2016-06-01

Mean Length of Utterance and Developmental Sentence Scoring in the Analysis of Children's Language Samples

Laurie Lynne Chamberlain
Brigham Young University

Mean Length of Utterance and Developmental Sentence Scoring

in the Analysis of Children's Language Samples

Laurie Lynne Chamberlain

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

Ron W. Channell, Chair
Christopher Dromey
Shawn L. Nissen

Department of Communication Disorders

Brigham Young University

June 2016

ABSTRACT

Mean Length of Utterance and Developmental Sentence Scoring
in the Analysis of Children's Language Samples

Laurie Lynne Chamberlain
Department of Communication Disorders, BYU
Master of Science

Developmental Sentence Scoring (DSS) is a standardized language sample analysis procedure that uses complete sentences to evaluate and score a child's use of standard American-English grammatical rules. Automated DSS software can potentially increase efficiency and decrease the time needed for DSS analysis. This study examines the accuracy of one automated DSS software program, DSSA Version 2.0, compared to manual DSS scoring on previously collected language samples from 30 children between the ages of 2;5 and 7;11 (years;months). The overall accuracy of DSSA 2.0 was 86%. Additionally, the present study sought to determine the relationship between DSS, DSSA Version 2.0, the mean length of utterance (MLU), and age.

MLU is a measure of linguistic ability in children, and is a widely used indicator of language impairment. This study found that MLU and DSS are both strongly correlated with age and these correlations are statistically significant, $r = .605$, $p < .001$ and $r = .723$, $p < .001$, respectively. In addition, MLU and DSSA were also strongly correlated with age and these correlations were statistically significant, $r = .605$, $p < .001$ and $r = .669$, $p < .001$, respectively. The correlation between MLU and DSS was high and statistically significant $r = .873$, $p < .001$, indicating that the correlation between MLU and DSS is not simply an artifact of both measures being correlated with age. Furthermore, the correlation between MLU and DSSA was high, $r = .794$, suggesting that the correlation between MLU and DSSA is not simply an artifact of both variables being correlated with age. Lastly, the relationship between DSS and age while controlling for MLU was moderate, but still statistically significant $r = .501$, $p = .006$. Therefore, DSS appears to add information beyond MLU.

Keywords: Developmental Sentence Scoring, automated language sample analysis, automated Developmental Sentence Scoring, mean length of utterance

ACKNOWLEDGMENTS

My sincerest thanks to my thesis committee members, Dr. Christopher Dromey, and Dr. Shawn Nissen. I am especially thankful for my remarkable thesis chair, Dr. Ron Channell, who kindly and patiently helped me throughout the writing of this thesis. I feel very grateful to have been able to work with him. Next, I would like to thank Lori Taylor Banta, Melissa L. Barber, and Elizabeth Chamberlain Mitchell for collecting and transcribing the language samples used in the present study. Also, I would like to thank Amy Chamberlain and Dr. Janelle Macintosh for the editing assistance they provided, as well as Dr. Christopher Macintosh for his help with the statistics used in this study. Lastly, I'm very grateful for my wonderful children, their exceptional spouses, and my beautiful grandchildren who have given me love and encouragement during my years as graduate student.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF APPENDICES

DESCRIPTION OF THESIS CONTENT AND STRUCTURE

This thesis, *Mean Length of Utterance and Developmental Sentence Scoring in the Analysis of Children's Language Samples*, is part of a larger research project, and all or part of the data from this thesis may be published as part of articles listing the thesis author as a co-author. The thesis itself is to be submitted to a peer-reviewed journal in speech-language pathology. An annotated bibliography is presented in Appendix A.

Introduction

Two techniques for the quantitative analysis of naturalistic language samples have been widely used in both research and clinical practice for over 40 years. These quantitative measures are the mean length of utterance (MLU; Brown, 1973) and Developmental Sentence Scoring (DSS; Lee, 1974). However, many questions regarding these measures still remain unasked and unanswered. DSS requires much more time and training than MLU, but does DSS offer enough additional insight about a child's language to justify this higher resource cost? If a fully automated version of DSS were available which overcame the additional resource cost of DSS, would this automated version offer similar insight? The present study attempts to address these questions.

Language Sample Analysis

Language sample analysis (LSA) is a method of childhood language assessment. Its purpose is to systematically assess, describe, and aid the clinician in understanding a child's expressive language abilities. Generally, interactive conversation allows the collection of a natural language sample. Therefore, LSA provides data that are more representative of the child's true linguistic ability than elicited language assessed from a standardized test.

Most procedures for conducting an LSA involve four steps: recording the conversation, transcription of the language sample, analysis, and interpretation. Information obtained with LSA is useful for diagnosing a language disorder and determining a treatment plan (Klee & Fitzgerald, 1985).

LSA is widely used among speech language pathologists (SLP) due to its clinical usefulness. Kemp and Klee (1997) reported results from a survey conducted on a representative sample of SLPs in the United States. Their objective was to assess the clinical practices of SLPs with regard to LSA use. Respondents stated that they used LSA for diagnosis, intervention, and screening of language disorders in children. While only 8% of the respondents reported that the use of LSA was mandated by their states, 85% of the 253 respondents reported using LSA (Kemp & Klee, 1997).

Westerveld and Claessen (2014) conducted a similar study. SLPs from Australia were surveyed to determine clinician opinions and practices of LSA. Items surveyed included the purpose of language sampling, elicitation methods, transcription, and analysis. Of the 257 respondents, 90.8% reported routine language sample collection and analysis. The primary reasons that the 8.2% of respondents did not use LSA were time constraints, lack of training, and a lack of computer hardware or software. Furthermore, 87% of respondents reported often or always using an informal LSA procedure, while only 37% reported often or always completing a detailed LSA. Time constraints were reported as the main obstacle to detailed LSA use. Overall, these findings were consistent with the Kemp and Klee (1997) findings.

The disadvantages to LSA include the amount of knowledge required, the lack of consistency in procedures used for collection and elicitation, and the difficulty in obtaining a representative sample. In the Kemp and Klee (1997) survey, 86% of respondents reported a lack of time as the most common reason not to use LSA. The amount of time needed to perform an LSA is considered the greatest disadvantage (Hux, Morris-Friehe, & Sanger, 1993).

Mean Length of Utterance

Though it had long been known that the average number of words in children's sentences increased as the child grew older, Roger Brown first added the insight that counting morphemes rather than whole words was a more sensitive approach to grammatical development (Brown, 1973). A morpheme is the smallest unit of meaning, and each word is made up of one or more morphemes (Turnbull & Justice, 2012). Morphemes are of two types: *free* morphemes can stand alone as words, and *bound* morphemes must be attached to another morpheme in order to be used in a word. Bound morphemes can be either derivational or inflectional. Derivational morphemes change a word into a different grammatical category (such as the *-ly* morpheme changing the noun *friend* into the adjective *friendly*) or into a different word, requiring a separate dictionary entry. Inflectional morphemes add information but don't change the word's grammatical category, such as the plural morpheme, which changes *cat* into *cats*. Brown's insight was that early grammatical development was reflected by the child's increasing skill at using inflectional morphemes and that the child's grammatical development could be measured and scaled relative to the average number of morphemes, both free and inflectional, used in utterances. Used as a clinical measure, MLU is the average length in morphemes of a child's utterance, obtained by using a language sample of 50-100 spontaneous utterances. The total number of morphemes is divided by the total number of utterances to calculate the MLU (Turnbull & Justice, 2012).

Many studies have investigated the reliability and validity of MLU (Chabon, Kent-Udolf, & Egolf, 1982; DeThorne, Johnson, & Loeb, 2005; Klee & Fitzgerald, 1985; Rice, Redmond, & Hoffman, 2006; Rondal, Ghiotto, Bredart, & Bachelet, 1987). In addition, Eisenberg, Fersko, and Lundgren (2001) reported that MLU is one way of measuring utterance length, and can be used to identify preschool children with language impairment. Furthermore, studies have concluded that MLU is widely used for the quantitative assessment of children's syntactic development (Hickey, 1991; Klee & Fitzgerald, 1985).

MLU can be, and these days usually is, calculated quickly and easily by computer software. For example, Systematic Analysis of Language Transcripts (SALT) is software that elicits, and analyzes language samples. Another possibility, Child Language Analysis (CLAN; MacWhinney, 2006), contains programs for analyzing language. Both SALT and CLAN are efficient methods for calculating the MLU of a language sample.

Developmental Sentence Scoring

Developed in the early 1970s, DSS is a method used to analyze children's language samples. The purpose of DSS is to evaluate and score the grammatical rules within complete sentences of children who speak Standard American English (SAE). Eight areas of grammatical development are examined and scaled from a spontaneous language sample containing at least 50 sentences. The areas examined include: (a) indefinite pronoun or noun modifier, (b) personal pronoun, (c) main verb, (d) secondary verb, (e) negative, (f) conjunction, (g) interrogative reversal in questions, and (h) wh- questions. A point value score, ranging from one to eight points, is awarded for each of one or more grammatical structures with a sentence. Higher point values are awarded to more advanced developing grammatical forms. Summing the points given to each utterance, and dividing this sum by the number of analyzed sentences obtain the DSS score.

Published surveys have indicated that DSS was the language sample analysis most commonly used by clinicians (Hux et al., 1993; Kemp & Klee, 1997), although there have not been any newer surveys done since the 1990s. Nevertheless, DSS continues to be used in research studies to quantify syntactic development (e.g., Leonard, Fey, Deevy, & Bredin-Oja, 2014; Smith, DeThorne, Logan, Channell, & Petrill, 2014). For example, Smith et al. reported a longitudinal view of school-age language outcomes of twins born prematurely versus a control group of twins born full term. The syntactic complexity of each participant's language was measured using DSS.

DSS has several strengths, including its ability to provide a numeric score with norm referencing, its usefulness in verifying a language problem quantitatively, and its ability to provide help in the description of development (Channell, 2003). Additionally, DSS provides information that is useful in making clinical decisions. It is a valuable instrument for the assessment of grammatical development, aiding in diagnostic judgments, assisting in treatment planning, and measuring treatment progress (Hughes, Fey, & Long, 1992). With DSS, clinicians can also compare data from a child to chronologically aged peers by referencing the child's DSS score to normative data. Lastly, DSS not only separates children with disordered language from children with typical language, but it can also isolate the particular area of difficulty of the language user (Johnson & Tomblin, 1975), which can assist the clinician in selecting treatment goals and assessing the effectiveness of treatment.

DSS has several limitations. One is that the sample size is small, which may lead to undependable results. The standard criterion of a normed test is 100 participants per age group. DSS data do not meet this standard, having only 20 participants per age group (Johnson & Tomblin, 1975). A second limitation is the lack of diversity of the normative group. Children who were white and middle-class made up the majority of participants used for DSS normative data. Consequently, comparisons should not be the sole basis of making a diagnostic judgment (Hughes et al., 1992). A third disadvantage of DSS is that the norms are older than the 7-year recommended maximum (Salvia & Ysseldyke, 2007). A fourth limitation of DSS is that considerable training and time is required of clinicians to conduct a DSS analysis (Lively, 1984). Fifth, it is possible that the required minimal sample size of 50 utterances is too small, resulting in unreliable results (Johnson & Tomblin, 1975). Sixth, Klee (1985) reported that the typical grammatical forms(years;months) developed by Lee (1974) might be inaccurate. A seventh disadvantage is that the same DSS score can represent many varying language profiles. For these reasons, a DSS score may oversimplify syntactic abilities if it is not analyzed further. While DSS can differentiate children with language disorders from typically developing children (Liles & Watt, 1984), it was not designed to analyze all aspects of a child's language. Therefore, it should not be used independently to determine whether or not a child is language impaired (Lee, 1974).

Fristoe (1979) reported the amount of time needed as the greatest disadvantage of DSS. Fristoe also reported that one hour is the recommended time required to obtain an adequate language sample, which then requires transcription and scoring. Long (2001) conducted a study of 256 students and practicing SLPs. The purpose of Long's study was to compare the time efficiency of manual and computerized procedures for phonological and grammatical analysis. Both MLU and DSS were included in the grammatical analyses. Long reported the mean length of time needed to score DSS on two different samples as 56.2 and 75 minutes. Furthermore, Long stated that without exception, computerized analyses were completed faster than manual analyses, and had better or equal levels of accuracy.

Automated DSS Analysis Software

Automated DSS programs may increase efficiency of DSS analysis, allowing its use in clinical and research settings on a more regular basis. There are several automated DSS analysis software programs, used with varying degrees of success. One program that is able to perform automated DSS analysis is Computerized Profiling (CP), which Stephen H. Long initially developed in 1986. The initial version had several disadvantages, including restrictions on maximum corpus size, misanalyses of multiple embedded clauses, and word truncation (Klee and Sahlie, 1987). To help reduce problems in the initial version of the CP program, a probabilistic automated grammatical tagging program, GramCats (Channell 1998), was integrated.

Long and Channell (2001) reported that CP could produce four grammatical analyses: MLU, DSS, Language Assessment, Remediation and Screening Procedure (LARSP; Crystal, 1982; Crystal, Garman, & Fletcher, 1989), and the Index of Productive Syntax (IPSyn; Scarsborough, 1990). Language samples were obtained from 69 typically developing children, speech-impaired children, and language-impaired children ranging from 2;6 to 7;10. An updated version of CP was used and a percentage of accuracy for the automated software was obtained by comparing the results of the automated analyses to results of the manual analyses. An accuracy rate of 89.9% was reported for the CP DSS analysis (Long & Channell, 2001).

Channell (2003) used 48 language samples collected from school-age children, 28 of whom had language impairment, to analyze the accuracy of automated DSS analysis obtained from CP. The accuracy rate of automated DSS scoring compared to manual DSS scoring was 78.2% (SD = 4.4). The accuracy rate of this study was 11.7% lower than the Long and Channell (2001) study; the lower level of accuracy was considered to result from the greater linguistic complexity of the samples used. There was a high correlation ($r = .97$, $p < .0001$) between the manual and CP-computed scores. The CP-computed scores were consistently higher than the manual-computed scores, and the difference was statistically significant, $p < .0001$.

Child Language Analysis (CLAN; MacWhinney, 2006) is another automated program, which computes grammatical analyses including MLU, type token ratio, DSS, and Index of Productive Syntax (IPSyn). Files must be in Codes for Human Analysis of Transcripts format to complete DSS analysis using CLAN. Also, the sample must be run through a morphological analysis program, and use the part of speech tagging program to code the sample for parts of speech. There is both an automatic and interactive mode in the DSS program (MacWhinney, 2006). Currently, there are no data regarding the accuracy of automated DSS analysis completed with CLAN.

Judson (2006) completed a study to determine the extent to which Developmental Sentence Scoring Automated (DSSA) could replace manual scoring. According to Judson the accuracy level of DSSA was approximately 86%. This percentage was considered “acceptable” according to criteria recommended by Hughes, Fey, Kertoy, and Nelson (1994) and Long and Channell (2001) for use as a clinical measure. The accuracy of DSSA was approximately 86%. Judson concluded that accuracy levels were sufficiently high to allow automated use of DSSA by clinicians as an alternative to manual DSS scoring when used for language sample analysis.

Relating MLU and DSS

Assessing and quantifying the level of a child’s syntactic development from spontaneous speech samples is important for speech-language pathologists. MLU is the most common measure used for this assessment, is conceptually simple, and can be accurately calculated by computer. DSS is another method that can be used to quantify syntactic complexity, has been widely used, and now can be calculated by computer.

Lee (1974) reported that the correlation between MLU and DSS was moderate, $r = .74$. However, little is known about how well DSS results compare to MLU and whether DSS offers additional insight into a child's grammatical development beyond the insight offered by MLU.

Rice et al. (2006) reported on two studies to examine the concurrent validity and temporal stability of MLU. In the first study, participants were selected from Rice, Wexler and Hershberger’s (1998) analysis of children’s knowledge of grammatical tense marking. Three groups were assessed during this study: 39 children with specific language impairment (SLI) having a mean chronological age (*M*) of 58 months (range = 52-68 months), 40 younger typically developing children used as a control group (*M* = 36 months, range = 30-44 months), and 45 children of the same age used as a control group (*M* = 60 months, range = 52-67 months). Each child with SLI had both expressive and receptive language impairment. Conversational language samples were obtained by using a variety of age-appropriate toys. Research assistants manually coded DSS scores, and MLU was obtained using the SALT software. Rice et al. (2006) reported a moderate correlation between MLU and DSS in the SLI group, $r = .56$, and a higher correlation between MLU and DSS in the control group, $r = .70$.

Relating MLU and Age

Many studies have reported the relationship between MLU and age with mixed results. Blake, Quartaro, and Onorati (1993) reported a significant correlation, $r = .70$, between MLU and age. Miller and Chapman (1981) conducted a study of 123 middle-to upper-middle-class Midwestern children, ages 1;5 - 4;11. Language samples were obtained while the children engaged in free play with their mothers. The study showed that the relationship between the variables of MLU and age were highly correlated, $r = .88$. However, despite the positive correlation between MLU and chronological age, same-age children have different MLUs (Miller & Chapman, 1981).

In contrast, Klee and Fitzgerald (1985) conducted a study to evaluate the grammatical performance and MLU of 18 typically developing children. They reported a low correlation between MLU and age, $r = .26$. This low correlation could be due to the homogeneity of the sample; subjects were chosen on the basis of restricted age, 2;1 to 3;1, as well as a restricted MLU range of 2.5-3.99.

Rice et al. (2006) also reported on the correlation between MLU and age in the first part of their study. The correlation between MLU and age was low for the SLI group, $r = .11$, indicating no association between MLU and age in the SLI group. However, the correlation between MLU and age was moderate for the control group, *r* = .51.

The second study that Rice et al. (2006) reported was longitudinal. Participants were the children who took part in the Rice et al. (1998) study of morphosyntax development. Participants included 38 children: 20 five-year olds with SLI and 18 children in the MLU-equivalent control group. Five years of language samples, at six-month intervals, were collected from each participant. MLU information was obtained from 205 conversational language samples. Throughout the five years, the two groups remained at comparable levels of MLU each time they were measured, indicating robustness in temporal stability of MLU matches. In general, MLU appears to be both reliable and valid as an index of general language development.

Goals of the Current Study

The present study examined the relation between manual DSS, automated DSS, MLU, and age. The following research questions were addressed:

- Are MLU and DSS developmentally sensitive in that each correlates with age?
- How high is the correlation between MLU and DSS, and could this correlation be an artifact of the fact that MLU and DSS are both correlated with age?
- Does DSS add information beyond MLU?

The present study also addressed how the answers to the three questions above change if the DSSA score is used instead of the manual DSS score.

Method

Participants

Conversational language samples previously collected from 30 (12 males and 18 females, 40% and 60%, respectively) children interacting with graduate student clinicians were used in this study. Language samples were collected from children living in a Brigham Young University family housing complex in Provo, Utah. Participants' ages ranged from 2;6 to 7;11. Three participants were included in each 6-month interval from 2;6 to 6;11 as well as three in the interval from 7;0 to 8;0. Parents of each child reported that the participants were typically developing, had no speech or language delay, and spoke English as their primary language. Also, each child passed a bilateral hearing screening at 15 dB HL. Each of the three graduate student clinicians collected a conversational language sample for one participant within every age interval. At least 200 intelligible utterances were collected in each sample. Neither adult utterances, nor child utterances containing one or more unintelligible words, were used in the sample. Student clinicians collected language samples in the participants' apartments using a variety of props to elicit conversation. These language samples were used in studies by Channell and Johnson (1999) and by Seal (2001).

Software

GramCats. The automated grammatical tagging software used was an updated version of the GramCats software evaluated and reported by Channell and Johnson (1999). The updated version determines and codes the grammatical category of words in running text by using information from two separate probability sources.

The first source determined the probability of a word being used as a particular part of speech, independent of context, by using its relative tag likelihood. An electronic dictionary built into the program contained the grammatical tag options and the relative frequencies of each tag option for each of over 20,000 English words, which had been automatically collected from manually tagged text. An unknown word was coded as a noun unless capitalized, in which case it was coded as a proper noun (Channell & Johnson, 1999).

The second source used was a probability matrix, which also had been automatically collected from manually tagged text. The matrix includes the frequency of each observed pair, divided by the second of those tags. Therefore, the information in the probability matrix is the probability of a tag coming after the prior tag (Channell & Johnson, 1999).

To test the accuracy of the GramCats program, Channell and Johnson (1999) compared automated grammatical tagging to manual tagging in conversational language samples. Thirty typically developing children ages 2;6 to 7;11 provided approximately 200 utterances. The average accuracy for tagging individual words was 95.1%. However, the accuracy of tagging entire utterances averaged 78%.

SALT. SALT (Miller and Chapman, 2000) includes a transcription editor, standard reports, and a reference database for comparison with typical peers. The SALT program can document language production in everyday speaking conditions by using a collection of representative language samples. The samples are transcribed and then compared to age or grade-matched typical speakers. SALT specifies areas of strength and weakness by calculating measures of syntax, semantics, rate, fluency, discourse, and errors. The profiles provided can help identify disordered language and can help SLPs develop language invention approaches. Additionally, SALT can compare performance in different sampling conditions to track change over time, in both primary and secondary languages. SALT can also compute a client's MLU in both words and morphemes (Miller and Chapman, 2000).

DSSA 2.0. Developmental Sentence Scoring Automated Version 2.0 (DSSA; Channell, 2016) is an updated version of the software that Judson (2006) used. Initially, the accuracy of this software was examined using language samples collected from participants including typical and language-impaired children. DSS was conducted, both manually and with DSSA, on 118 language samples obtained from 99 children between the ages of three and 11 (Judson, 2006). The manual coding was assumed to be accurate, and the accuracy of the DSSA software was determined from percent agreement between the manual coding and DSSA scores. The accuracy level of the single-child corpora was 82.7%, *SD* = 3.67, while the accuracy for the between-child corpora was 85.99%, *SD* = 5.05. Lower accuracy was found for children with language impairment (84%) and with language samples having lower manual DSS scores (Judson, 2006). Although the accuracy in each grammatical form category varied, the overall accuracy of automated DSS analysis in Judson's study was moderately high and was comparable to previous studies.

Procedure

Approximately 200 intelligible utterances were used from each language sample. GramCats was used to code the samples for grammatical category information. DSS was performed twice on each of the 30 samples, once manually and once using the DSSA Version 2.0 software. Manual interrater reliability was established by a second clinician analyzing 10% of the samples; the level of agreement was 97% (Seal, 2001).

Transcripts of speech samples from the 30 participants were manually coded according to SALT specifications. Data were entered into the SALT program and errors were manually corrected. Participants' MLU scores were obtained from SALT.

Analysis

Statistical analyses were conducted using SPSS version 23. Descriptive statistics were calculated for the study variables. MLU scores obtained from SALT were correlated with manual DSS scores, DSSA scores, and age. Partial correlations between MLU, DSS, and DSSA controlling for age were calculated. Partial correlations between DSS, DSSA, and age controlling for MLU were calculated.

Results

The present study addressed three research issues for both DSS and DSSA: the correlation of the measure with age, the correlation of the measure with MLU, and the developmental sensitivity of the measure beyond MLU. The ages and scores of all participants on each measure are listed in Appendix B.

Correlation With Age

The first focus of the present study was to look at MLU and DSS and determine if they were developmentally sensitive in that each measure correlated with age. Pearson product-moment correlations between age, MLU, and DSS are reported in Table 1. As Table 1 shows, MLU and DSS are both strongly correlated with age and these correlations are statistically significant, $r = .605$, $p < .001$ and $r = .723$, $p < .001$, respectively. As age increases, a child's MLU and DSS also tend to increase.

Table 1

Correlations Among Mean Length of Utterance (MLU), Manual Developmental Sentence Scoring (DSS), and Developmental Sentence Scoring Automated (DSSA)

	Age	MLU	DSS	DSSA
Age	-	.605**	.723**	.669**
MLU		-	.873**	.875**
DSS			-	.985**
DSSA				-

Note: MLU = mean length of utterance in morphemes; DSS = manual Developmental Sentence Scoring; DSSA = Developmental Sentence Scoring Automated Version 2.0 (DSSA; Channell, 2016).
** $p < .01$

When using DSSA instead of DSS to examine developmental sensitivity, the findings were almost identical. Correlations among age, MLU, and DSSA were also shown in Table 1. MLU and DSSA were both strongly correlated with age, and these correlations were statistically significant, $r = .605$, $p < .001$ and $r = .669$, $p < .001$, respectively. Thus, DSSA is developmentally sensitive as it also correlated with age, as did DSS.

Correlation of DSS and DSSA With MLU

The second research question considered in this study was the strength of the correlation between MLU and DSS and whether or not the correlation was due only to the fact that both correlated with age. As reported in Table 1, the correlation between MLU and DSS was high and statistically significant, $r = .873$, $p < .001$.

To address the second part of this research question, the partial correlation between MLU and DSS while controlling for age was calculated. The partial correlations are reported in Table 2. As seen in Table 2, when controlling for age, the partial correlation between MLU and DSS is still strong and statistically significant $r = .792$, $p < .001$. When the shared correlation with age was removed, the correlation between MLU and DSS decreased only slightly, dropping from $r = .873$ to $r = .792$. This suggests that the correlation between MLU and DSS is not simply an artifact of both measures being correlated with age.

Table 2

Partial Correlations Controlling for Age

	MLU	DSS	DSSA
MLU	-	.792**	.794**
DSS		-	.976**
DSSA			-

Note: MLU = mean length of utterance in morphemes; DSS = manual Developmental Sentence Scoring; DSSA = Developmental Sentence Scoring Automated Version 2.0 (DSSA; Channell, 2016).
** Significant $p < .01$

When using DSSA instead of DSS the findings were again almost identical. The correlation between MLU and DSSA was strong and statistically significant, $r = .875$, $p < .001$. To address whether this correlation was merely an artifact of the shared correlation of these variables with age, the partial correlation between MLU and DSSA while controlling for age was calculated. This partial correlation is reported in Table 2. When the shared correlation with age was removed, the correlation between MLU and DSSA decreased only slightly, dropping from $r = .875$ as reported in Table 1 to $r = .794$ as reported in Table 2. This suggests that the correlation between MLU and DSSA is not simply an artifact of both variables being correlated with age.

Information Beyond MLU

The third research question asked if DSS added information beyond MLU. If MLU were held constant, would DSS still correlate with age? To address this question, the partial correlation between DSS and age while controlling for MLU was calculated. The partial correlations are reported in Table 3. As can be seen in Table 3, when MLU is held constant, DSS still correlates with age. The relationship between DSS and age while controlling for MLU was moderate but still statistically significant $r = .501$, $p = .006$. Therefore, DSS appears to add information beyond MLU.

Table 3

Partial Correlations Controlling for MLU

	DSS	DSSA	Age
DSS	-	.936**	.501**
DSSA		-	.361
Age			-

Note: MLU = mean length of utterance in morphemes; DSS = manual Developmental Sentence Scoring; DSSA = Developmental Sentence Scoring Automated Version 2.0 (DSSA; Channell, 2016).
** Significant $p < .01$

The third research question also asks whether DSSA adds information beyond MLU. If MLU were held constant, would DSSA still correlate with age? To address this question, the partial correlation between DSSA and age while controlling for MLU was calculated. The partial correlations were reported in Table 3. When MLU is held constant, DSSA correlated less strongly with age than did the DSS. The relationship between DSSA and age while controlling for MLU was lower and not statistically significant, $r = .361$, $p = .055$. Therefore, DSSA may not add information beyond MLU like DSS does.

Discussion

The purpose of this study was to consider the following questions:

- Are DSS, DSSA, and MLU developmentally sensitive in that each correlates with age?
- How high is the correlation between DSS, DSSA, and MLU, and could the correlation between DSS, DSSA, and MLU be an artifact of the fact that DSS, DSSA, and MLU are each correlated with age?

• Do DSS and DSSA add information beyond that of MLU?

Results of the first research question showed that DSS and MLU are developmentally sensitive in that each measure correlates with age. There is a strong correlation of both DSS and MLU with age, and the correlations were statistically significant. Therefore, it appears that as age increases DSS and MLU also increase.

Findings were nearly identical when DSSA was used instead of DSS. MLU and DSSA are both strongly correlated with age and these correlations were statistically significant. Like DSS, DSSA also is developmentally sensitive as it also correlates with age.

The second research area addressed the correlation of the measure with MLU. The correlation between MLU and DSS was strong and statistically significant. There was only a slight decrease in correlation between MLU and DSS when the correlation with age was removed, suggesting that the correlation between MLU and DSS in not simply an artifact of both being correlated with age. Findings were also nearly identical when using DSSA instead of DSS in the second research area; the correlation between MLU and DSSA was strong and statistically significant. As with DSS, it appears that the correlation between MLU and DSSA is not simply an artifact of both variables being correlated with age.

To answer the third research question as to whether DSS adds information beyond that of MLU, the partial correlation between DSS and age while controlling for MLU was calculated. While controlling for MLU, the relationship between DSS and age was moderate but statistically significant, indicating that DSS appears to add information beyond MLU. To determine if DSSA adds information beyond MLU, the partial correlation between DSSA and age while controlling for MLU was calculated. There was a lower relationship between DSSA and age while controlling for MLU, and results were not statistically significant, indicating DSSA may not add information beyond MLU as does DSS.

Several studies have been completed that are similar to the current study. These include Long and Channell (2001), Channell (2003), and Judson (2006).

Long and Channell (2001) studied the accuracy of four automatic language analysis procedures obtained with the CP software. The four language analyses included MLU, LARSP, IPSyn, and DSS. In contrast, the present study performed language analysis with DSSA and focused on DSS and MLU scores. Long and Channell reported that time was a major factor stated by clinicians as a reason for not using language sample analysis and that software can compete with results produced manually. This study also concluded that the accuracy rates of DSSA are comparable to the accuracy rates of DSS. Therefore, both studies agree that the results of automatic language analysis are essentially equivalent to those from manual language analysis, and this approach could be a beneficial timesaver for clinicians.

Channell (2003) conducted a study to determine the accuracy of automated DSS analysis performed by the CP software. In addition to the overall data of the manually coded DSS, Channell reported the per-category and point-level levels of agreement, misses (false negatives), intrusions (false positives), and percentages of correct tagging. The goal of Channell's study was to provide a baseline for future software comparison, and to improve the use of automated DSS software by informing clinicians about its areas of strength and weakness.

The current study included information about automated DSS and MLU scores obtained with DSSA software. Utterances in Channell's (2003) study were DSS-coded, both manually and with the CP software. This study also included manually obtained DSS scores, but differed in that DSSA software was used for the automated DSS and MLU analysis. Channell studied the accuracy of the DSS software and listed the areas of the analysis that may have had more errors and therefore require more thorough scrutiny. The current study focused on the extent to which DSS, DSSA and MLU were developmentally sensitive, how high the correlations between DSS, DSSA, and MLU were, and whether DSS or DSSA added information beyond MLU.

Judson (2006) conducted a study on the accuracy of automated DDS entitled DSSA (Version 1.0). Like Judson, the current study also used previously collected language samples from typically developing children. Unlike Judson, this study implemented a newer version of automated DSSA: DSSA (Version 2.0). When compared to manual scores, DSS scores obtained via DSSA (Version 1.0) differed by less than one point, indicating that existing DSS norms may be applicable to DSSA (Version 1.0). Both the present study and Judson's study reported that accuracy levels of DSSA were sufficiently high to allow clinicians to use the automated analysis as an alternative to manual DSS scoring.

Limitations of the present study should be considered when interpreting the results. These limitations include the size and diversity of the sample population. Additionally, only samples from typically developing children were used in the present study. Furthermore, the limitations of DSS previously mentioned also apply to DSSA. Future studies may include a larger sample population obtained from more diverse backgrounds, as well as the inclusion of children with atypical development.

The use of DSSA has a clinical advantage over manual DSS due to its efficiency in analyzing language samples. Many authors have reported the importance of comprehensive language analysis in determining specific treatment goals for clients (Crystal, 1982; Fey, 1986; Lund & Duncan, 1993). As previously mentioned, clinicians do not often use DSS for language analysis due to the time required to learn and administer the DSS procedures. The timesaving factor of DSSA allows DSS to be performed with greater ease; DSSA software analysis takes less than two seconds to analyze 200 utterances. Results from this study indicate that clinicians may use DSSA with confidence because both DSS and DSSA were highly correlated. Therefore, DSSA offers the timesavings needed by clinicians with heavy caseloads to complete analysis of client language samples.

In summary, the present study provided information about the relationship among DSS, DSSA, and MLU, finding that MLU and DSS are each both strongly correlated with age, and that MLU and DSSA are also strongly correlated with age. The present study also found that the high correlation between MLU and DSS was independent of both measures being correlated with age. While there is room for improvement in DSSA, the present study corroborated earlier suggestions (e.g. Judson, 2006) that DSSA might be beneficial both clinically and in research settings, due its timesavings and its ability to achieve moderately high levels of accuracy. Additionally, the correlation between MLU and DSSA was high, and the correlation between these two variables was not simply an artifact of them both being correlated with age. Lastly, the relationship between DSS and age while controlling for MLU was moderate, but still statistically significant, suggesting that DSS appears to add information beyond MLU. These findings offer additional insight for the use of MLU, DSS, and DSSA in both clinical and research settings.

References

Blake, J., Quartaro, G., & Onorati, S. (1993). Evaluating quantitative measures of grammatical complexity in spontaneous speech samples. *Journal of Child Language, 20,* 139-152. doi.org/10.1017/S0305000900009168

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

Chabon, S. S., Kent-Udolf, L., & Egolf, D. B. (1982). The temporal reliability of Brown's Mean Length of Utterance (MLU-M) measure with post-stage V children. *Journal of Speech, Language, and Hearing Research, 25,* 124-128. doi:10.1044/jshr.2501.124

Channell, R. W. (1998). *GramCats*. (Version 1.0) [MS-DOS Computer program]. Provo, UT: Department of Audiology and Speech-Language Pathology, Brigham Young University.

Channell, R. W. (2003). Automated developmental sentence scoring using Computerized Profiling software. *American Journal of Speech-Language Pathology, 12,* 369-375. doi:10.1044/1058-0360(2003/082)

Channell, R. W. (2016). *DSSA*. Version 2.0 [Computer program]. Provo, UT: Department of Communication Disorders, Brigham Young University.

Channell, R. W., & Johnson, B. W. (1999). Automated grammatical tagging of child language samples. *Journal of Speech, Language, and Hearing Research, 42*, 727- 734. doi:10.1044/jslhr.4203.727

Crystal, D. (1982). *Profiling linguistic disability.* London, England: Edward Arnold.

Crystal, D., Garman, M., & Fletcher, P. (1989). *The grammatical analysis of language disability: A procedure for assessment and remediation* (2nd ed.). London, England: Cole and Whurr.

DeThorne, L. S., Johnson, B. W., Loeb, J. W. (2005). A closer look at MLU: What does it really measure? *Clinical Linguistics & Phonetics*, *19*, 635-648. doi:10.1080/02699200410001716165

Eisenberg, S. L., Fersko, T. M., & Lundgren, C. (2001). The use of MLU for identifying language impairment in preschool children: A review. *American Journal of Speech-Language Pathology, 10,* 323-342. doi:10.1044/1058-0360(2001/028)

Fey, M. E. (1986). *Language intervention with young children.* Boston, MA: College-Hill Press.

Fristoe, M. (1979). Developmental sentence analysis. In F. L. Darley (Ed.), *Evaluation of appraisal techniques in speech and language pathology* (pp. 15-17). Reading, MA: Addison Wesley.

Hickey, T. (1991). Mean length of utterance and the acquisition of Irish. *Journal of Child Language, 18*, 553-569. doi.org/10.1017/S0305000900011247

Hughes, D. L., Fey, M. E., Kertoy, M. K., & Nelson, N. W. (1994). Computer-assisted instruction for learning Developmental Sentence Scoring: An experimental comparison. *American Journal of Speech-Language Pathology, 3*, 89-95. doi:10.1044/1058-0360.0303.89

Hughes, D. L., Fey, M. E., & Long, S. H. (1992). Developmental Sentence Scoring: Still useful after all these years. *Topics in Language Disorders, 12*(2)*,* 1-12. doi.org/10.1097/00011363-199202000-00003

Hux, K., Morris-Friehe, M., & Sanger, D. D. (1993). Language sampling practices: A survey of nine states. *Language, Speech, and Hearing Services in Schools, 24,* 84-91. doi:10.1044/0161-1461.2402.84

Johnson, M. R., & Tomblin, J. B. (1975). The reliability of Developmental Sentence Scoring as a function of sample size. *Journal of Speech and Hearing Research, 18,* 372-380. doi:10.1044/jshr.1802.372

Judson, C. (2006). *Accuracy of automated developmental sentence scoring software* (Master's thesis). Retrieved from http://scholarsarchive.byu.edu/etd/788/

Kemp, K., & Klee, T. (1997). Clinical language sampling practices: Results of a survey of speech-language pathologists in the United States. *Child Language Teaching and Therapy, 13,* 61-176. doi:10.1177/026565909701300204

Klee, T. (1985). Clinical language sampling: Analyzing the analysis. *Child Language Teaching and Therapy, 1* (2), 182-198. doi:10.1177/026565908500100206

Klee, T., & Fitzgerald, M. D. (1985). The relation between grammatical development and mean length of utterance in morphemes. *Journal of Child Language, 12,* 251-269. doi:http://dx.doi.org/10.1017/S0305000900006437

Klee, T., & Sahlie, E. (1987). Review of Computerized Profiling. *Child Language Teaching and Therapy, 3*(1)*,* 87–93. doi:10.1177/026565908700300106

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

Leonard, L. B., Fey, M. E., Deevy, P., & Bredin-Oja, S. L. (2014). Input sources of third person singular –s inconsistency in children with and without specific language impairment. *Journal of Child Language, 42,* 786-820. doi.org/10.1017/S0305000914000397

Liles, B. Z., & Watt, J. H. (1984). On the meaning of "language delay." *Folia Phoniatric, 36,* 40-48. doi:10.1159/000265719

Lively, M. A. (1984). Developmental Sentence Scoring: Common scoring errors. *Language, Speech and Hearing Services in Schools, 15,* 154-168. doi:10.1044/01611461.1503.154

Long, S. H. (1986). Computerized Profiling (Version 1.0) [Computer software]. Arcata, CA: Author.

Long, S. H. (2001). About time: a comparison of computerized and manual procedures for grammatical and phonological analysis. *Clinical Linguistics & Phonetics, (15)*5, 399-426.

Long, S. H., & Channell, R. W. (2001). Accuracy of four language analysis procedures performed automatically. *American Journal of Speech-Language Pathology, 10*, 180-188. doi:10.1044/1058-0360(2001/017)

Lund, N. J., & Duncan, J. F., (1983). *Assessing children's language in naturalistic contexts.* Englewood Cliffs, NJ: Prentice-Hall.

MacWhinney, B. (2006). *CLAN Manual*. Retrieved from: http://childes.psy.cmu.edu/manuals/CLAN.pdf

Miller, J. F., & Chapman, R. S. (1981). The relation between age and mean length of utterance in morphemes. *Journal of Speech and Hearing Research, 24,* 154-161. doi:10.1044/jshr.2402.154

Miller, J. F., & Chapman, R. S. (2000). *Systematic Analysis of Language Transcripts (SALT*, Version 6.1) Windows; [Computer software]. Madison, WI: Language Analysis Laboratory, Waisman Center on Mental Retardation and Human Development.

Rice, M. L., Redmond, S. M., & Hoffman, L. (2006). Mean length of utterance in children with specific language impairment and in younger control children shows concurrent validity and stable and parallel growth trajectories. *Journal of Speech, Language, and Hearing Research, 49,* 793-808. doi:10.1044/1092-4388(2006/056)

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. doi:10.1044/jslhr.4106.1412

Rondal, J. A., Ghiotto, M., Bredart, S., & Bachelet, J. (1987). Age-relation, reliability and grammatical validity of measures of utterance length. *Journal of Child Language, 14,* 433-446. doi:org/10.1017/S0305000900010229

Salvia, J., & Ysseldyke, J. E. (2007). *Assessment in special and inclusive education* (p. 113). Boston, MA: Houghton Mifflin.

Scarborough, H. S. (1990). Index of Productive Syntax. *Applied Psycholinguistics, 11*, 1- 22. doi:org/10.1017/S0142716400008262

Seal, A. (2001). *Scoring sentences developmentally: An analog of developmental sentence scoring.* (Master's thesis) Brigham Young University, Provo, UT. Retrieved from http://scholarsarchive.byu.edu/etd/1141/

Smith, J. M., DeThorne, L. S., Logan, A. R. L., Channell, R. W., & Petrill, S. A. (2014). Impact of prematurity on language skills at school age. *Journal of Speech, Language, and Hearing Research, 57,* 901-916. doi:10.1044/1092-4388(2013/12-0347)

Turnbull, K. L. P., & Justice, L. M. (2012). *Language development from theory to practice.* Upper Saddle River, NJ: Pearson Education.

Westerveld, M. F., & Claessen, M. (2014). Clinician survey of language sampling practices in Australia. *International Journal of Speech-Language Pathology, 16*, 242-249. doi: 10.3109/17549507.2013.871336

Appendix A: Annotated Bibliography

Channell, R. W. (2003). Automated Developmental Sentence Scoring using Computerized Profiling software. *American Journal of Speech-Language Pathology, 12,* 369-375. doi:10.1044/1058-0360(2003/082)

Focus: The quantification of expressive syntax development has often been accomplished using Developmental Sentence Scoring (DSS). A transcript can be entered as a computer text file and be assessed with DSS. The purpose of this study was to determine the accuracy of automated DSS analysis and to decide which parts require increased correction due to more errors.

Method: Thirty samples containing a total of 6,891 utterances were collected from children in the Reno, Nevada, area as part of the Fujiki, Brinton, and Sonnenberg (1990) study. Ten children from each of three groups provided the samples. Ten children with language impairment were matched to typically developing children that were language score similar, as well as to 10 children who were typically developing and similar in chronological age. Eighteen additional samples were obtained from children in the Jordan School District in Salt Lake County, Utah. These samples provided an additional 2,193 utterances.

Procedure: The DSS techniques specified by Lee (1974) were used to manually score the utterances in each sample. Computerized Profiling (CP) was also used to score each sample for DSS.

Results: The mean accuracy rate for all samples was 78.2% (SD = 4.4). There was a high correlation (r = .97, p < .0001) between the manual and CP-computed scores. The CP-computed scores were consistently higher than the manually computed scores, and the difference was statistically significant (p < .0001).

Discussion: Hughes et al. (1994) suggested 80% as an acceptable level of skill for effective clinical use of DSS; the observed accuracy of analysis for the samples in this study averaged just below that suggested level. The program made two types of errors. First, misses (false negatives), where the manual analysis indicated an utterance as having an item that the program didn't. Second, intrusions (false positives), where the program coded a cell that wasn't in the manual analysis. Approximately 3% of utterances were completely omitted by the CP analysis; this accounted for many of the misses noted. Sample size might be one possibility why the accuracy of automated DSS is less than outstanding; accuracy was lower in samples with fewer utterances. Also, it is possible that the software had a higher rate of accuracy on samples from children who used more advanced grammatical forms. However, the relationship between the number of utterances in samples and the manually calculated scores had an even stronger relationship. The use of partial correlations to remove some of the redundant information in these three variables demonstrated that the relationship between the developmental score and the number of utterances continued to be strong. It was recommended that clinicians continue to check and correct the program's output due to the current level of accuracy of CP's automated DSS analysis.

Relevance to my study: This study focused on the accuracy of automated DSS analysis, which is the basis of one aspect of my study. The accuracy of the CP software can be compared with the accuracy of the software I am analyzing. Also, my study will use the same basic procedure.

Long, S. H., & Channell, R. W. (2001). Accuracy of four language analysis procedures performed automatically. *American Journal of Speech-Language Pathology, 10*(2), 180-188. Retrieved from http://search.proquest.com/docview/204264448

Focus: Clinical use of language sample analysis has been recommended for speech-language pathologist for at least 40 years. Several comprehensive procedures for language grammar analysis were developed in the 1970s. However, two surveys conducted in the 1990s discovered that half of the speech-language pathologists working in preschool and school settings reported that the extensive time needed to analyze language samples led to the procedure being used infrequently. Computer analysis of language samples is fast but its accuracy has been studied much. The purpose of this study is to explore the accuracy of automatic language analysis.

Method: *Child language samples*: Sixty-nine conversational language samples from four sources were used for analysis. These language samples represented a range of ages, national dialects, levels of linguistic development, and diagnostic categories. These diversities imitated the range of challenges existing in clinical language analysis. Typically developing American children provided 30 samples. Australian children playing with their mothers in clinical environments provide 17 samples. Twelve Canadian children diagnosed with Specific Language Impairment also provided language samples. American children who were diagnosed with Specific Expressive Language Impairment provided 10 samples.

Computerized analysis: Relevant modules of Computerized Profiling (CP) were used for all computerized language analysis. Four language analyses were performed with two conditions. First, all coding and tabulation was done by CP. Second, codes generated by CP were reviewed by two judges. The mean length of utterance (MLU) was computed from complete and intelligible utterances. Language Assessment, Remediation and Screening Procedure (LARSP) codes were generated for each utterance in all samples. Developmental Sentence Scoring (DSS) scores were calculated from utterances with a subject-predicate structure in each sample. Index of Productive Syntax (IPSyn) was created on the identification of syntactic types in each sample.

Results: Automatic and corrected summary scores were very similar. MLU had the highest degree of accuracy, 99.4%, across all groups. IPSyn and DSS had degrees of accuracy of 95.8% and 89.8% respectively.

Discussion: A comparison of the accuracy findings in this study with the standards for coding reliability in child language research can be used in assessing the usefulness of automatic language analysis. The evaluation of interrater reliability has no absolute standards. However, it was suggested that levels of agreement between coders in children language research is deemed acceptable if they are greater that .85, good if they are greater than .90 and excellent if they are greater than .95. Using this standard, the analyses generated automatically in this study showed a range of reliability when compared to corrected analysis. LARSP showed an acceptable level of agreement, IPSyn and DSS showed a good level of agreement and MLU showed an excellent level of agreement.

Relevance to my study: This article discussed the accuracy of automated language analyses including DSS and MLU, which are part of my research focus. Results indicated that automated DSS procedures need further improvement to reach an acceptable level of accuracy. My study includes the use of a newly developed automated DSS software program, and will determine whether or not it has increased accuracy of DSS analysis.

Long, S. H. (2001). About time: Comparison of computerized and manual procedures for grammatical and phonological analysis. *Clinical Linguistics & Phonetics, 15*, 399-426. doi:10.1080/02699200010027778

Focus: The American Speech-Language-Hearing Association identified "organizing and time management" as one of the nine skills that graduate students in speech-language pathology need to learn. The focus of this paper was to determine if computerization of language analysis samples would reduce the time of analysis, which could make the procedures more clinically manageable. Most authors agree that language sample analysis is time consuming. Language sample analysis requires several tasks which include recording the conversation, transcription, analysis, and interpretation. The time required to examine a variety of phonological and grammatical analyses was examined on samples that varied in length and complexity.

Method: *Participants:* Two hundred fifty-six students and practicing speech-language pathologists (SLP) from the United States and Australia participated in this study. Each SLP had received university-level instruction, varying from two months to eleven years, on the analysis procedures prior to participation. Participants chose the type and number of analyses they performed and were asked to choose only analyses they were familiar and comfortable performing.

Language Samples: All language samples were obtained during conversational interactions. The phonological analyses were mostly broad phonetic transcriptions obtained from three samples. Three samples were also used to perform grammatical analyses and were typed according to conventional orthographies.

Manual analysis procedures: Each participant was given the printed transcript, an instruction packet, a set of forms to use while recording and tabulating during analysis, and a time log for each manually analyzed sample. Starting and stopping times were recorded to the nearest minute for each analysis.

Computer analysis procedures: Prior to the analysis done for this study at least one complete analysis was performed by each participant. Appropriate modules of Computerized Profiling were used to perform all language analyses. Time logs were used to record starting and stopping times.

Phonological analysis: Ten phonological analyses were performed including: type-token ratio, variability analysis, homonymy analysis, word shape analysis, vowel inventory, consonant inventory, vowel target analysis, consonant target analysis, percentage consonant correct, and phonological process analysis. Each analysis was timed separately.

Grammatical analysis: The following five analyses were performed: MLU, number of syntactic types, LARSP, the Developmental Sentence Score, and the Index of Productive Syntax.

Order of analyses: To allow for direct comparison of manual and computer times, each participant analyzed the same transcript of every language sample twice. Each transcript was analyzed once by hand, and once by computer. It was anticipated that the computer analysis would be more time efficient. To bias the study against this, the computer analyses were always performed first. This ensured that any advantage obtained by previous exposure to the sample would act as a means of decreasing manual analysis times.

Results: *Accuracy of analyses:* This study provided a well-defined picture of the relative accuracy of manual and computerized analysis. The computerized procedure received 8.8 of ten accuracy points for phonological analysis averaged across six participants. The computerized procedure received 4.7 of five possible points averaged across 30 participants.

Efficiency of analyses: Although always time consuming, the length of time needed for a comprehensive phonological analysis varied greatly according to the type of sample being analyzed. Samples ranged from just over three hours to nearly 10 hours to analyze by hand. Computerized analyses were completed faster than manual analyses in each of the 136 analyses performed.

Discussion: The direct measure of time savings was not a focus of this study, however, the efficiency of performing productivity analyses on the computer appears beyond question. The results of this study indicated that language analysis done by hand will not be regularly possible in most clinical schedules due to the time required.

Relevance: This study reports the difficulty that speech-language pathologists have in analyzing language samples by hand due to the amount of time required. Results indicated that automated DSS procedures need additional improvement to reach an acceptable level of accuracy. The current study assesses the accuracy of fully automated DSS analysis and whether it has increased to an acceptable level of accuracy due to improvements in the new software.

Kemp, K., & Klee, T. (1997). Clinical language sampling practices: Results of a survey of speech-language pathologists in the United States. *Child Language Teaching and Therapy, 13,* 166-176. doi:10.1177/026565909701300204

Introduction: Naturally-occurring conversational speech has been used for many years by clinicians and teachers to assess children's language abilities. Clinical language sample analysis provides the opportunity to study a child's linguistic system during communicative interaction. This article reported the results of a survey whose purpose was to determine how speech-language pathologists use clinical language sample analyses, and to find which problems occur during their use. The authors considered previous studies which surveyed clinical practices with the objective of concluding whether any changes in clinical practice have occurred since the last clinician survey. No research on this topic had been done based on a national random sample of speech-language pathologies in the United States.

Method: Surveys were sent throughout the United States to 500 randomly selected speech-language pathologist from 3952 preschool settings. All participants were listed in the current ASHA directory. Each of the 253 respondents reported holding the Certificate of Clinical Competence in Speech-Language Pathology. The median caseload size reported by the clinicians was 26, and each clinician reported that they primarily worked with pre-school children who have language disorders.

Results: Assessment by means of language sample analysis was reported by 85% of those surveyed, and 97% reported the use of standardized tests. It appeared that the clinicians used language sample by choice, because only 8% of those surveyed reported that language samples were state mandated. Nearly all (92%) of the respondents stated that language sample analysis was used for diagnostic purposes. Other reasons for using language sample analysis were intervention (77%), post-intervention (64%) and screening (44%). Fifteen percent of respondents reported not using language sample analysis. The most common reasons given for not using language sample analysis were lack of time (86%), lack of computer resources (40%), lack of training and expertise (16%) and financial constraints (15%). Ninety-five percent of respondents reported transcribing their own language samples. Two-thirds of the clinicians reported that they would send recorded samples to a lab for transcription if it were available and affordable. Nearly half (48%) of the clinicians in the survey preferred non-standardized forms of language sample analysis. DSS was reported as the standardized procedure used most often; 35% of respondents reported using it. Lahey's Content/Form/Use analysis was used by 29% of those surveyed, and Assigning Structural Stage was used by 17% of the respondents.

Discussion: Participants for this survey were sampled randomly and anonymously. Just over half of the questionnaires distributed were returned. Also, every geographic region of the continental United States was represented in this survey. Therefore, the authors are reasonably confident that the results of their survey are representative of the views of ASHA-certified speech-language pathologists who work with pre-school age children. Eighty-five percent of respondents reported using some form of language sample for clinical assessment, although they were not legally obligated to do so. This suggests that most clinicians consider language sample analysis important during clinical assessment. It was noted in this survey that language sample analysis is usually done by hand from transcription analysis. Only 8% of clinicians reported using computer-assisted language sample analysis.

Relevance to my study: This article indicated that DSS was the most commonly used standardized form of analysis among clinicians working with pre-school children. Although clinicians view language sample analysis as important, they often do not have time to use these analyses due to the time constraints of heavy caseloads. My study is evaluating the possibility of reducing the amount of time needed for language sample analysis by using automated DSS.

Channell, R. W., & Johnson, B.W. (1999). Automated grammatical tagging of child language samples. *Journal of Speech, Language, and Hearing Research*, *42*, 727-734. doi:10.1044/jslhr.4203.727

Introduction: “Tagging,” or the automated grammatical categorization of words, has been reported in recent studies to have significant levels of accuracy of agreement with manual tagging from words used in a variety of texts. The purpose of this study was to examine the accuracy of a computer program in automatically tagging transcriptions of children’s spoken language.

Methods: Conversational language samples previously collected from 30 typically developing children interacting with graduate students were used in this study. Ages ranged from 2;6 to 7;11 (years;months) with ages being spread evenly across the age continuum. Approximately 200 intelligible utterances were in each sample. Neither adult utterances nor child utterances containing one or more unintelligible words were tagged. The automated grammatical tagging software used was GramCats, which determines the grammatical category words in running text by using information from two separate probability sources. The first source is an electronic dictionary used for relative tag likelihood information. The second source is a probability matrix used for tag transition likelihood information.

Procedure: The first author manually tagged the language samples. This study used 75 word-level grammatical tags. Each sample was also tagged using the GramCats software, and was compared with the manually tagged version of that sample on both a word-by-word and utterance-by-utterance basis.

Results: The accuracy rates for automated grammatical tagging yielded word-by-word accuracy rates that ranged from 92.9% to 97.4% (M = 95.1%, SD = 1.2%). For an utterance-by-utterance agreement each automated tag must agree with each manual tag in an utterance; the utterance agreement ranged from 60.5% to 90.3% (M = 77.7%, SD = 7.9).

Discussion: An overall accuracy of 95% for word-by-word grammatical category tagging has been reported by previous studies using probabilistic methods (Church, 1998; DeRose 1988). A similar level of overall accuracy for word-by-word tagging of the naturalistic language of normally developing preschool and younger school-age children was found in this study. However, this study suggested that further research of probabilistic grammatical analysis was warranted due to the extension of findings from edited, adult written text to naturalistic child language samples. A next logical step would be the evaluation of accuracy in tagging language samples from children in which language impairment has been identified. Whole-utterance tagging accuracy was lower than the word-by-word tagging accuracy, suggesting that additional improvement is required to obtain automated analysis tagging of utterances that will avoid the need for manual post editing.

Relevance to my study: The results of this study reported that the reliability of automated grammatical tagging was high (95%) indicating that automated grammatical tagging software has the potential to achieve levels of reliability similar to human analysts. Also, GramCats is a component of the DSSA 2.0 software that will be used in my study.

Gavin, W. J., & Giles, L. (1996). Sample size effect on temporal reliability of language sample measures of preschool children. *Journal of Speech and Hearing Research, 39,* 1258-1263. doi:10.1044/jshr.3906.1258

Introduction: This study provides information about the temporal reliability of four quantitative language sample measures: total number of words (TNW), number of different words (NDW), mean length of utterance in morphemes (MLU-m), and mean syntactic length (MSL). The validity and reliability of these measures must be determined empirically if language samples are to be used diagnostically.

Method: Twenty children, who had passed a pure-tone audiometric screening, were used in this study. The children, 15 males and 5 females, ranged from 31 to 46 months of age. A language laboratory designed as a playroom was used to conduct all evaluations. Each child was tested twice with the two sessions occurring at the same time of day. The participants received a hearing screening and tympanometry, Form L of the Peabody Picture Vocabulary Test-Revised (PPVT-R), and the Reynell Developmental Language Scales. The assessment session also included a 20-minute parent-child free play used to obtain an audio-recorded language sample. No observers were in the room while the recordings were being made, and the same toys were provided at each session. Four trained undergraduate students, blind to the purpose of the study, transcribed the language samples directly from audiotape into computer files. The systematic Analysis of Language Transcripts (SALT) was used to transcribe both the caregiver and child utterances. Inter-transcriber reliability was assessed by a trained graduate student who transcribed a randomly selected 5-minute segment of each language sample. Two types of sample sizes were used: (a) time based (12 or 20 minutes), and (b) utterance based (consisting of 25-175 intelligible and complete utterances).

Results: The size of the language sample was the dependent factor of the temporal reliability coefficients. The temporal reliability of the TNW was found to be inadequate. MLU-m and MSL exceeded the minimum r criterion of .71 in both timed samples. TNW did not meet this criterion at either time based sample. NDW only met the criterion in the 20-minute sample. Minimally acceptable temporal reliabilities for NDW, MLU-m, and MSL were indicated for samples 20-minutes in length. In samples greater than or equal to 100 complete and intelligible utterances reliability generally increased with all four measures exceeding the *r* criterion of .71. A more persuasive diagnostic criterion of a coefficient greater than .90 wasn't reached until the sample size reached 175 complete and intelligible utterances for each of the four measures.

Conclusion: The language sample measures of NDW, MLU-m, and MSL for children in this age range, if they are obtained from parent-child conversations of at least 175 utterances, have sufficiently high temporal reliability for both diagnostic and research tasks. Smaller sample sizes, with lower reliability levels, may be adequate for use by clinicians to track a client's progress during intervention.

Relevance to the current study: Gavin and Giles discuss the reliability of MLU when used diagnostically and in research. SALT software was used in this study to compute language production; SALT was used in my study in the same way.

Hughes, D. L., Fey, M. E., Kertoy, M. K., & Nelson, N. W. (1994). Computer-assisted instruction for learning Developmental Sentence Scoring: An experimental comparison. *American Journal of Speech-Language Pathology, 3,* 89-95. doi:10.1044/1058-0360.0303.89

Introduction: The purpose of this study was to use two methods as a means of comparing students' learning of the Developmental Sentence Scoring (DSS) procedure. The methods used were classroom-based tutorial (CBT) and computer-assisted instruction (CAI). One major advantage of CAI is the degree of flexibility it allows students. When using CAI, students can complete exercises at their own convenience. CAI is also advantageous to instructors because it frees up time that would be used for instruction and exam grading. Practicing clinicians can use CAI to learn new techniques and reinforce old ones. Generally, instructors teach DSS by having their students read important information about the technique. Students also listen to in-class lectures and complete practice exercises. The authors collectively agreed that teaching DSS took valuable time away from discussing other analysis matters, as well as other important subjects dealing with childhood language disorders.

Method: Fifty-five speech-language pathology students participated in this study, which took place over an 8-week period. The participants were selected from speech-language programs located in Michigan, Kansas, and in Ontario, Canada. Prior to the study none of these students had scored sentences using the procedure. During the first week each of the participants were assigned to read chapter 4 in Lee's (1974) original text. They also attended a 2-hour introductory lecture taught by the co-authors of this report. The introductory lecture contained three parts: first, an explanation of Lee's rules for transcribing and segmenting utterances; second, definitions and examples of each of the eight categories; third, criteria for assigning sentence points. The participants were then randomly assigned to one of two treatment groups which consisted of CBT or CAI. Two samples, designated Quiz 1 and Quiz 2 were used to determine the effectiveness of the CBT and CAI for teaching the DSS procedure used for language sample analysis. Fifty consecutive segmented utterances from the same child were used as the two samples. At each of the three sites, half of the students received Quiz 1 as the pre-test and Quiz 2 as the post-test. The order was reversed for the other half of the students.

Results: There were no significant differences between Quiz 1 and Quiz 2 as evidenced by the *t*-tests for independent samples ($t < -1.8$, $p > 0.09$). A two-way repeated measures of analysis of variance was used to measure participants' performance. The between-subjects factor was instruction and the within factor was pre- and post-tests. The number correct out of 212 was the dependent measure. The two methods did not differ in their effects on student performance as evidenced by the lack of main effect for instruction method, $F(1,53) = .69$, $p = .41$. The participants' post-test scores were classified into percent correct ranges to provide an overview of the participants' performance. It was reported that 93% of the participants in this study scored an accuracy of 80% or greater. This indicated an acceptable level of skill for effectively using DSS in a clinical setting.

Discussion: There was no significant difference in the results of the participants learning to use a language analysis procedure through CAI versus traditional CBT. Despite their pre-test scoring abilities, the participants obtained near ceiling levels of scoring after receiving instruction about DSS. Possible advantages of using CAI include automatic and immediate feedback provided to students. Also, CAI is more convenient for students, allowing them to practice at their own pace and when it is an opportune time for them. Additionally, instructors using CAI will have significant time savings.

Relevance to the current study: This article discusses two methods of learning DSS, and indicated that both are effective ways of learning how to perform DSS analysis. Clinicians can reach acceptable levels of scoring accuracy with extensive practice. Hughes et al. suggest an accuracy rate of 80% as an acceptable level for clinical use. Therefore, the program used in my study should, at a minimum, reach this level of accuracy.

Johnson, M. R., & Tomblin, J. B. (1975). The reliability of Developmental Sentence Scoring as a function of sample size. *Journal of Speech and Hearing Research, 18,* 372-380. doi:10.1044/jshr.1802.372

Introduction: For many years statisticians have known that as the sample size increased, the reliability of that measure also increases. Developmental Sentence Scoring (DSS) requires a language sample of at least 50 utterances. Knowing what the sample size must be to achieve a particular level of reliability is important for clinicians using any quantitative measure of spontaneous language such as DSS. The purpose of this study was to provide information regarding the reliability of DSS with sample sizes larger and smaller than 50 utterances.

Procedure: Fifty children were selected from the University of Iowa Institute for Child Development Preschool. Their ages ranged from 4;8 (years;months) to 5;8. There were two criteria for subject selection. First, the children had to be monolingual and second, they had to have normal hearing. The stimuli selected to elicit language samples were chosen because they were of interest to children in that age range. The stimuli consisted of two sets of questions, two types of picture stimuli and a variety common household tools. The language samples were obtained individually with only the child and experimenter present. Although the overall schedule was structured, the atmosphere of the sessions was casual and conversational. The five tasks were given in a randomized order to each subject. Utterances were recorded, beginning with the presentation of the first task, until 60 acceptable complete sentences were obtained. As described by Lee and Canter (1971) the experimenter transcribed the first 50 complete, consecutive, different, and intelligible sentences. After transcription, 25 sentences were randomly selected from the body of the 50 sentences. Next, DSS scoring was done on each of the 25 sentence samples after which each sentence was divided into segments of five sentences. Each unit of five sentences was considered one response segment.

Results: An analysis of variance was used to find estimates of reliability for the measures from score values. First, a response segments by subjects analysis of variance was performed for each measure's score values. Reliability estimates were then obtained from the mean squares provided by each analysis. Reliabilities were estimated from sample sizes of 5 to 250 sentences. As the sample size increased, the estimated reliability values increased for all scoring categories.

Discussion: The standard error of measurement may have greater importance to clinicians because this information has more usefulness for interpreting individual scores due to its expression in score points rather that relative terms. Determining the appropriate sample size for DSS may be aided by use of the standard error of measurement, which has norms given in percentile. According to Lee and Carter (1971) intervention is needed for any DSS score below the 10th percentile. DSS does not separate children with a language disorder from typically speaking children. Instead, it is used as a measure to isolate specific areas of difficulty for a language user. The age range and stimulus materials used limit the results of this study. Furthermore, a study completed by Lee and Koenigsknecht (1971) indicated a propensity for reliabilities of all DSS measures to increase with age, which indicates that reliabilities at various ages may be different. Estimated data on reliability and standard error of measurement from this study may be used as a general guide for other age groups and different stimulus materials.

Relevance to the current study: This article provided information regarding the reliability of DSS analysis with sample sizes larger and smaller than 50 utterances. While DSS is a valuable tool, it is important to keep in mind its limitations. My study included some samples that were less that the 175 sentences recommended by Johnson and Tomblin. Therefore, there is a possibility that the DSS scores do not completely represent each child's true ability. Also, it is important to keep in mind that DSS scores should not be the only factor used in making clinical decisions.

Hughes, D. L., Fey, M. E., & Long, S. H. (1992). Developmental Sentence Scoring: Still useful after all these years. *Topics in Language Disorders, 12*(2), 1-12. doi.org/10.1097/00011363-199202000-00003

Introduction: Developmental Sentence Scoring (DSS) is a measure of spoken syntax used with children who speak Standard American English (SAE). Initially, DSS comprised scores based on the following eight grammatical categories: (1) indefinite pronoun/noun modifiers, (2) personal pronouns, (3) main verbs, (4) secondary verbs, (5) negatives, (6) conjunctions, (7) wh-questions, and (8) interrogative reversals. In addition to scores awarded to any of these eight categories, a sentence point (SP) was given for sentences that were grammatically and semantically correct. In 1974 Lee revised the DSS system and assigned a developmental value for each category; these values ranged from 1-8. Fifty sentences were averaged for the DSS score. Guidelines, which were available for clinicians, were provided by 200 children from whom data were obtained. Twenty children for each 6-month interval from 2;0 to 6;11 (years;months) were used for data collection. The standard criteria of 100 subjects per age group was not met. However, the data are helpful in provided an approximation of a child's functioning when compared to that of other children.

Why DSS?: There are three factors that make DSS a valuable clinical and research tool. First, DSS is a numeric variable. This variable can be compared with previous and later scores from the same child. DSS can also be compared to the scores of other children. Second, DSS can aid the clinician in making diagnostic judgments because it provides some developmental data. Third, DSS can be used as a method of organization when asking and answering clinical questions. The drawback of these features is that a specific score can have various meanings. For example, two children may greatly differ in their abilities yet achieve the same score. Also, a child may receive many sentence points despite his or her simple yet grammatically correct sentences. Later, the same child may be awarded fewer points for a more complex sentence because it contains errors.

Some Uses of DSS: First, DSS can be helpful in making diagnostic judgments. Often clinicians only report the Mean Length of Utterance (MLU) for a language sample. However, information obtained from DSS when used in conjunction with MLU provides valuable quantitative support based on qualitative analyses which can be used in making clinical judgments. A second use of DSS is goal selection and treatment planning. DSS can be useful for goal selection and treatment planning in the following ways:

1. A clinician may select a grammatical target by noting the frequency of the attempt marks for each category. This will inform the clinician of the grammatical targets the child is attempting to produce but is doing so incorrectly.

2. Point values can help the clinician choose a selection of a group of forms that are more developmentally complex. This may be needed when many low-scoring forms are correctly produced while higher scoring forms are infrequent.

3. Treatment goals can be chosen by analyzing sentence point errors, which may reveal error patterns.

4. Examining the regularity with which errors occur in each category may lead a clinician to bring about infrequently used forms in a child's language sample. Essentially, a DSS can lead

to a hypothesis about the nature of the child's impairment when grammatical forms present, absent, infrequently used or produced in error are analyzed. This information can also aid in goal selection and therapy planning.

Some difficulties with DSS: Some rules within DSS that are either counter-intuitive or likely to result in undue skewing. The measurement of a child's level of grammatical development is the principal goal of DSS. Some rules are inconsistent with that goal.

Relevance to the current study: This article explains why DSS is still a useful clinical and research tool. My study will evaluate a program for computerized DSS analysis, which could potentially decrease the time needed for clinicians to perform DSS analysis.

Lively, M. A. (1984). Developmental Sentence Scoring: Common scoring errors. *Language, Speech and Hearing Services in Schools, 15,* 154-168. doi:10.1044/0161-1461.1503.154

Introduction: Developmental Sentence Scoring (DSS) is a popular and commonly used method of analyzing preschool children's morphologic and syntactic development. DSS was developed by Laura Lee and her colleagues at Northwestern University as a means of quantifying the grammatical structure of young children's expressive language. Additionally, DSS can aid in determining intervention goals and in evaluation of children's process during intervention. Significant study and actual practice is needed to learn to score language samples accurately. Although most improve rapidly, most graduate student clinicians have difficulty when first learning DSS scoring; additional practice, close supervision and instructional feedback aid in the improvement of DSS scoring. This article was written in an attempt to assist clinicians by identifying common problems and scoring errors when learning the DSS procedure.

The author described 10 common problem areas and scoring errors:

1. Determining an appropriate 50-response language sample: DSS instructions state that a sample should contain 50 different utterances. Often student clinicians will record the same utterance more than one time. Also, only complete sentences should be used as samples. A complete sentence is defined by Lee (1974) as one which contains a noun/pronoun and verb in

subject-predicate relationship. Student clinicians often include phrases known as “presentences” which are often uttered by preschool language-disordered children. Presentences omit copular forms of *to be* and main verbs comprising *have*.

2. Awarding the sentence point: Only completely grammatically and semantically correct sentences, are awarded points. Errors are often made by student clinicians as to whether or not to a sentence should receive a point.

3. Using attempt marks and incomplete designations: An “attempt mark” is given to utterances attempted that are not grammatically incorrect or semantically incorrect. Incorrect utterances are often given a score by student clinicians instead of an attempt mark. An “incomplete” is awarded to utterances which are incomplete on the surface level but are conversationally appropriate.

4. Indefinite pronouns and noun modifiers: Persons learning DSS frequently make two types of errors in this category. First, and most frequent, is when words listed on the DSS protocol in the Indefinite Pronoun/ Noun Modifier category are scored in error because they are functioning as adverbs which are not scored in the DSS.

5. Main verbs: According to the author’s experience, errors in scoring main verbs occur significantly more than any other category. Students learning DSS are strongly encouraged to carefully study the main verbs section of the DSS scoring instructions.

6. Secondary verbs: Often, the reason for errors in this category is that the scorer does not notice that a secondary verb is present. Errors often occur when the clinician fails to recognize the infinite marker *to*.

7. Negatives: Misunderstanding of what scores a 1 and what scores a 7 is the most common error in this category. Rarely to errors involve *can’t* and *don’t* which are always scored a 4, or *isn’t* and *won’t* which are always scored a 5.

8. Conjunctions: The most common errors in this category are failure to score conjunctions which begin sentences if they begin an independent clause, confusion between wh-conjunctions and wh-pronouns, and mistakes related to the rules for dividing sentences which contain multiple *ands*.

9. Interrogative reversals: Generally, the most common error in this category happens when wh-questions are scored and student clinicians forget to score the subject-verb inversion (interrogative reversal). Generally, yes/no questions are scored correctly.

10. Wh-questions: This category generally does not pose much difficulty to student clinicians.

Conclusion: Much time and effort is needed to improve accuracy of the DSS procedure. Given practice and experience, most clinicians can rapidly improve their productivity and accuracy.

Relevance to the current study: All of the samples in my study were manually and automatedly scored for DSS. Lively reports the many errors which many DSS learners make, which I should consider in the manual scoring of my samples. Also, if a fully-automated DSS software program had a sufficient level of accuracy, the frequency of human errors would be reduced.

Long, S. H., & Masterson, J. J. (1993). Computer technology: Use in language analysis. *ASHA, 35(9),* 40-51.

What is CLA software?: Computerized Language Analysis (CLA) is divided into two groups. First, computerized phonological analysis (CPA) which are programs to perform phonological analyses of phonetic transcription data. Second, language sample analysis (LSA) which yields semantic, syntactic, or pragmatic analyses of written transcripts. Analyses for all CLA programs are built on a particular model of language structure. For example, most LSA programs differentiate intentional versus unintentional speech. Unintentional speech segments are known as "mazes." Theses mazes contain revisions, filled pauses, and repetitions, which reflect difficulty in language construction. Furthermore, for proper analysis clinicians using CPA must fully understand the program's model of phonological structure as well as how to correctly enter data.

What CLA can software do?: CLA can be helpful in planning intervention and evaluating clients who have various types of language disorders. Whether completed by hand or by computer, LSA produces criterion-referenced results, which can aid in determining skills to target during intervention. CLA aids in forming the basis of ethical language intervention by informing the clinician about the client's patterns of learning, competencies and areas of deficit.

Clinicians can complete language sample analyses faster using CLA than by hand. Articulation tests are analyzed quickly by CLA software. Programs with phonetic dictionaries can analyze connected speech quickly. Otherwise, connected speech takes more time to analyze. Time constraints of clinicians performing detailed analyses by hand are overcome when using CLA. For example, most clinicians learn to calculate language indices and construct linguistic profiles for their clients as part of their professional education. Due to the time that these procedures require, they are rarely used by practicing clinicians. This is not the case when using CLA.

What CLA software cannot do: Using a computer for language analysis does not ensure that the results are correct. For example, CLA cannot assist in the orthographic or the phonetic transcription of a client's language. Also, data incorrectly entered into a computer will result in an inaccurate analysis. In general, language analyses are too complex to be conducted exclusively by a computer algorithm. Therefore, it is important to keep in mind that CLA does not produce indisputable accuracy. Human input is important to ensure the yielded results are valid. Valid results are accomplished in two ways. First, provide the software more linguistic information with which to work. This is done through advanced coding of the transcripts prior to submitting them for analysis. Second, the clinician is asked to approve the computer's decision after transcripts are tentatively coded. Incorrect codes from the computer must be changed. This is especially valuable during complex semantic or syntactic analysis which require many individual decisions. For example, an analysis completed by LARSP may require more than 40 codes per utterance. Much time can be saved if CLA can correctly generate 75% of these codes. However, the human user is responsible for linguistic judgments. The human user must be capable of generating the same analysis by hand in order for CLA to produce an accurate analysis. Currently, CLA is limited to its ability to calculate measures we consider clinically significant such as MLU and PCC. It cannot interpret the results of language sample analyses or other types of clinical data.

Conclusion: CPA and LSA programs can increase clinical efficiency by enabling clinicians to analyze language samples at a level that would be difficult or impossible without their use.

Relevance the current study: This article is about computerized language analysis which is another term used for automated language analysis. The issues discussed, such as time and accuracy are motivating concerns for my study.

Hux, K., Morris-Friehe, M., & Sanger, D. D. (1993). Language sampling practices: A survey of nine states. *Language, Speech, and Hearing Services in Schools, 24,* 84-91. doi:10.1044/0161-1461.2402.84

Introduction: There is general agreement among clinicians that language sampling should play an important role in assessment. However, researchers and clinicians are mindful of the disadvantages of performing language sampling. Disadvantages include: the amount of time required, the expertise needed, difficulty in obtaining a language sample that is indicative of the client's ability, and the lack of procedural consistency within and between professionals for elicitation and collection. This study was conducted to survey the collection and analysis of language samples by speech-language pathologists working in schools.

Method: The basis of data collection was a survey consisting of 51 questions. Each survey consisted of three sections: (a) the respondent's background information, (b) practices regarding language sampling procedures, and (c) attitudes about language sampling procedures. The researchers of this study sought input from eight speech-language pathologists and four school administrators who were asked to review the survey for clarity, completeness and relevancy. The final version of the survey included feedback from these professionals. Surveys were sent to 500 speech-language pathologists working in10 Midwestern states (Colorado, Illinois, Iowa, Kansas, Montana, Nebraska, North Dakota, South Dakota, Utah, and Wyoming). Fifty participants were randomly chosen from each of the 10 states. Participant names were chosen from the state's professional organization and from personnel lists from the State Department of Education. The final participant pool consisted of 239 subjects from nine states. North Dakota did not meet the return rate of 40% and was excluded from the study.

Results: Most of the respondents (92%) worked in public school settings additionally (67%) held certificates of clinical competence from American Speech–Language–Hearing Association

(ASHA). Respondent caseloads varied from 15 to more than 75 students, with more than half (54.8%) having caseloads between 30 and 60 clients. The majority of the services provided were articulation and language cases, with language cases making up the greatest percentage. There were definite trends in the responses to questions asked about age and severity of clients. The survey indicated a preference for the use of non-standardizes language sample analysis. The two types of information obtained most frequently were mean length of utterance (81%) and qualitative language descriptors (80%).

Discussion: School-based speech-language pathologists often use language sample analyses to supplement standardized assessments and to plan client treatments. Neither state education agencies nor local school districts mandated language sampling for 82% of the respondents. However, the speech-language pathologists seemed to value language sampling information. These data are reassuring and indicate the commitment of speech-language pathologists to provide quality assessments irrespective of administrative guidelines.

Relevance to current study: Information in this article addressed many of the reasons that clinicians have difficulty with language samples. The use of automated DSS, which is one focus of my study, will provide a time-saving factor which allows clinicians to more readily use language sample analysis.

Klee, T. (1985). Clinical language sampling: Analyzing the analysis. *Child Language Teaching and Therapy, 1*(2), 182-198. doi:10.1177/026565908500100206

Introduction: Identifying and evaluating language disorders in children has shifted from dependence on tests and elicitation procedures to a more naturalistic approach of examining the child's actual linguistic production obtained in conversational setting. This study presented 14 clinical procedures which can be used for language sample analysis (LSA). Several of these procedures were reviewed and evaluated.

Mechanics of LSA: There are four phases used in conducting a clinical LSA:

1. Recording a conversation. Both the actual recording and the conversational interaction should be considered. Choose a context which accurately reflects the child's conversational linguistic and communicative abilities. High quality recording will aid in transcription.

2. Transcription. A visual record is made from the audio-recording. Only those trained in the study of child language should perform the transcription, however a more efficient mode is to use a transcription machine. The four levels of transcription recognized are broad morphemic, narrow morphemic, broad phonetic and narrow phonetic.

3. Analysis. After the transcription is completed, choose the most appropriate means of analysis which will best reveal the child's problematic linguistic areas.

4. Interpretation. Form a hypothesis by examining the scope and consistency of patterns in the child's language. The clinician moves from objective information to a subjective interpretation. Goals for intervention are established during this phase.

Language Sample Analysis: Language can be broadly divided into the categories of structure and use. Language structure can again be divided into the domains of phonology, semantics and grammar. Since Lee (1974) first standardized linguistic analysis, more than a dozen clinical linguistic analyses have been published. Most of these analyses are grammatically based. However, assessments for phonology and semantics now exist. Phonology is the study of the sound systems of language. It includes both segmental and non-segmental aspects of the sound systems. Presently, there are five procedures used to provide a clinical assessment of phonology.

Semantics is defined as the study of the meaning of language. Linguistic meaning can be further subdivided into lexical semantics and relational (or discourse) semantics. There are also five procedures used to provide clinical assessment in the area of semantics. Grammar is comprised of syntax and morphology. Syntax studies the rules governing how words can be combined to form larger units of speech. Morphology studies the form and structure of words. Currently clinicians use six assessment procedures for clinical analysis.

The Developmental Sentence Scoring (DSS) technique (Lee & Canter, 1971; Lee, 1974) is the only norm-referenced procedure. DSS is useful in establishing a baseline to use when determining intervention goals. A disadvantage of DSS is that the developmental classification used within some of the eight grammatical categories is not congruent with the current child language acquisition research. Also, DSS does not differentiate lexical elements with a different level of syntax. Language Assessment, Remediation and Screening Procedure (LARSP; Crystal, Garman, & Fletcher, 1976) uses an adult grammatical framework to provide a developmental description of a child's language. LARSP aims at analyzing each utterance in a sample and provides a criterion-referenced analysis. LARSP provides the ability to analyze clauses, phrases and word structure, while DSS mainly provides a phrase level analysis. Systematic Analysis of Language Transcripts (SALT; Miller & Chapman, 1983) is a fully-automated linguistic analysis computer program which provides analyses directly from the language transcript. SALT includes the benefit of *ad hoc* analysis, by means of the SEARCH program, which is specified by the clinician. Using SALT and LARSP together allows practically all levels of grammar to be investigated.

Relevance to current study: This study discusses some of the advantages and disadvantages of DSS; clinicians should be aware of these when using DSS in a clinical setting. Klee discusses the reasoning for clinical use of language sample analysis such as DSS. My study reports information about automated DSS analysis which would reduce the time needed to complete a language sample analysis. This time saving factor will allow language sample analysis to be used more frequently by clinicians.

Smith, J. L., DeThorne, L. S., Logan, J. A. R., Channell, R. W., & Stephen, A. P. (2014). Impact of prematurity on language skills at school age. *Journal of Speech, Language, and Hearing Research, 57,* 901–916. doi:10.1044/10924388(2013/12-0347)

Introduction: Approximately 500,000 babies are born prematurely each year in the United States. Premature infants are at an increased risk for many morbidities including hearing and vision deficits, impaired neurodevelopment, as well as behavior problems. Although rates of multiple births, which are associated with the increase of premature births, have increased in recent decades, advances in neonatal intensive care has led to a decreased mortality rate for premature babies. Currently, approximately 85% of very low birth weight (VLBR) babies survive to be discharged from the hospital. However, the population of impaired survivors has increased due to the incidence of neurodevelopmental consequences remaining constant even though more babies are surviving. The purpose of this study was to address the scarcity of discourse-based language outcomes of prematurely born children.

Method: Participants: The participants for this study were obtained from the Western Reserve Reading Project (WRRP; Petrill et al., 2006) longitudinal study, which assessed the abilities of children's reading, mathematics, and related skills. Data were obtained from 368 same-sex twins living primarily in Ohio. Each set of twins began participation when they were in either kindergarten or first grade. Children from the WRRP sample were chosen to participate in the premature group if they met one of two criteria. First, if he or she had a VLBR of less than 1,500 g. Second, if he or she was born at or fewer than 32 weeks' gestation. Fifty-seven children (19 boys and 38 girls) met these criteria. The control group consisted of children born at least 37 weeks' gestation with no perinatal complications reported.

General procedure: Two WRRP examiners visited families in their homes each year beginning when the participants were an average of six years-old. Additional longitudinal data were obtained at approximately seven, eight, and 10 years of age. The examiners collected data, which included a conversational language sample, at Year 1 (age 7). The Year 2 visit (age 8) included measures of reading ability, conversational language sampling as well as other measures of language ability. The Test of Narrative Language (TNL; Gillam & Pearson, 2004) was administered at Year 3. The TNL provides both a standardized score for narrative ability and a narrative language sample for analysis.

Language Sample Procedure: Fifteen-minute conversational language samples were collected.

Semantic Measures: Systematic Analysis of Language Transcripts (SALT) was used to calculate the number of different words (NDW) and the number of total words (NTW). The transcriptions were obtained from the first 100 utterances the children produced at Year 1 and Year 2, and the first 50 utterances spoken at Year 3. The use of low-frequency vocabulary was used in this study because children demonstrate less ability with these words. The first 100 utterances produced by a participant at Year 1 and Year 2, and the first 50 utterances produced at Year 3 were used to calculate the NDW and NTW via SALT.

Syntactic Measures: All language samples containing complete and intelligible utterances were analyzed to provide the mean length of utterance in C-units (MLU-C). MLU-C can differentiate children with varying language abilities and note developmental change through the school years.

Results: Although in some cases the differences were small, the control group outperformed the premature group in production of all target structures of growth in semantic and syntactic measures. The control group also outscored the premature group on performance of standardized tests.

Discussion: The results revealed that prematurely born school-age children are outperformed by peers born at full term on standardized tests. Results of this study were consistent with existing literature reporting that premature children are within the lower end of normal range and not outside it.

Relevance to current study: This article included DSS as a research method to evaluate children's language abilities. Although the clinical use of DSS has declined, this study demonstrates that DSS is still used in research settings. Automated DSS analysis, a focus of my study, would be useful to researchers.

Yoder, P. J., Spruytenburg, H., Edwards, B., & Davies, B. (1995). Effect of verbal routine contexts and expansions on gains in the mean length of utterance in children with developmental delays. *Language, Speech, and Hearing Services in Schools, 26,* 21-32. doi:10.1044/0161-1461.2601.21

Introduction: An important goal for children with developmental delay is to increase their length of utterance. Young children with longer utterances express greater grammatical and semantic information. Nelson (1989) suggested that children develop sematic relations and syntactic knowledge when aided by adult expansions. Adults use expansion by providing utterances after a child's, referring to the central relationships and events of a child's utterance and increasing the semantic or syntactic complexity of the communication. The purpose of this study was to assess the hypothesis that a child's mean length of utterance (MLU) is increased by verbal routines and expansions. Although the subjects varied in chronological age, mental age, mental development index, receptive language age and productive language age, each child scored in the borderline or mild mentally impaired range.

Method: Three male and one female participants were included in this study. Each participant attended a university-based preschool for children with developmental delays. A multiple-baseline-across-subjects design was used to assess the effect of intervention on MLU. Independent variables were verbal routines and adult expansions of participant nominative utterances. All sessions took place in a play laboratory. A baseline MLU was established prior to the intervention phase which consisted of four weekly sessions. To enable the participant to develop a verbal routine, each child was repeatedly exposed to the same book. Next, the child was asked questions about the pictures on the page. After pausing for the child's response, adults were instructed to use complete sentences to expand the child's non-imitative utterances. Participants' MLU in morphemes was calculated using the Systematic Analysis of Language Transcripts (SALT) program. Participants were referred to as "cases" as follows: (a) particular cases within subject B are referred to B-1 or B-2 depending on whether the data came from the first or second book; (b) cases with only one book are referred to by their subject ID alone (e.g., A, C, and D).

Results: Strong evidence was seen on the intervention effect on case A, moderately strong evidence of intervention effect on case B-1, and strong evidence of an intervention effect on case C. Results for B-2 and D lacked confidence in interpretation.

Discussion: There was stronger backing for an intervention effect on generalized MLU for cases A, B-1, and C. Although there was no baseline for the interventions sessions, it is probable that the increase in the children's non-imitative utterances obtained during the intervention sessions was due to expansions and/or the repeated experience to the same book which lead to verbal routines. However, it should be kept in mind that MLU does not allow one to differentiate between memorized phrases that may have been learned during the intervention sessions and novel combinations of words. To identify which children will benefit most from expansions embedded in routine interactions further research is warranted. The five cases presented in this study are not sufficient to adequately study increases in aptitude as a result of intervention.

Relevance to current study: This study used MLU as a means of scoring children's syntax. My study also includes the use of MLU as one way of measuring children's syntax. The authors of this study used SALT to calculate the MLU of each child; the software program SALT was also used in my study.

Vaughn-Cooke, F. B. (1983). Improving language assessment in minority children. *American Speech-Language-Hearing Association (ASHA),25,* 29-34.

Introduction: According to the American Speech-Language-Hearing Association, a dialectal variety of English is not a disorder. However, it is possible for dialect speakers to have a disorder within the dialect. Currently, there are no generally recognized standardized methods for assessing linguistic ability for persons who speak nonstandard English dialects. This study presented seven proposed alternatives to inappropriate tests for nonstandard English speakers.

1. Standardize existing tests on non-mainstream speakers: This alternative has been used by several researchers including Evard and McGrady (1974), Evard and Sabers (1974), and Evard and Sabers (1979). Evard and McGrady (1974) used non-mainstream speakers in Arizona to standardize the Templin-Darley Tests of Articulation, and the Auditory Association and Grammatic Closure Subtests of the Illinois Test of Psycholinguistic Abilities (ITPA). Two problems have been reported with this adapted standardization. First, low norms were reported. For example, the norms on the Grammatic Closure Subtest were much lower for Black non-mainstream speakers than for standard-English speakers. Second, most standardized language tests are created to expose what a child knows about Standard English. Therefore, children learning non-standard English are at a disadvantage on these tests, and the tests should not be considered valid or appropriate.

2. Include a small percentage of minorities in the standardization sample when developing a test: The standardization of the ITPA features similar problems addressed above. The ITPA normative sample included approximately 4% Black children. This percentage was lower than in the communities from which they were selected and also lower than the nationwide percentage.

Weiner and Hoock (1973) reported that nothing was accomplished, in terms of validity, by including a small percentage of Blacks in this sample.

3. Modify or revise existing tests in ways that will make them appropriate for non-mainstream English speakers: Nelson (1976) and Hemingway, Montague, and Bradley (1981) implemented this alternative. Nelson (1976) modified Developmental Sentence Scoring (DSS) (Lee, 1974) in an effort to make DSS appropriate for Black non-mainstream English speakers. To achieve this goal a thorough knowledge of Black English is necessary. Test modifiers must obtain a comprehensive knowledge non-mainstream dialects before revisions are begun.

4. Utilize a language sample when assessing the language of non-mainstream speakers: Many researchers have recommended the use of this non-standardized alternative to assessing the language of minority children. At least two problems hinder the use of language sample analysis (LSA) for assessing the language of minority children. First, LSA does not provide pertinent information needed to determine if a child's language is normal. Standardized tests are needed in conjunction with LSA. Second, results from LSA require an interpretation within a developmental framework. Currently, most of the LSA performed on non-mainstream speakers are interpreted according to standards established by middle-class white children.

5. Utilize criterion-referenced measures when assessing the language of non-mainstream speakers: Although criterion-referenced testing has an important role in language assessment and intervention, currently its use is one of the difficulties attendant with LSA. Criterion-referenced tests should not be seen as a viable alternative in assessment until after more research has been conducted on the language development of non-mainstream speakers.

6. Refrain from using all standardized tests that have not been corrected for test bias when assessing the language od non-mainstream speakers: A task force on language and communication skills recommended that the following test should not be use when assessing Black English speakers: (a) Peabody Picture Vocabulary Test; (b) Houston Test of Language Development; (c) Utah Test of Language Development; (d) Grammatic Closure Subtest of the

ITPA; (e) DSS; (f) Templin Darley Tests of Articulation; and (g) Wepman Test of Auditory Discrimination.

7. Develop a new test which can provide a more appropriate assessment of the langue of non-mainstream speakers: Many researches believe that this is only solution to the assessment problem. Drumwright et al. (1973) developed the language-based Black Intelligence Test of Cultural Homogeneity. Scores from 200 Black and White high students showed that the Black students performed better on the test than the White indicating that tests developed for purpose of assessing specific knowledge of one cultural group are not suitable for other cultural groups.

Discussion: The alternative approaches listed depict an accurate, but dismal picture. Test developers, researchers and clinicians must increase their efforts to improve assessment for non-mainstream speakers.

Relevance to current study: This study presented ways to improve language assessment for nonstandard English speakers, including DSS. The time saver factor of the automated DSS in my study would make language sample analysis more expedient for clinicians.

Miles, S., Chapman, R., & Sinberg, H. (2006). Sampling context affects MLU in the language of adolescents with Down Syndrome. *Journal of Speech, Language, and Hearing Research, 49,* 325-337. doi:1092-4388/06/4902-0325

Introduction: Persons with Down Syndrome (DS) often display certain language phenotypes including delays in expressive syntax, inaccuracies of grammatical morpheme omission and use, and decreased intelligibility. Mean length of utterance (MLU) is generally used as a measure of expressive language during conversations with clients having DS. The purpose of this study was to describe the procedures used to explain an unexpected finding. Namely, in conversations without picture support, adolescents with DS had a lower MLU that their typically developing peers, but did not have lower a MLU during narratives when wordless picture books were used.

Method: This study included 28 children, adolescents, and young adults; 14 individuals with DS and 14 typically developing (TD) individuals. Ages ranged from 12;10 (years;months) to 21 years. Each participant participated in a seven-session study of word learning and narrative development in adolescents and young adults with DS. Language samples (LS) were obtained on the first of the seven sessions. Wordless picture books were used to elicit two narratives. Spontaneous LS were obtained with an interview format by introducing topics of personal interest. After narratives and interviews were transcribed, the data were entered into the Systematic Analysis of Language Transcripts (SALT).

Results: MLU-Narrative was significantly higher than the MLU-Interview for the group with DS. There was no significant difference in the MLU-Interview results between the DS group and the TD group, $M = 5.69$, $SD = 1.89$, $p < .01$, and $M = 6.19$, $SD = 1.79$, $p < .01$, respectively. The DS-Narrative group was significantly higher than the DS-Interview group $M = 5.69$, $SD = 1.89$, $p = .03$, and $M = 4.38$, $SD = 1.56$, $p = .03$, respectively.

Discussion: The TD group showed no difference in MLU obtained via narrative or interview. The DS group had a higher MLU in the narrative context versus the interview. The use of pictures, rather that just using narrative, increased the MLU scores for the DS group. Clinically, the use of narrative rather than conversational samples when assessing expressive language for persons with DS allow clinicians to more effectively determine the extent of an individual's skill.

Relevance to current study: Spontaneous language samples were used in this study, as well as in my study, to determine the participants' MLU.. This study used SALT as a means of obtaining MLU as was also done in the current study.

Reed, V. A., MacMillan, V., & McLeod, S. (2001). Elucidating the effects of different definitions of 'utterance' on selected syntactic measures of older children's language samples. *Asia Pacific Journal of Speech, Language, and Hearing, 6*(1), 39-45. doi:10.1179/136132801805576842

Introduction: Clinicians and researchers often use language samples (LS) to examine the spoken language abilities of children. However, there is considerable variation in its implementation. The purpose of this exploratory study was to examine the effects of the utterance definition for T-unit, C-unit, Developmental Sentence Scoring (DSS), and Tone unit.

Method: Ten typically developing male children between the ages of 11;0 (years;months) and 11;11 participated in this study. Each participant was from English speaking, middle socioeconomic families in Sydney, Australia. To elicit samples each child was seen individually, and a competitive element was introduced. For example, the children were asked to assist in an "important university assignment," and asked to retell the "very best story" they could and to explain "very well" how to play a game. *Frog's Night Out* (Glenn & McLeod, 1993) was used for story retell, and the children were asked to explain how to play Monopoly for the game explanation task. There were seven measures of syntactic ability calculated: (a) mean length of utterance (MLU) in words; (b) MLU in morphemes; (c) number of dependent clauses; (d) number of independent clauses; (e) number of dependent clauses per utterance; (f) number of independent clause per utterance; and (g) number of utterances in the sample.

Results: A one-way ANOVA with repeated measures was conducted for each of the seven language measures to determine if there was a statistically significant main effect of definition. The effect of definition was statistically significant, $p < 0.01$, for each language measure. *Post hoc* analyses were completed to identify the definitional source(s) of difference and the means indicated the direction(s) of difference, and revealed the following patterns of results.

First, LS which were segmented according to DSS definition were different from the LS segmented according to the other three definitions. Both MLU in words and morphemes had longer utterance measures when the DSS definition was implemented. Also, the DSS definition resulted in more dependent and independent clause per utterance than the other three definitions. Also, measuring results from T-unit and C-unit definitions led to significantly different results. Furthermore, the effect of T-unit definition influenced the number of dependent and independent clauses in the samples; with these two measures fewer dependent clauses and more independent clauses were identified in the samples when they were segmented according to the T-unit definition rather that the other three definitions.

Discussion: This study brought to light some of the possible effects of varied definitions of utterance which lead to different segmentations of LS when measuring the syntax of older children. When writing reports, authors can help their readers interpret results by explicitly communicating the utterance definitions used in determining the presented results.

Relevance to current study: Reed et al. discuss the importance of language sample analysis as a means of examining the syntax of children. My study also discusses the importance of language sample analysis. This study included SALT as a means of obtaining participants' MLU, which was also done in the current study.

Holdgrapher, G. (1995). Comparison of two methods for scoring syntactic complexity. *Perceptual and Motor Skills, 81*(2), 498. http://dx.doi.org/10.2466/pms.1995 .81.2.498

Introduction: Scoring syntactic complexity of language samples is accomplished with only a few methods. Two methods considered to be sensitive to individual differences in language acquisition are Developmental Sentence Scoring (DSS) and the Index of Productive Syntax (IPS).

Method: One hundred utterances were obtained from 29 preterm children between ages 3;7 (years;months) and 5;0 years. Language samples were obtained during 15 minutes of play. Ten of the subjects were at greater risk for language delay due to neurological uncertainties. Computerized Profiling was used for sample transcription. Twenty percent of the sample were randomly selected to determine inter-rater reliability, which exceeded 90% for word by word comparisons. Computer-assisted analysis provided summary scores for both the DSS and IPS.

Results: As expected, the Pearson correlation between the scores was moderate (r =58, $p < .001$). The IPS was the only procedure to differentiate the typical versus neurologically suspect groups ($t = 2.8$, $p < .009$).

Discussion: According to the results of this study, IPS is more sensitive that the DSS.

Relevance to current study: This study used Developmental Sentence scoring as a means of measuring syntactic complexity as was also done in the current study. Computerized profiling was used in this study as well as in my study as a means of sample transcription.

Eisenberg, S. L., Fersko, T. M., & Lundgren, C. (2001). The use of MLU for identifying language impairment in preschool children: A review. *American Journal of Speech-Language Pathology, 10*(4), 323-342. doi:10.1044/1058-0360(2001/028)

Introduction: It is common for speech-language pathologists to be asked to determine if a child has a language disorder. There are many norm-referenced tests available to aid in this decision. However, some children cannot be assessed with formal measures. A recent survey reported that 93% of speech-language pathologists use language sample analysis (LSA), with mean length of utterance (MLU) being the most widely used (91%) procedure in LSA. The following are suggested guideline for evaluating assessment tools: (a) clear definition of purpose; (b) sufficient description of administration and scoring procedures; (c) sufficient description of the normative sample; (d) appropriate reference data; (e) evidence of reliability; (f) evidence of validity. This article summarizes available information on MLU.

Purpose: The intended purpose of an assessment instrument is necessary to evaluate its adequacy. There are three aspects to purpose: (a) domain; (b) population; (c) assessment aim. The domain or trait is what is being measured. It is important not to overemphasize the trait being measured when defining a domain. Therefore, MLU should be considered as one of several possible ways to measure utterance length instead of as a measurement of morphosyntax. Eisenberg, Fersko and Lundgren referred to two sets of MLU reference data as their population. First, Miller and Chapman (MC; 1981) who reported MLU data for children ages 18-60 months. Second, Leadholm and Miller (LM; 1992) who reported MLU data for children ages 3-13 years of age. The following assessment aims have been suggested for MLU: (a) to diagnose or identify a language impairment; (b) to determine stage or overall level of language development; (c) to guide further langue assessment; (d) to compare language use across situations (e) to measure change in language impairment.

Administration and Scoring Procedures: Most text books recommend obtaining language samples in two contexts. Speech-language pathologists should know and follow administration and scoring procedures. Both MC and LM used conversation sampling procedures. When obtaining language samples, specific sample size, setting, participant, instructions given to those interacting, the activity, and materials are needed for acceptable standardization. Criterion for scoring items should be provided for a standardized test.

Normative Sample: For clinicians to determine how representative a sample is for a certain child or a type a child, the normative sample must be described sufficiently. For example, to study the relationship between MLU and age MC pooled data from five different studies as a means of producing a research report rather than develop local norms. LM sought to develop local norms by included 100 children ages 3-5 years.

Reference Data and Interpretation: Means and standard deviations were reported for all participants. On its own, MLU is not interpretable as it needs norm-referenced data. Klee et al. year had a cutoff of -1.5 *SD*, but at this cutoff the sensitivity was only 63%. Therefore, it could not be concluded that children with a MLU higher than that cutoff rate have normal language. However, an MLU below that cutoff rate may support a diagnosis of language impairment.

Reliability: Both consistency of administration and scoring are factors of examiner reliability and agreement. As reported by MC, inter-examiner agreement for utterance segmentation ranged from 85-95%.

Validity: Generally, validity is defined as the extent to which a text measures what it claim to measure. Rather than using this definition which implies that validity is an inherent trait, another view is that validity is more a matter of how test results are used rather than something the test does or does not have. Brown (1973) did not define or provide operational criteria for identifying the term *utterance.* This is cause for concern being that the number of utterances is necessary for the MLU calculation.

Discussion: MLU should be used as a way of measuring utterance length rather than as a measure of syntactic development. MLU is capable of identifying some, but not all, language impaired preschool children. To identify the majority of children that are not language impaired, a limit can be set. Therefore, a low MLU is supportive of a language impairment diagnosis.

However, an MLU above the set limit does not automatically preclude a child from having an impairment.

Relevance to current study: This study examines the use of MLU for identifying language impairment. MLU is one of the measures used in my study to assess the syntax of children. Eisenberg et al. report that a low MLU is indicative of language impairment as is mentioned in my study.

Appendix B: DSS Scores from Manual Analysis and Automated Analysis

ID	Age in Months	Number of Sentences	DSS	DSSA	MLU
1	30	141	5.71	6.45	4.06
2	30	117	5.79	5.68	3.26
3	33	129	5.96	5.82	4.26
4	35	168	5.93	5.98	4.56
5	37	144	6.22	6.15	3.74
6	39	101	3.67	4.87	3.9
7	45	186	7.46	7.48	5.54
8	45	185	7.10	7.08	5.58
9	46	152	7.04	6.18	5.03
10	53	148	8.82	9.02	5.15
11	56	161	10.11	10.03	5.17
12	59	138	10.54	10.28	6.07
13	59	182	9.09	8.65	5.61
14	62	162	8.25	8.22	5.16
15	62	168	6.84	6.38	4.46
16	64	161	7.32	7.14	5.23
17	65	135	7.41	7.45	4.29
18	65	187	11.50	11.51	6.89
19	66	163	8.36	8.08	4.97
20	68	151	7.86	7.4	4.28
21	69	177	11.29	10.82	5.66
22	72	149	9.17	8.26	4.94
23	75	190	9.28	8.67	6.27
24	77	195	12.29	11.9	6.64
25	79	160	6.96	6.57	4.35
26	79	167	8.42	7.78	5.2
27	84	149	8.38	8.03	5.11
28	91	160	8.92	8.39	5.08
29	94	195	13.81	13.72	7.37
30	95	189	13.41	13.03	6.54

Note: DSS = developmental sentence score; DSSA = Developmental Sentence Scoring Automated version 2.0 (DSSA; Channell, 2016); MLU = mean length of utterance in morphemes.

Appendix C: DSS Scoring Chart (from Lee, 1974)

SCORE	INDEFINITE PRONOUNS OR NOUN MODIFIERS	PERSONAL PRONOUNS	MAIN VERBS	SECONDARY VERBS	NEGATIVES	CONJUNCTIONS	INTERROGATIVE REVERSALS	WH-QUESTIONS
1	it, this, that	1st and 2nd person: I, me, my, mine, you, your(s)	A. Uninflected verb: *I see* you. B. copula, is or 's: *It's* red. C. is + verb + ing: He *is coming.*		it, this, that + copula or auxiliary is, 's, + not: It's *not* mine. This is *not* a dog. That is *not* moving.		Reversal of copula: *Isn't it* red? *Were they* there?	
2		3rd person: he, him, his, she, her, hers	A. -s and -ed: *plays, played* B. irregular past: *ate, saw* C. Copula: *am, are, was, were* D. Auxiliary *am, are, was, were*	Five early-developing infinitives: I wan*na see* (want *to see)* I'm gon*na see* (going *to see)* I got*ta see* (got *to see)* Lemme [to] see (let me [*to*] *see)* Let's [to] play (let [us *to*] *play)*				A. who, what, what + noun: *Who* am I? *What* is he eating? *What book* are you reading? B. where, how many, how much, what . . . do, what . . . for *Where* did it go? *How much* do you want? *What* is he *doing?* *What* is a hammer *for?*
3	A. no, some, more, all, lot(s), one(s), two (etc.), other(s), another B. something, somebody, someone	A. Plurals: we, us, our(s), they, them, their B. these, those		Non-complementing infinitives: I stopped *to play.* I'm afraid *to look.* It's hard *to do* that.		and		
4	nothing, nobody, none, no one		A. can, will, may + verb: *may go* B. Obligatory do + verb: *don't go* C. Emphatic do + verb: I *do see.*	Participle, present or past: I see a boy *running.* I found the toy *broken.*	can't, don't		Reversal of auxiliary be: *Is he* coming? *Isn't he* coming? *Was he* going? *Wasn't he* going?	
5		Reflexives: myself, yourself, himself, herself, itself, themselves		A. Early infinitival complements with differing subjects in kernels: I want you *to come.* Let him [*to*] *see.* B. Later infinitival complements: I had *to go.* I told him *to go.* I tried *to go.* He ought *to go.* C. Obligatory deletions: Make it [*to*] *go.* I'd better [*to*] *go.* D. Infinitive with wh-word: I know what *to get.* I know how *to do* it.	isn't, won't	A. but B. so, and so, so that C. or, if		when, how, how + adjective *When* shall I come? *How* do you do it? *How big* is it?
6		A. Wh-pronouns: who, which, whose, whom, what, that, how many, how much I know *who* came. That's *what* I said. B. Wh-word + infinitive: I know *what* to do. I know *who(m)* to take	A. could, would, should, might + verb: *might come, could be* B. Obligatory does, did + verb C. Emphatic does, did + verb			because	A. Obligatory do, does, did: *Do they* run? *Does it* bite? *Did*n't *it* hurt? B. Reversal of modal: *Can you* play? *Won't it* hurt? *Shall I* sit down? C. Tag question: It's fun, *isn't it?* It isn't fun, *is it?*	
7	A. any, anything, anybody, anyone B. every, everything, everybody, everyone C. both, few, many, each, several, most, least, much, next, first, last, second (etc.)	(his) own, one, oneself, whichever, whoever, whatever Take *whatever* you like.	A. Passive with *get*, any tense Passive with *be*, any tense B. must, shall + verb: *must come* C. have + verb + en: *I've eaten* D. have got: I'*ve got* it.	Passive infinitival complement: With *get:* I have *to get dressed.* I don't want *to get hurt.* With *be:* I want *to be pulled.* It's going *to be locked.*	All other negatives: A. Uncontracted negatives: I can *not* go. He has *not* gone. B. Pronoun-auxiliary or pronoun-copula contraction: I'm *not* coming. He's *not* here. C. Auxiliary-negative or copula-negative contraction: He was*n't* going. He has*n't* been seen. It could*n't* be mine. They are*n't* big.			why, what if, how come how about + gerund *Why* are you crying? *What if* I won't do it? *How come* he is crying? *How about* coming with me?
8			A. have been + verb + ing had been + verb + ing B. modal + have + verb + en: *may have eaten* C. modal + be + verb + ing: *could be playing* D. Other auxiliary combinations: *should have been sleeping*	Gerund: *Swinging* is fun. I like *fishing.* He started *laughing.*		A. where, when, how, while, whether (or not), till, until, unless, since, before, after, for, as, as + adjective + as, as if, like, that, than I know *where* you are. Don't come *till* I call. B. Obligatory deletions: I run faster *than* you [run]. I'm *as big as* a man [is big]. It looks *like* a dog [looks] C. Elliptical deletions (score 0): That's *why* [I took it]. I know *how* [I can do it]. D. Wh-words + infinitive: I know *how* to do it. I know *where* to go.	A. Reversal of auxiliary have: *Has he* seen you? B. Reversal with two or three auxiliaries: *Has he been* eating? *Could*n't *he have* waited? *Could he have been* crying? *Would*n't *he have been* going?	whose, which, which + noun *Whose* car is that? *Which book* do you want?