Identifying Risk for Specific Language Impairment with Narrow and Global Measures of Grammar

Sofía M. Souto, **Laurence B. Leonard**, and **Patricia Deevy**
Purdue University, West Lafayette, IN

Abstract

In this study, we ask: (1) whether measures of the developmental level of the tense/agreement morphemes used by children have diagnostic value, as has been found for tense/agreement consistency; and (2) whether global measures of accuracy can be applied to children four and five years of age. The spontaneous speech samples of 112 four- and five-year-olds with SLI or typical language were analyzed. Group differences were seen for the developmental level of the children's tense/agreement morpheme use, but diagnostic accuracy did not reach acceptable levels, in contrast to a measure of tense/agreement consistency applied to the same data. A global measure of grammatical accuracy was found to be useful, but more appropriate for screening children already viewed as at risk for language difficulties. These findings suggest that an extended period of tense/agreement inconsistency may be more central to SLI than alternative measures related to tense/agreement morphology.

In recent years, several measures of spontaneous speech have shown good diagnostic accuracy in distinguishing preschool-age children with specific language impairment (SLI) from their typically developing (TD) peers. One type of measure focuses on tense/agreement morphology. This area of grammar has been the center of much research on SLI. Many tense/agreement morphemes are used less consistently by children with SLI than by same-age peers and even younger children matched for mean length of utterance (e.g. Cleave & Rice, 1997; Conti-Ramsden, Botting, & Faragher, 2001; Hoover, Storkel, & Rice, 2012; Rice & Wexler, 1996) and verb inventory (Leonard, Miller, & Gerber, 1999). These morphemes include third person singular *–s*, past tense *–ed*, both copula and auxiliary *is, are, am, was, were* and, in some studies, auxiliary *do, did*, and *does*. Composites representing children's use of all of these morphemes combined have shown good sensitivity and specificity for the preschool period (Bedore & Leonard, 1998; Rice, 2003). Along with composite measures, formal tests that provide composite information for tense/agreement morpheme use have also been developed (Rice & Wexler, 2001).

Although composite measures of tense/agreement use in spontaneous speech have diagnostic value, they are greatly affected by differences in the number of obligatory contexts found for each morpheme. For example, in the "finite verb morphology composite"

used by Bedore and Leonard (1998), the number of obligatory contexts for all tense/agreement morphemes are combined and divided into the total number of tense/agreement morphemes actually produced. Therefore, the child's ability with the morphemes having the most obligatory contexts (e.g. copula *is*) will have a larger influence on the composite score than the child's ability with the morphemes having relatively few obligatory contexts (e.g. past tense *–ed*). Complicating interpretation further is the possibility that some morphemes may be produced as part of an unanalyzed whole. For example, the production of frequent forms such as *it's* and *that's* will result in the child being credited with copula *is* use, but these forms might overestimate the child's ability with this morpheme. If they are produced frequently, the child's tense/agreement composite score is likely to be misleading.

Hadley and her colleagues (Hadley & Holt, 2006; Hadley & Short, 2005) developed a measure of spontaneous tense/agreement morpheme use that avoids these difficulties by emphasizing the diversity of contexts in which these morphemes are used and excluding contexts that are often associated with unanalyzed productions. For example, as part of the "productivity score" used by these investigators, a child could receive five points for using copula *is* in combination with five different subjects (e.g. *The cup's over here*, *Mom's at work, That bus is yellow*, *My doggie's hungry, Is this yours?*) but instances in which the copula is contracted with a pronoun (as in *She's not here* and *It's blue*) are excluded.

Measures of this type show a clear developmental trajectory of tense/agreement morpheme use in the early years (Hadley & Holt, 2006) and seem to be relatively successful in identifying children at risk for SLI (Hadley & Short, 2005). Application of these measures to four- and five-year-olds seems largely successful in distinguishing children with SLI from their typically developing age-mates (Gladfelter & Leonard, 2013).

One characteristic of tense/agreement morpheme use that has not yet been examined is the developmental level of the morpheme produced. For example, in both the finite verb morphology composite and the productivity score, forms such as copula *is* and copula *are* have the same status. Yet, it is likely that copula *are* is a later attainment than *is*. One goal of the present study is to determine whether a measure that provides different scores according to the presumed developmental level of the tense/agreement morphemes might serve as a useful diagnostic measure for children four and five years of age. Toward this end, we make use of the Main Verb category of Developmental Sentence Scoring (DSS; Lee, 1974) – a measure of grammatical development in spontaneous speech that predates the recent interest in tense/agreement morphology. The Main Verb category consists of a wide range of tense/agreement forms that include not only the well-studied morphemes such as copula *is*, third person singular *–s*, and past tense *–ed*, but also modal auxiliaries (e.g. *can, will, could, should*), and auxiliary *have* forms, among others. Of particular interest is the fact that tense/agreement morphemes belonging to the same paradigm (e.g. the copula *be* paradigm) are assigned higher or lower scores depending on their developmental level. For example, use of copula *is* earns a score of 1 but use of copula *are* earns a score of 2; use of auxiliary *do* earns a score of 4 but use of auxiliary *does* earns a score of 6. Because all entries in the Main Verb category of the DSS involve the functional category Agreement (AGR) and/or Tense (T) in current linguistic theory, they appear to be worthy of closer investigation in any populations known to have difficulty with tense/agreement morphemes.

A second goal of the present study is to determine whether a more global measure of grammar in spontaneous speech might have diagnostic utility. In a study of three-year-olds, Eisenberg and Guo (2013) found that two measures of overall grammatical accuracy showed adequate sensitivity and specificity in distinguishing typically developing children from children with presumed language impairments. One of these measures was the percentage of utterances that earned a "sentence point" based on the DSS. A sentence point is an extra point that is awarded to a sentence if it is free of grammatical errors and contains both a subject and a verb. In the present study, we determine whether the sentence point of the DSS also serves to distinguish SLI and typically developing groups at four and five years of age.

We assumed that the two types of measures of interest in the present study would be used for quite different purposes, even if both were found to possess adequate sensitivity and specificity. The global, sentence point measure makes no distinction among levels of complexity or types of error. A sentence point is awarded to grammatical sentences that can range from *I see you* to *The dog that was covered in mud was chased by the horse*. Similarly, a sentence point is denied regardless of whether the error occurred in a lengthy sentence such as *Yesterday both my brother and my sister take the cows from the barn to the pasture* or in a very immature sentence such as *Me take that*. Such rough measures might be useful tools for initially identifying children whose language should be assessed in greater detail. However, the sentence point scores will not be helpful in identifying areas of special weakness or possible targets for intervention.

In contrast, a measure of the developmental level of a child's tense/agreement morpheme use would seem to offer not only a means of identifying a child at risk for SLI, but also an impression of the developmental level of specific tense/morphemes produced by the child. Used in conjunction with the finite verb morphology composite, such a measure might indicate whether children begin to produce later-developing tense/agreement morphemes while still showing inconsistency with earlier developing morphemes. Such information could have theoretical importance, because it is not yet clear whether children's transition from treating tense/agreement as optional to treating tense/agreement as obligatory is influenced in part by the size of a child's tense/agreement morpheme inventory.

A final goal of the present study was to evaluate the diagnostic properties of the overall DSS score (Lee, 1974). This longstanding measure has been translated into a computer-assisted version (Computerized Profiling; Long, Fey, & Channell, 2000), appears to show acceptable interjudge reliability (Fey, Cleave, Long, & Hughes, 1993) and continues to be used as a primary or supplemental measure to document language impairment in investigations of SLI (e.g. Leonard, Camarata, Pawłowska, Brown, & Camarata, 2008). The overall DSS score reflects the contribution of not only the Main Verb category and sentence point already discussed, but also the categories Indefinite Pronouns/Noun Modifiers, Personal Pronouns, Secondary Verbs, Negatives, Conjunctions, Interrogative Reversals, and Wh-Questions. To be sure, our theoretical motivation behind selecting the DSS was its provision of points that differ according to the presumed developmental level of the tense/agreement morphemes used. However, to our knowledge, the overall DSS score has not yet been assessed in terms of sensitivity and specificity. For this reason, it seemed opportune to provide such an assessment here.

Method

Participants

The spontaneous speech samples from 112 children provided the data for this study. All of the children had been participants in studies conducted by the Child Language Laboratory at Purdue University. These studies involved experimental tasks that focused on the children's grammatical, lexical, nonword repetition, or attention skills (Deevy, Wisman Weil, Leonard, & Goffman, 2010; Finneran, Francis, & Leonard, 2009; Leonard, Deevy et al., 2007; Leonard, Deevy, Wong, Stokes, & Fletcher, 2007).The spontaneous speech samples analyzed in this study were collected prior to administration of the experimental tasks to document the children's mean length of utterance and determine whether their developmental level was appropriate for participation in the experiments.

Sixty of the children ranged from 48 to 59 months of age; hereafter, these children are referred to as the four-year-olds. The remaining 52 children ranged in age from 60 to 70 months, and are referred to as the five-year-olds. All children were monolingual English speakers, passed a bilateral hearing screening at 500, 1000, 2000, and 4000 Hz, and scored above 85 on the Columbia Mental Maturity Scale (CMMS; Burgemeister, Blum, & Lorge, 1972), a test of nonverbal intelligence. In addition, for all children, parent reports indicated no history of neurological impairment. Twenty-eight four-year-olds and 22 five-year-olds met the criteria for SLI. These children were enrolled in language treatment programs or were scheduled to enter such a program. Spontaneous speech samples from these particular children were selected from the larger database because these children met our gold standard for language impairment, defined as a z-score of −2.20 (1st percentile) or lower on the Structured Photographic Expressive Language Test—Second Edition (SPELT-II; Werner & Kresheck, 1983), a test shown to have good sensitivity and specificity (Plante & Vance, 1994). This version of the SPELT was the version in use at the time of the original data collection. The cutoff of −2.20 was selected because no typically developing child from the same database scored lower than a z-score of −0.99.

The remaining 62 children were typically developing (TD). Their language development was proceeding as expected according to parent report. These children's z-scores on the SPELT-II ranged from −0.88 (19th percentile) to +2.33 (99th percentile). Thirty-two of the TD children were four-year-olds; as a group, these children were closely matched in age to the four-year-olds with SLI. The similarity in ages can be seen in Table 1. The remaining 30 TD children were five years of age and were closely matched in age to the five-year-olds with SLI.

Within both the SLI and TD groups at each age level, one-half of the children with SLI and one-half of the TD children were selected at random and designated the "original" SLI-TD pair. The remaining children were designated the "replication" SLI-TD pair. The breakdown of children in each of these pairs is also shown in Table 1. Analyses of the original pairs were conducted and identical analyses were then performed for the replication pairs to determine the degree to which the findings were consistent.

Spontaneous Speech Samples

Each child's spontaneous speech sample was collected during interactions with the experimenter while engaged in play with age-appropriate toys. The experimenter primarily allowed the child to initiate conversation, but occasionally implemented prompts to promote further spontaneous language production. All children produced a minimum of 100 spontaneous utterances. The children's samples were transcribed using the Systematic Analysis of Language Transcripts software (SALT; Miller & Chapman, 2000). The SALT transcriptions were converted using the Computerized Profiling software (CP) for DSS (Lee, 1974) analysis. Scoring rules described by Lee (1974) were followed to obtain a DSS score, but where rules were ambiguous, those proposed by Hughes, Fey, and Long (1992) were implemented. Following DSS guidelines, all measures were calculated using the first 50 utterances that contained both a subject and a verb. Specific scoring details are discussed below for each measure employed.

Measures

Mean tense/agreement developmental score—Two of the principal measures of interest in this study were those that provided a developmental score for each use of a tense/agreement form in a child's speech sample. These scores came from the Main Verb category of the DSS categorization scheme. The Main Verb category is one of eight categories employed by DSS to arrive at an overall DSS score. The other categories are Indefinite Pronouns/Noun Modifiers, Personal Pronouns, Secondary Verbs, Negatives, Conjunctions, Interrogative Reversals, and Wh-questions. The scores within the Main Verb category range from a score of 1 (e.g. for copula *is*) to a score of 8 (e.g. for a modal in combination with present perfect, as in *could have eaten*). With the exception of correct zero-marked lexical verbs (e.g. *want* in *I want this*) that earn a score of 1, all verb forms that are scored in the Main Verb category require the use of an overt form marking agreement and/or tense. One measure of interest was each child's mean tense/agreement developmental score, calculated by adding the Main Verb category scores for each of the 50 utterances and dividing the total by the number of utterances that earned at least a score of 1 for this category. Note that this measure excluded errors, to render the measure as distinct as possible from the more traditional finite verb morphology composite that takes errors into consideration. An example involving the computation of the mean tense/agreement developmental score is shown in the Appendix.

Mean of the five highest tense/agreement developmental scores—Because the developmental level of the children's use of tense/agreement forms was the essential point of interest, we recognized that an alternative measure of tense/agreement developmental level might be important. The preceding measure constitutes a mean across 50 utterances. Frequent use of low-scoring but fully grammatical tense/agreement forms as in *She sees me*, and *That's mine* could lower the mean and as a result obscure a child's ability to use more advanced, but less frequently occurring forms. Accordingly, as a second measure of developmental level, we calculated the mean of each child's five highest scoring tense/agreement forms. The Appendix provides an example. This measure is not intended to reflect a child's customary developmental level, but provides an estimation of the developmental level at which the child can operate in select instances.

Finite verb morphology composite score—We also computed the finite verb morphology composite of Bedore and Leonard (1998) as a means of determining whether our sample of children was representative of the wider preschool SLI and TD populations. The finite verb morphology composite has shown good diagnostic accuracy across studies and we expected the same would be true for the children in the present study. This measure could help us determine if any lack of group differences or poor diagnostic accuracy on the part of the two measures of tense/agreement developmental level might be attributable to the measures themselves and not to the accidental selection of an unrepresentative sample of children.

The finite verb morphology composite was computed by totaling the number of obligatory contexts for third person singular –*s*, past tense –*ed*, copula *is, are, am,* and auxiliary *is, are, am,* and calculating the percentage of these contexts in which the morpheme was actually produced. For these calculations, we used all utterances beginning with the child's first utterance in the sample to the point at which the 50th utterance containing a subject plus verb was identified for the purpose of the just-described tense/agreement developmental level scores. This resulted in a somewhat smaller sample size than is customarily used to calculate the finite verb morphology composite. However, if a larger sample size had been used, it could not be determined if the success of this measure relative to the developmental level scores were due to the measures themselves or to the advantage of a larger sample size.

Sentence point—Another goal of this study was to determine whether a global measure of grammatical accuracy would function well in the four- and five-year-old age range. Eisenberg and Guo (2013) successfully employed the sentence point from the DSS for this purpose for a sample of children three years of age. We also used the DSS sentence point for this purpose. For each sentence produced by the child, a sentence point was awarded if and only if the sentence was fully grammatical, regardless of whether it employed simple or complex morphosyntax. Any grammatical error was sufficient to deny a sentence point; errors were not limited to the grammatical forms represented in the eight DSS categories. Thus, a sentence point would be withheld not only for, say, an utterance such as *Her broke the window* (an error from the Personal Pronoun category), but also utterances such as *Dad built new birdhouse* and *Mom ate two apple* that involve errors of grammatical forms (articles, noun plural inflections) that are not represented in the DSS categories. The sentence point score was obtained by adding the total number of sentence points earned by the child and dividing this number by 50, the number of utterances employed.

Overall DSS score—Because we employed verb forms from the Main Verb category of the DSS for our tense/agreement developmental level measures and the sentence point of the DSS for our global measure of grammatical accuracy, it seemed opportune to assess the diagnostic accuracy of the overall DSS score as well. To our knowledge, this assessment has not yet been made. The overall DSS score is computed by totaling the number of points earned for each utterance, including the sentence point, and dividing the total by 50, the number of utterances used in the sample.

Reliability

To assess scoring reliability, we randomly selected the spontaneous speech samples from 16 children, equally divided according to age (four-, five-year-olds), diagnostic group (SLI, TD), and type of pair (original, replication). An independent judge scored all sentences in each of these samples according to tense/agreement development score, and also identified those sentences containing the five highest tense/agreement developmental scores in each sample. Reliability for awarding a sentence point was also calculated. Sentence-by-sentence agreement between the original judge and independent judge for the tense/agreement developmental score was 95%. Agreement between the two judges in the identification of the five highest tense/agreement developmental scores was 88%. For the sentence point, agreement was 94%.

Results

Mean Tense/Agreement Developmental Score

The first analysis of interest was a comparison of the children's mean tense/agreement developmental level scores – scores that reflect the average point value of the tense/agreement verb forms actually produced by the child, with no deductions for errors. For each age group, we compared the SLI and TD groups in the original and replication pairs, using a two-factor analysis of variance (ANOVA) with diagnostic group (TD, SLI) and pair (original, replication) as between-subjects factors. For the four-year-olds, no main effect was found for diagnostic group, $F (1, 56) = 1.14$, $p = .29$, $\eta^2_p = .02$, pair, $F (1, 56) = 1.15$, $p = .29$, $\eta^2_p = .02$, or the interaction between these two factors, $F (1, 56) = 1.16$, $p = .29$, $\eta^2_p = .02$. A summary of the children's scores can be seen in Table 2.

The same type of ANOVA for the five-year-olds showed a main effect for diagnostic group, $F (1, 48) = 10.71$, $p = .002$, $\eta^2_p = .18$. Neither pair, $F (1, 48) = 1.47$, $p = .23$, $\eta^2_p = .03$, nor the interaction, $F (1, 48) = 0.76$, $p = .39$, $\eta^2_p = .02$, pair, was significant. A summary of the children's scores appears in Table 2.

Diagnostic accuracy was examined through logistic regression analysis. Given the absence of a difference for diagnostic group among the four-year-olds, sensitivity and specificity were not expected to be satisfactory. For five-year-olds, however, diagnostic accuracy might have been better as clear group differences were found for this age group. However, as can be seen in Table 2, diagnostic accuracy was not impressive for the five-year-olds. In particular, sensitivity for the original pair of five-year-olds was inadequate.

Mean of the Five Highest Tense/Agreement Developmental Scores

The four-year-old TD children showed significantly higher scores than the children with SLI based on the mean of the five highest tense/agreement developmental level scores, $F (1, 56) = 10.92$, $p = .002$, $\eta^2_p = .16$. No difference was found between the original and replication pairs, $F (1, 56) = 1.92$, $p = .17$, $\eta^2_p = .03$. However, there was a diagnostic group $\times$ pair interaction, $F (1, 56) = 7.67$, $p = .008$, $\eta^2_p = .12$. Post-hoc testing at the .05 level indicated that the children in the original TD group scored higher than the children with SLI in the original group, but the children in the replication TD group did not differ from the children

with SLI in the replication group. Although, as expected, the TD children in the original and replication groups did not differ, this was not true for the children with SLI. The children with SLI in the original group had significantly lower scores than those in the replication group. Table 3 provides a summary of the children's scores.

Group differences were clearer for the five-year-olds, $F(1, 48) = 34.23$, $p < .001$, $\eta^2_p = .42$. No difference was seen for pair, $F(1, 48) = 1.12$, $p = .29$, $\eta^2_p = .02$, or for the interaction, $F(1, 48) = 0.52$, $p = .48$, $\eta^2_p = .01$. A summary can be seen in Table 3.

As can be seen in Table 3, the diagnostic accuracy for four-year-olds was unsatisfactory. For the replication pair, sensitivity was strikingly poor. Although diagnostic accuracy was adequate for the five-year-old replication pair, sensitivity did not achieve adequacy for the original pair.

Finite Verb Morphology Composite Scores

Although the two types of measures that provided developmental levels of tense/agreement morpheme use showed generally higher scores for the TD children, diagnostic accuracy was not impressive. To gain a clearer perspective on this finding, however, it seemed necessary to document that our participants were representative of the wider literature. Toward this end, we also compared the children on the more traditional finite verb morphology composite. This composite involves a smaller collection of tense/agreement morphemes, but treats all of these morphemes equally. Furthermore, unlike the measures considered thus far, scores are lowered when children fail to produce these morphemes in obligatory contexts. To form a proper comparison with the other measures, it was necessary to compute the finite verb morphology composite on the same range of utterances considered for these other measures. Usually, this composite is calculated on larger samples sizes, so the outcome in this instance was not certain.

The ANOVA on the data from the four-year-olds revealed a significant main effect for diagnostic group, $F(1, 56) = 90.21$, $p < .001$, $\eta^2_p = .62$. Nonsignificant differences were seen for both pair, $F(1, 56) = 0.77$, $p = .38$, $\eta^2_p = .01$, and the interaction, $F(1, 56) = 0.20$, $p = .66$, $\eta^2_p = .01$. Nearly identical findings were seen for the five-year-olds, with diagnostic group showing a significant main effect, $F(1, 48) = 64.45$, $p < .001$, $\eta^2_p = .57$, and neither a significant effect for pair, $F(1, 48) = 0.75$, $p = .39$, $\eta^2_p = .02$, nor for the interaction, $F(1, 48) = 0.63$, $p = .43$, $\eta^2_p = .01$. Table 4 provides a summary.

Diagnostic accuracy was good for the four-year-olds, as can be seen in Table 4. For the five-year-olds, the sensitivity and specificity values can be considered acceptable, with the sensitivity for the five-year-old replication pair of 82% representing the lowest value.

Mean Sentence Point

A second goal of the study was to determine whether a global measure of grammatical accuracy might be suitable as an initial tool for identifying risk for language impairment, following findings reported by Eisenberg and Guo (2013) for younger children. We employed the DSS sentence point for this purpose.

For four-year-olds, a significant main effect was observed for diagnostic group, $F(1, 56) = 127.28$, $p < .001$, $\eta^2_p = .69$, with no difference for either pair, $F(1, 56) = 0.75$, $p = .39$, $\eta^2_p = .01$, or the diagnostic group × pair interaction, $F(1, 56) = 0.05$, $p = .83$, $\eta^2_p = .01$. The picture was more complicated for the data from the five-year-olds. A main effect for diagnostic group was seen, favoring, as expected, the TD children, $F(1, 48) = 106.95$, $p < .001$, $\eta^2_p = .69$. However, both a main effect for pair, $F(1, 48) = 4.62$, $p = .04$, $\eta^2_p = .09$, and a marginally significant interaction, $F(1, 48) = 3.62$, $p = .06$, $\eta^2_p = .07$, were also seen. Post-hoc testing indicated that the interaction could be attributed to significantly higher scores for the children with SLI in the original pair than the children with SLI in the replication pair. This can be seen quite readily in Table 5. However, this interaction did not affect the comparisons between the TD and SLI groups; for both the original pair and the replication pair, the TD children scored significantly higher than the children with SLI.

Likewise, the interaction seen for the sentence point data for the five-year-olds had no ill effects on diagnostic accuracy. As shown in Table 5, diagnostic accuracy was good for the four-year-olds and even better for the children five years of age.

Overall DSS Score

The final goal of the study was to evaluate the diagnostic accuracy of the overall DSS score – the score given to a child's language sample based on all of the grammatical categories included on the DSS along with the sentence point. Table 6 provides a summary of the results. Diagnostic group differences were evident for the four-year-olds, $F(1, 56) = 49.06$, $p < .001$, $\eta^2_p = .47$. The original pair and replication pair did not differ, $F(1, 56) = 0.13$, $p = .72$, $\eta^2_p = .01$; likewise, the diagnostic group × pair interaction was nonsignificant, $F(1, 56) = 0.06$, $p = .81$, $\eta^2_p = .01$. Similar results were obtained for the five-year-olds, with a significant main effect for diagnostic group, $F(1, 48) = 37.57$, $p < .001$, $\eta^2_p = .44$, and no significant difference for either pair, $F(1, 48) = 1.31$, $p = .26$, $\eta^2_p = .03$, or the diagnostic group × pair interaction, $F(1, 48) = 0.13$, $p = .72$, $\eta^2_p = .01$.

Unfortunately, diagnostic accuracy for the overall DSS score was not as encouraging as the group differences might suggest. In particular, for both four- and five-year-olds, sensitivity for the original pair fell short of adequacy levels. This can be seen in Table 6.

To gain additional perspective on our findings for the overall DSS score, we compared the mean overall DSS scores of our TD participants with those in Lee's (1974) normative sample. The mean overall DSS score for Lee's four-year-olds was 8.04 (SD = 1.64); for our four-year-old TD participants (original and replication combined), the mean was 8.37 (SD = 1.87). The effect size for this comparison (d = 0.19) does not quite reach the level considered to be a small effect size (Cohen, 1988). This value means that 85.5% of the scores from the two are overlapping. For five-year-olds, Lee's sample and ours overlapped to an even greater extent. Lee's five-year-olds had a mean of 9.19 (SD = 1.90), whereas our five-year-old TD participants had a mean of 9.17 (SD = 1.91), resulting in an effect size of 0.01. Given the similarity between our TD participants and the normative sample of Lee, we have no reason to suspect that our children were unrepresentative.

Discussion

The design of this study seemed quite suitable for assessing the diagnostic accuracy of the measures of interest. Children were initially identified as exhibiting a language impairment or typical language based on a recognized "gold standard" – the SPELT-II – which was the version of the SPELT in use at the time of the original data collection. In addition, at each age level of four and five years, the SLI and TD groups in each pair were closely matched for age; likewise the original and replication pairs were well-matched. Following DSS guidelines, spontaneous samples were limited to 50 utterances. Although larger samples might have produced a different outcome, we saw that for most of the measures employing these 50-utterance samples as the source of data, rather large differences between the TD and SLI groups were seen.

The first goal of this study was to determine whether measures of the developmental level of the tense/agreement morphemes used would yield accurate classifications of children with SLI or typical language development. One such measure – the mean tense/agreement developmental level score – was not satisfactory in this regard. The scoring for this measure did not penalize children for omitting tense/agreement morphemes in obligatory contexts. However, it is possible that this measure was not diagnostic because all children – those with SLI and those with typical language – may use many low-scoring tense/agreement morphemes. For example, assume that in a 50-utterance sample a TD child showed no omissions and used 35 examples of morphemes earning 1 point (such as copula *is*), 10 examples of morphemes earning 2 points (such as past tense *–ed* and copula *am*), and 5 examples of morphemes earning 4 points (such as the modals *can* and *will*). The child's mean would be 75/50 or 1.50. Now assume that in a 50-utterance sample a child with SLI used tense/agreement morphemes in only 25 (50%) of the 50 obligatory contexts, including 15 examples of morphemes earning 1 point, and 10 examples earning 2 points. This child's mean would be 35/25 or 1.40. Note that although the latter child produced nothing earning higher than 2 points (and was very inconsistent in those morphemes used), she did not differ dramatically from the TD child in the mean tense/agreement developmental level score. The similarity in scores is owing to the relatively high frequency of low-scoring morphemes used by both children. In certain speaking contexts, even adult speakers might show use similar to this TD child.

Given the observations above, the second measure involving developmental level – the mean of the five highest tense/agreement developmental level scores – would seem to hold greater promise. This measure is not unduly influenced by the frequent use of tense/agreement morphemes that earn relatively few points. In the example used above, the TD child would have a mean score of 4 (thanks to the 5 productions of modal auxiliaries) whereas the child with SLI would have a mean score of only 2 (by taking 5 of her productions of morphemes that earn her highest score of 2).

However, in practice, the mean of the five highest tense/agreement developmental scores did not prove as discriminating as had been expected. Group differences for both the original and replication pairs were seen only for the five-year-olds, and, for the original five-year-old pair, sensitivity was not adequate. Because the findings were more encouraging for our five-

year-old participants than for our four-year-olds (see Table 3), it is possible that this particular measure will prove diagnostically useful at somewhat older ages. Research exploring this possibility would be worthwhile.

Diagnostic accuracy was quite good for the finite verb morphology composite score, a measure that is driven largely by whether children are consistent in using tense/agreement morphemes in obligatory contexts. It is noteworthy that this result occurred considering that, by using the same range of utterances used for the DSS-related measures, the finite verb morphology composite was operating with a smaller sample size than is customarily used.

Together the findings indicate that the primary feature distinguishing SLI from TD groups at four and five years of age is probably not the degree to which the children produce tense/ agreement morphemes beyond the earlier-developing forms. Rather, the important feature appears to be how reliably the children produce these earlier-developing forms when the sentence context obligates them.

We do not presume that children with SLI acquire later-developing tense/agreement morphemes as readily as their TD peers. However, at least in a sample of 50 utterances, their highest-scoring morphemes are not sufficiently different from those seen of TD children to be diagnostically dependable.

If correct, this interpretation is in line with the view proposed by Rice, Wexler, and their colleagues (Rice & Wexler, 1996; Rice, Wexler, & Cleave, 1995; Rice, Wexler, & Hershberger, 1998) that the grammars of preschool-age children with SLI incorporate the features of tense and agreement but treat these features as optional. Given that group differences were not great for the mean of the highest five tense/agreement developmental scores, children with SLI seem to be capable of using later-developing morphemes to some extent.

A treatment implication of this finding is that a greater premium might be placed on assisting children in producing tense/agreement morphemes more consistently. Expanding the children's repertoire to include later-developing tense/agreement morphemes may certainly be helpful, but may not prove to be the children's greatest stumbling block.

Once treatment goals can move beyond helping a child gain greater consistency of tense/ agreement morpheme use, the goal of fostering the tense/agreement ability assessed by Hadley and her colleagues (Hadley & Holt, 2006; Hadley & Short, 2005) might be preferable to teaching later-developing morphemes. Recall that Hadley and her colleagues employed measures that credited children according to the variety of sentence contexts in which the morphemes appeared. Although developmental level is not considered (e.g. auxiliary *was* has the same value as copula *is*), and errors do not enter into the calculation, Gladfelter and Leonard (2013) found that these measures met appropriate levels of sensitivity and specificity for four- and five-year-old children. For a comprehensive and diagnostically useful picture of tense/agreement morpheme use by preschoolers with SLI, then, a combination of the finite verb morphology composite measure and the measures of Hadley and colleagues would seem to be most informative.

A second goal of the study was to determine whether a single, global measure of grammatical accuracy might serve as an initial means of identifying four- and five-year-olds at risk for SLI. Eisenberg and Guo (2013) evaluated two global measures of accuracy and found that each was suitable for use with a group of three-years-olds. One of these was the DSS sentence point that was employed in the present study. We found that for both four- and five-year-olds, the sentence point exhibited very good diagnostic accuracy in distinguishing children with SLI from their typically developing peers.

Our view that the DSS sentence point can be a useful early indicator of risk for SLI is not a claim that this measure will meet the criterion as a screening measure for the general four- and five-year-old population. Our study – like so many others that have evaluated diagnostic accuracy – made use of children who were already identified (by some gold standard) as exhibiting SLI or typical language development. However, a screening instrument for children in the general population – whose language status might be completely unknown – requires taking into account the prevalence of SLI, estimated to be 7% (Tomblin et al., 1997). An accurate screening tool for this purpose must be relatively successful in selecting those 7% who are at risk for the disorder, while also passing the remaining 93% of children. For example, using the 93% sensitivity and 94% specificity values seen for the original pair of four-year-olds (see Table 5), one can calculate an adjusted post-test probability of disorder based on prevalence, following the procedures of Sackett, Haynes, Guyatt, and Tugwell (1991). The result indicates that, whereas before determining the DSS sentence point the probability of identifying a child with SLI is 7% (the prevalence of the disorder), this probability increases to 54% once the mean sentence point is determined to be below the established cutoff. Such a large increase in probability from 7% to 54% is regarded as clinically important (Straus, Richardson, Glasziou, & Haynes, 2005). However, many children who failed the screening based on this cutoff would be found to be functioning at typical levels of language after further testing.

Although the higher standard for universal screening is not met by the DSS sentence point, we must mention that, to our knowledge, there is no screening measure of preschool-age language ability that meets this standard. Even the best of the measures will result in a very large number of children who fall below the cutoff but who will, upon further testing, prove to be within the typical range of language development. Future research might reveal that a global measure such as the mean sentence point, when used in conjunction with another (yet-to-be-developed) measure, can yield an accuracy level that can be safely used with the general preschool population.

A third goal of this study was to evaluate the diagnostic accuracy of the overall DSS score, long used in the study of childhood language disorders. Using this measure, we found very large differences between the SLI and TD groups at both four and five years of age. These differences were evident for the replication pair as well as the original pair. Such findings reinforce the view that the overall DSS score is a useful measure for studying differences at the group level. We also suspect that this measure can provide considerable descriptive information for clinical purposes.

Unfortunately, sensitivity did not meet acceptable levels at either age. We do not believe that this finding can be attributed to the selection of unrepresentative children for our study. As noted in the Results, our TD children showed significant overlap in scores with the children in Lee's (1974) normative sample.

The diagnostic accuracy findings were somewhat surprising given that our participants with SLI were dramatically lower than our TD participants on the SPELT-II. Children with significant grammatical difficulties will not score well on the SPELT-II. However, although the SPELT-II, like the overall DSS score, reflects a wide range of grammatical details, its scoring of each item is all-or-none. In contrast, on the DSS, any utterance produced by the child is likely to earn at least some minimal number of points. For example, an utterance such as *They gonna see the movie* will earn points for the Personal Pronoun and Secondary Verb categories even though no points will be earned for the Main Verb category and no sentence point will be awarded. The fact that even ungrammatical utterances can earn points toward the overall DSS score may have narrowed the differences between the children with SLI and the TD children. Consistent with this possibility is our earlier-discussed finding that the DSS sentence point – a measure that is sensitive to any and all grammatical errors – showed very good diagnostic accuracy.

In summary, the results of this study are important in showing that different grammatical measures – even those that pertain to the same broader category of tense/agreement morphology – do not yield equivalent results for children with SLI. Based on the developmental level of the tense/agreement morphemes used, preschoolers with SLI are not highly distinguishable from their typically developing peers. Other measures of tense/agreement morpheme use – such as measures of consistency and productivity – would seem to be more diagnostic, and probably strike closer to the children's core difficulties. At a more global level, the simple identification of grammatical errors has considerable promise for the screening of preschoolers for whom there is already suspicion of language difficulties. The DSS remains a useful tool, given that its scoring includes the sentence point and one can use the 50-utterance format as a basis for computing consistency of tense/agreement morpheme use. The overall DSS score continues to be viable for comparisons at the group level, but for four- and five-year-olds, its sensitivity, in particular, has shortcomings.

Acknowledgments

Appendix. An example of scores assigned from the Main Verb Category of Developmental Sentence Scoring to arrive at the mean tense/agreement

**developmental level and the mean of the five highest tense/agreement
developmental scores based on a 50-uterance sample**

	Utterance	Main Verb Category Example	Points Earned	Among the Five Highest
1	These are koosh balls.	copula *are*	2	
2	But I was gonna bring Guess Who.	auxiliary *was*	2	
3	I don't know your names.	obligatory *do*	4	
4	I can read.	modal *can*	4	
5	I am higher than them.	copula *am*	2	
6	I forgot that.	irregular past	2	
7	I've seen a treehouse before.	*have* + verb + *-en*	7	7
8	What does this go on?	obligatory *does*	6	6
9	He's climbing in the tree.	auxiliary *is*	1	
10	I will make him taller.	modal *will*	4	
11	He is sitting on a dragon.	auxiliary *is*	1	
12	I ate that one already.	irregular past	2	
13	I messed up.	past tense *-ed*	2	
14	Why is that mouse sad?	copula *is*	1	
15	What are these?	copula *are*	2	
16	She's a baby.	copula *is*	1	
17	We could go there.	modal *could*	6	6
18	We borrowed one.	past tense *-ed*	2	
19	She needs a tail.	third person singular *-s*	2	
20	I'll be the cowgirl.	modal *will*	4	
21	This is the kid.	copula *is*	1	
22	I would have gone with her.	modal + *have* + verb + *-en*	8	8
23	She's riding the horse.	auxiliary *is*	1	
24	She took the rope.	irregular past	2	
25	We were locking it.	auxiliary *were*	2	
26	This must be his.	*must* + verb	7	7
27	John used that one.	past tense *-ed*	2	
28	But Daddy holds me.	third person singular *-s*	2	
29	I can't get this on.	modal *can*	4	
30	He does like it.	emphatic *does*	6	
31	There's a tree.	copula *is*	1	
32	Mine were bigger.	copula *were*	2	
33	I did that.	irregular past	2	
34	She sees me on the tv.	third person singular *-s*	2	
35	It's a Cinderella thing.	copula *is*	1	
36	It's hurting my eyes.	auxiliary *is*	1	
37	I can do it by myself.	modal *can*	4	
38	That was my horse.	copula *was*	2	

	Utterance	Main Verb Category Example	Points Earned	Among the Five Highest
39	Did you see that movie?	obligatory *did*	6	
40	She's gonna be this for Halloween.	auxiliary *is*	1	
41	We played that game already.	past tense *-ed*	2	
42	It's not a real horse.	copula *is*	1	
43	I made a shark right now.	irregular past	2	
44	It's my brother's shirt.	copula *is*	1	
45	I got them because of my costume.	irregular past	2	
46	I also saw those two foxes.	irregular past	2	
47	Will she go too?	modal *will*	4	
48	He likes snacks.	third person singular *-s*	2	
49	This snake is coiling up.	auxiliary *is*	1	
50	It is the green one.	copula *is*	1	
	Mean tense/agreement developmental score:		2.64	
	Mean of the five highest tense/agreement scores:			6.80

References

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

Burgemeister, B.; Blum, L.; Lorge, J. Columbia Mental Maturity Scale. New York: Harcourt Brace Jovanovich; 1972.

Cleave P, Rice M. An examination of the morpheme *be* in children with specific language impairment: The role of contractibility and grammatical form class. Journal of Speech, Language, and Hearing Research. 1997; 40:480–492.

Cohen, J. Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Lawrence Erlbaum; 1988.

Conti-Ramsden G, Botting N, Faragher B. Psycholinguistic markers for specific language impairment. Journal of Child Psychology and Psychiatry. 2001; 42:741–748. [PubMed: 11583246]

Deevy P, Weil L, Leonard L, Goffman L. Extending use of the NRT to preschool-aged children with and without specific language impairment. Language, Speech, and Hearing Services in Schools. 2010; 41:277–288.

Eisenberg S, Guo LW. Differentiating children with and without language impairment based on grammaticality. Language, Speech, and Hearing Services in Schools. 2013; 44:20–31.

Fey M, Cleave P, Long S, Hughes D. Two approaches to the facilitation of grammar in language-impaired children: An experimental evaluation. Journal of Speech and Hearing Research. 1993; 36:141–157. [PubMed: 7680731]

Finneran D, Francis A, Leonard L. Sustained attention in children with specific language impairment. Journal of Speech, Language, and Hearing Research. 2009; 52:915–929.

Gladfelter A, Leonard L. Alternative tense and agreement morpheme measures for assessing grammatical deficits during the preschool period. Journal of Speech, Language, and Hearing Research. 2013; 56:542–552.

Hadley PA, Holt JK. Individual differences in the onset of tense marking: A growth-curve analysis. Journal of Speech, Language, and Hearing Research. 2006; 39:984–1000.

Hadley PA, Short H. The onset of tense marking in children at risk for specific language impairment. Journal of Speech, Language, and Hearing Research. 2005; 48:1344–1362.

Hoover J, Storkel H, Rice M. The interface between neighborhood density and optional infinitives: Normal development and specific language impairment. Journal of Child Language. 2012; 39:835–862. [PubMed: 22123500]

Hughes DL, Fey ME, Long SH. Developmental sentence scoring: Still useful after all these years. Topics in Language Disorders. 1992; 12(2):1–12.

Lee, L. Developmental sentence analysis. Evanston, IL: Northwestern University Press; 1974.

Leonard L, Camarata S, Pawłowska M, Brown B, Camarata M. The acquisition of tense and agreement morphemes by children with specific language impairment during intervention: Phase 3. Journal of Speech, Language, and Hearing Research. 2008; 51:120–125.

Leonard L, Deevy P, Kurtz R, Krantz L, Owen A, Polite E, Elam D, Finneran D. Lexical aspect and the use of verb morphology by children with specific language impairment. Journal of Speech, Language, and Hearing Research. 2007; 50:759–777.

Leonard L, Deevy P, Wong AMY, Stokes S, Fletcher P. Modal verbs with and without tense: A study of English- and Cantonese-speaking children with specific language impairment. International Journal of Language and Communication Disorders. 2007; 42:209–228. [PubMed: 17365094]

Leonard LB, Miller C, Gerber E. Grammatical morphology and the lexicon in children with specific language impairment. Journal of Speech, Language, and Hearing Research. 1999; 42:678–689.

Long, S.; Fey, M.; Channell, R. Computerized profiling (CP) Version 9.2.7. Cleveland, OH: Department of Communication Sciences, Case Western Reserve University; 2000.

Miller, J.; Chapman, R. Systematic Analysis of Language Transcripts (Research Version 6.1a) [Computer Software]. Madison: University of Wisconsin-Madison, Language Analysis Laboratory; 2000.

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

Rice, ML. A unified model of specific and general language delay: Grammatical tense as a clinical marker of unexpected variation. In: Levy, Y.; Schaeffer, J., editors. Language competence across populations: Toward a definition of specific language impairment. Mahwah, NJ: Erlbaum; 2003. p. 63-95.

Rice ML, Wexler K. Toward tense as a clinical marker of specific language impairment in English-speaking children. Journal of Speech, Language, and Hearing Research. 1996; 39:1239–1257.

Rice, ML.; Wexler, K. Rice/Wexler Test of Early Grammatical Impairment. San Antonio, TX: Psychological Corporation; 2001.

Rice ML, Wexler K, Cleave P. Specific language impairment as a period of extended optional infinitive. Journal of Speech and Hearing Research. 1995; 38:850–863. [PubMed: 7474978]

Rice ML, Wexler K, Hershberger S. Tense over time: The longitudinal course of tense acquisition in children with specific language impairment. Journal of Speech, Language, and Hearing Research. 1998; 41:1412–1431.

Sackett, D.; Haynes, R.; Guyatt, G.; Tugwell, P. Clinical epidemiology. Boston, MA: Little, Brown; 1991.

Straus, S.; Richardson, W.; Glasziou, P.; Haynes, R. Evidence-based medicine: How to teach and practice EBM (Third Edition). London, England: Elsevier; 2005.

Tomblin JB, Records N, Buckwalter P, Zhang X, Smith E, O'Brien M. The prevalence of specific language impairment in kindergarten children. Journal of Speech, Language, and Hearing Research. 1997; 39:1284–1294.

Werner, EO.; Kresheck, JD. Structured Photographic Expressive Language Test – II. DeKalb, IL: Janelle Publications; 1983.

Table 1

The ages (in months) of the four- and five-year-old children in the original and replication pairs.

	Original Pair			Replication Pair		
	N	Mean (SD)	Range	N	Mean (SD)	Range
Four Years						
SLI	14	53.29 (2.67)	48–58	14	53.29 (3.47)	48–58
TD	16	53.38 (3.54)	48–59	16	53.56 (3.33)	48–59
Five Years						
SLI	11	64.91 (3.39)	60–70	11	64.45 (3.27)	60–70
TD	15	64.67 (3.37)	60–70	15	64.67 (3.31)	60–70

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 2

The mean tense/agreement developmental scores for the four- and five-year-old children in the original and replication pairs.

	SLI		TD		Sensitivity	Specificity
	M	(SD)	M	(SD)		
Four Years						
Original	1.60	(0.26)	2.08	(0.39)	79%	81%
Replication	1.97	(0.34)	1.88	(0.29)	50%	75%
Five Years						
Original	1.92	(0.19)	2.15	(0.35)	64%	80%
Replication	1.73	(0.34)	2.12	(0.39)	82%	80%

Table 3

The mean of the five highest tense/agreement developmental scores earned by the four- and five-year-old children in the original and replication pairs.

	SLI		TD		Sensitivity	Specificity
	M	(SD)	M	(SD)		
Four Years						
Original	3.74	(1.00)	5.23	(0.85)	71%	69%
Replication	4.76	(1.04)	4.89	(0.89)	14%	94%
Five Years						
Original	4.33	(0.49)	5.53	(0.89)	73%	87%
Replication	3.91	(1.10)	5.45	(0.76)	82%	87%

Table 4

The finite verb morphology composites of the four- and five-year-old children in the original and replication pairs.

	SLI		TD		Sensitivity	Specificity
	M	(SD)	M	(SD)		
Four Years						
Original	56.93	(23.91)	96.97	(4.27)	93%	94%
Replication	62.26	(21.05)	98.70	(2.44)	93%	100%
Five Years						
Original	70.00	(18.82)	97.46	(4.41)	91%	93%
Replication	63.69	(21.97)	97.19	(3.12)	82%	93%

Table 5

The mean sentence points earned by the four- and five-year-old children in the original and replication pairs.

	SLI		TD		Sensitivity	Specificity
	M	(SD)	M	(SD)		
Four Years						
Original	0.60	(0.14)	0.91	(0.08)	93%	94%
Replication	0.57	(0.24)	0.92	(0.04)	100%	100%
Five Years						
Original	0.70	(0.09)	0.93	(0.03)	100%	100%
Replication	0.59	(0.18)	0.92	(0.05)	100%	100%

Table 6

The overall DSS scores earned by the four- and five-year-old children in the original and replication pairs.

	SLI		TD		Sensitivity	Specificity
	M	(SD)	M	(SD)		
Four Years						
Original	5.53	(1.24)	8.24	(1.37)	79%	94%
Replication	5.58	(0.87)	8.50	(2.31)	93%	94%
Five Years						
Original	6.67	(1.06)	9.36	(2.31)	72%	87%
Replication	5.97	(1.26)	8.99	(1.46)	82%	87%