinicians not confine grammatical assessment or intervention to Brown's 14 grammatical morphemes (Brown, 1973) or mean length of utterances. The present study provides clinicians with a framework to evaluate children's global grammatical skills (i.e., PGU) and to identify grammatical errors beyond Brown's 14 grammatical morphemes for intervention. PGU also has the potential to assist clinicians in determining the presence/absence of LI and evaluating the treatment progress for children between 4 and 9 years of age. Despite the usefulness of PGU, it is important for clinicians to know that grammatical accuracy is only one aspect of grammatical proficiency. Grammatical proficiency also encompasses other aspects, such as productivity and complexity (Schuele, 2013). Clinicians may need to focus on different aspects of grammatical proficiency for children at different levels to best facilitate their grammatical development.

Acknowledgments

The development of the Edmonton Narrative Norms Instrument was supported by a grant from the Children's Health Foundation of Northern Alberta. 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, Katelynn Imagna, Kayla Kuehlewind, and Sanjana Nair for coding the data.

References

Aiken, L. R., & Groth-Marnat, G. (2006). *Psychological testing and assessment* (12th ed.). Boston, MA: Allyn & Bacon.

Arndt, K. B., & Schuele, C. M. (2012). Production of infinitival complements by children with specific language impairment. *Clinical Linguistics & Phonetics, 26*(1), 1–17.

Bishop, D. V. M., Snowling, M. J., Thompson, P. A., Greenhalgh, T., & the CATALISE Consortium. (2016). CATALISE: A multinational and multidisciplinary Delphi consensus study. Identifying language impairments in children. *PLOS ONE, 11*(7), e0158753.

Bishop, D. V. M., Snowling, M. J., Thompson, P. A., Greenhalgh, T., & the CATALISE-2 Consortium. (2017). Phase 2 of CATALISE: A multinational and multidisciplinary Delphi consensus study of problems with language development: Terminology. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 58*(10), 1068–1080.

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.

Bock, K., & Levelt, W. (1994). Language production: Grammatical encoding. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 945–984). San Diego, CA: Academic Press.

Brown, R. (1973). *A first language: The early stages*. Cambridge, MA: Harvard University Press.

Charest, M., & Johnston, J. R. (2011). Processing load in children's language production: A clinically oriented review of research. *Canadian Journal of Speech-Language Pathology and Audiology, 35*(1), 18–31.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Dawson, J., Stout, C., Eyer, J., Tattersall, P., Fonkalsrud, J., & Croley, K. (2005). *Structured Photographic Expressive Language Test–Preschool 2 (SPELT-P2)*. DeKalb, IL: Janelle Publications.

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

Dollaghan, C. A., & Horner, E. A. (2011). Bilingual language assessment: A meta-analysis of diagnostic accuracy. *Journal of Speech, Language, and Hearing Research, 54*(4), 1077–1088.

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*(5), 1330–1349.

Eisenberg, S. L. (2003). Production of infinitival object complements in the conversational speech of 5-year-old children with language impairment. *First Language, 23*(3), 327–341.

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*(1), 20–31.

Eisenberg, S. L., & Guo, L.-Y. (2015). Sample size for measuring grammaticality in preschool children from picture-elicited language samples. *Language, Speech, and Hearing Services in Schools, 46*(2), 81–93.

Eisenberg, S. L., Guo, L.-Y., & Germezia, M. (2012). How grammatical are 3-year-olds? *Language, Speech, and Hearing Services in Schools, 43*(1), 36–52.

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.

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

Gillam, R. B., Loeb, D. F., Hoffman, L. M., Bohman, T., Champlin, C. A., Thibodeau, L., … Friel-Patti, S. (2008). The efficacy of Fast ForWord language intervention in school-age children with language impairment: A randomized controlled trial. *Journal of Speech, Language, and Hearing Research, 51*(1), 97–119. https://doi.org/10.1044/1092-4388(2008/007)

Grela, B. G., & Leonard, L. (1997). The use of subject arguments by children with specific language impairment. *Clinical Linguistics & Phonetics, 11*(6), 443–453.

Grela, B. G., Rashiti, L., & Soares, M. (2004). Dative prepositions in children with specific language impairment. *Applied Psycholinguistics, 25*(4), 467–480.

Guo, L.-Y., & Schneider, P. (2016). Differentiating school-aged children with and without language impairment using tense and grammaticality measures from a narrative task. *Journal of Speech, Language, and Hearing Research, 59*(2), 317–329.

Hadley, P. A. (1998). Language sampling protocols for eliciting text-level discourse. *Language, Speech, and Hearing Services in Schools, 29*(3), 132–147.

Hewitt, L. E., 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*(3), 197–213.

Ibbotson, P. (2013). The scope of usage-based theory. *Frontiers in Psychology, 4*(255), 1–15.

Individuals with Disabilities Education Act, 20 U.S.C. § 1400 (2004).

Johnston, J. R. (2006). *Thinking about child language: Research to practice*. Eau Claire, WI: Thinking Publications.

Kleinbaum, D. G., Kupper, L. L., Muller, K., & Nizam, A. (1998). *Applied regression analysis and other multivariable methods* (3rd ed.). Pacific Grove, CA: Brooks/Cole Publishing.

Lee, L. 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.

Leonard, L. B., Haebig, E., Deevy, P., & Brown, B. (2017). Tracking the growth of tense and agreement in children with specific language impairment: Differences between measures of accuracy, diversity, and productivity. *Journal of Speech, Language, and Hearing Research, 60*(12), 3590–3600. https://doi.org/10.1044/2017_jslhr-l-16-0427

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

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

MacDonald, M. C., & Christiansen, M. H. (2002). Reassessing working memory: Comment on Just and Carpenter (1992) and Waters and Caplan (1996). *Psychological Review, 109*(1), 35–54.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). New York, NY: Psychology Press.

McCauley, R. J., & Swisher, L. (1984). Use and misuse of norm-referenced tests in clinical assessment: A hypothetical case. *Journal of Speech and Hearing Disorders, 49*(4), 338–348.

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine, 8,* 283–298.

Miller, J., & Chapman, R. S. (2000). Systematic Analysis of Language Transcripts [Computer software]. Madison: University of Wisconsin.

Moore, M. E. (2001). Third person pronoun errors by children with and without language impairment. *Journal of Communication Disorders, 34*(3), 207–228.

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*(5), 1033–1047. https://doi.org/10.1044/1092-4388(2005/072)

Paul, R. (2007). *Language disorders from infancy through adolescence: Assessment & intervention* (3rd ed.). St. Louis, MO: Elsevier Mosby.

Pavelko, S. L., Owens, J. R. E., Ireland, M., & Hahs-Vaughn, D. L. (2016). Use of language sample analysis by school-based SLPs: Results of a nationwide survey. *Language, Speech, and Hearing Services in Schools, 47*(3), 246–258. https://doi.org/10.1044/2016_lshss-15-0044

Pawłowska, M. (2014). Evaluation of three proposed markers for language impairment in English: A meta-analysis of diagnostic accuracy studies. *Journal of Speech, Language, and Hearing Research, 57*(6), 2261–2273. https://doi.org/10.1044/2014_jslhr-l-13-0189

Plante, E., & Vance, R. (1994). Selection of preschool language tests: A data-based approach. *Language, Speech, and Hearing Services in Schools, 25*(1), 15–24.

Redmond, S. M. (2016). Language impairment in the attention-deficit/hyperactivity disorder context. *Journal of Speech, Language, and Hearing Research, 59*(1), 133–142.

Rice, M. L., Wexler, K., & Hershberger, S. (1998). Tense over time: The longitudinal course of tense acquisition in children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 41,* 1412–1431. https://doi.org/10.1044/jslhr.4106.1412

Sackett, D. L. (1991). *Clinical epidemiology: A basic science for clinical medicine* (2nd ed.). Boston, MA: Little, Brown and Company.

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

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

Schneider, P., & Hayward, D. (2010). Who does what to whom: Introduction of referents in children's storytelling from pictures. *Language, Speech, and Hearing Services in Schools, 41*(4), 459–473. https://doi.org/10.1044/0161-1461(2010/09-0040)

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*(4), 224–238.

Schuele, C. M. (2013). Beyond 14 grammatical morphemes: Toward a broader view of grammatical development. *Topics in Language Disorders, 33*(2), 118–124.

Schuele, C. M., & Tolbert, L. (2001). Omissions of obligatory relative markers in children with specific language impairment. *Clinical Linguistics & Phonetics, 15*(4), 257–274.

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*(2), 324–339.

Semel, E., Wiig, E. H., & Secord, W. A. (1995). *Clinical Evaluation of Language Fundamentals–Third Edition (CELF-3)*. San Antonio, TX: The Psychological Corporation.

Semel, E., Wiig, E. H., & Secord, W. A. (2003). *Clinical Evaluation of Language Fundamentals–Fourth Edition (CELF-4)*. 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.

Southwood, F., & Russell, A. F. (2004). Comparison of conversation, freeplay, and story generation as methods of language sample elicitation. *Journal of Speech, Language, and Hearing Research, 47*(2), 366–376.

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*(10), 741–756. https://doi.org/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*(1), 61–72. https://doi.org/10.1044/0161-1461(2006/007)

Statistics Canada. (n.d.). *Canada dimensions: The people [on-line]*. Retrieved from http://www.statcan.ca

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

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*(4), 922–937. https://doi.org/10.1044/1092-4388(2008/068)

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

Tomblin, J. B., & Nippold, M. (2014). Features of language impairment in the school years). In J. B. Tomblin (Ed.), *Understanding individual differences in language development across the school years* (pp. 79–116). New York, NY: Psychology Press.

Watkins, R. V., & Rice, M. L. (1991). Verb particle and preposition acquisition in language-impaired preschoolers. *Journal of Speech and Hearing Research, 34*(5), 1130–1141.

Westerveld, M. F., & Gillon, G. T. (2010). Profiling oral narrative ability in young school-aged children. *International Journal of Speech-Language Pathology, 12*(3), 178–189.

Wiig, E. H., Secord, W. A., & Semel, E. (1992). *Clinical Evaluation of Language Fundamentals–Preschool (CELF-P)*. San Antonio, TX: The Psychological Corporation.

Wiig, E. H., Secord, W. A., & Semel, E. (2004). *Clinical Evaluation of Language Fundamentals Preschool–Second Edition (CELF-P2)*. San Antonio, TX: Pearson.