2002

Scoring Sentences Developmentally: An Analog of Developmental Sentence Scoring

Amy Seal
Brigham Young University - Provo

SCORING SENTENCES DEVELOPMENTALLY: AN ANALOG OF DEVELOPMENTAL SENTENCE SCORING

by

Amy Seal

A thesis submitted to the faculty of

Brigham Young University

in partial fulfillment of the requirements for the degree of

Master of Science

Department of Audiology and Speech-Language Pathology

Brigham Young University

April 2001

BRIGHAM YOUNG UNIVERSITY

GRADUATE COMMITTEE APPROVAL

of a thesis submitted by

Amy Seal

This thesis has been read by each member of the following graduate committee and by majority vote has been found to be satisfactory.

Date	Ron W. Channell, Chair
Date	Bonnie Brinton
Date	Martin Fujiki

BRIGHAM YOUNG UNIVERSITY

As chair of the candidate's graduate committee, I have read the thesis of Amy Seal in its final form and have found that (1) its format, citations, and bibliographical style are consistent and acceptable and fulfill university and department style requirements; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the graduate committee and is ready for submission to the university library.

Date

Ron W. Channell
Chair, Graduate Committee

Accepted for the Department

Ron W. Channell
Graduate Coordinator

Accepted for the College

Robert S. Patterson
Dean, David O. McKay School of Education

ABSTRACT

SCORING SENTENCES DEVELOPMENTALLY: AN ANALOG

OF DEVELOPMENTAL SENTENCE SCORING

Amy Seal

Department of Audiology and Speech-Language Pathology

Master of Science

A variety of tools have been developed to assist in the quantification and analysis of naturalistic language samples. In recent years, computer technology has been employed in language sample analysis. This study compares a new automated index, Scoring Sentences Developmentally (SSD), to two existing measures. Eighty samples from three corpora were manually analyzed using DSS and MLU and the processed by the automated software. Results show all three indices to be highly correlated, with correlations ranging from .62 to .98. The high correlations among scores support further investigation of the psychometric characteristics of the SSD software to determine its clinical validity and reliability. Results of this study suggest that SSD has the potential to compliment other analysis procedures in assessing the language development of young children.

ACKNOWLEDGMENTS

A project of this magnitude is certainly not undertaken or completed by a single individual. Although, my name is listed as the sole author, this work is result of the efforts of many. I would like to thank my parents, Gregory and Suzanne Seal, for their love and support. Their constant encouragement kept me going when stress, fatigue, or procrastination threatened to halt my progress. I want to thank my fellow students for their friendship. I couldn't have done it with out them. They know who they are!
The late nights, marathon study sessions, and equally lengthy study breaks made this journey so much more enjoyable! And finally, to Dr. Ron Channell, I want to express my sincere thanks for the endless hours he spent answering my questions, talking me through details, and editing my fledgling drafts. His sense of humor and positive outlook kept me motivated and smiling (most of the time) through the months of research and writing. Without his guidance, skill, and patience, this work never would have moved beyond page one.

TABLE OF CONTENTS

LIST OF TABLES

Introduction

A variety of useful tools and indices for language sample analysis have been developed to assist in the quantification of natural, spontaneous language. The ability to quantify language provides a basis for collecting normative data and making developmental comparisons (Bennett-Kastor, 1988; Miller, 1991). Quantified descriptions of language can be useful in providing baseline information prior to developing appropriate intervention goals (Klee & Paul, 1981; Klee, 1985). Normative data are also valuable for measuring progress during intervention and comparing treatment outcomes (Hughes, Fey, & Long, 1992; Lee, 1974). Existing quantification measures range from frequency count procedures such as Mean Length of Utterance (MLU; Brown, 1973), to scored indices of grammatical complexity such as Developmental Sentence Scoring (DSS; Lee, 1974) and the Index of Productive Syntax (IPSyn; Scarborough, 1990).

For more than 30 years, MLU has been used as a measure of grammatical development. The correlation between MLU and the acquisition of grammatical morphemes has been verified (de Villiers & de Villiers, 1973; Klee & Fitzgerald, 1985; Rondal, Ghiotto, Bredart, & Bachelet, 1986). However, the validity of MLU beyond the age of two or three (Bennet-Kastor, 1988; Klee, Schaffer, May, Membrino, & Mougey, 1989; Rondal et al., 1986) and its sensitivity to syntactic development (Klee et al., 1989) have been called into question. Despite these criticisms MLU maintains widespread clinical use (Kemp & Klee, 1997; Muma, Pierce, & Muma, 1983).

DSS is the most commonly recognized formal procedure for grammatical language sample analysis. Although the DSS procedure is more than 20 years old, it continues to be recognized as a valid, reliable tool for obtaining information about grammatical development (Hughes et al., 1992). Reportedly DSS is the tool most frequently employed by clinicians practicing language sample analysis (Hux, Morris-Friehe, & Sanger, 1993; Kemp & Klee, 1997). While DSS enjoys clinical popularity, the procedure is not without its limitations. The reliability of DSS scores using only the recommended 50-utterance sample has proven to be problematic (Johnson & Tomblin, 1975). In addition, DSS does not account for incomplete utterances and emerging forms in the scoring procedure.

Automated versions of DSS have been developed to facilitate more efficient grammatical analysis. As with most language sample analysis tools, DSS is time-consuming and requires clinician skill and training (Hux et al., 1993; Kemp & Klee, 1997). In order to decrease these time and resource demands, programs such as Computerized Language Analysis (CLAN; MacWhinney; 1991) and Computerized Profiling (CP; Long & Fey, 1993) were developed to perform automated DSS analysis. However, the accuracy of these programs is variable at best. Both CLAN and CP display low accuracy rates in certain grammatical categories (Boyce, 1995) and are unable to detect subtle nuances of DSS scoring such as correctness of use (e.g. pronoun gender agreement). In addition, there are elements of DSS that do not lend themselves to automation at all, including attempt marks and sentence points. The absence of these DSS features raises the question as to whether the analyses performed by existing programs can truly be termed DSS. In order to obtain a complete and accurate DSS analysis, the clinician must make corrections and additions to the generated data. Since DSS output from CLAN and CP requires manual correction, both programs can be classified as only “semi-automated” (Baker-Van Den Goorbergh, 1994).

Current views maintain that fully automated programs (i.e. programs which do not require clinician assistance beyond the initial input of the transcript) are not yet practical (Baker-Van den Goorbergh, 1994; Long & Fey, 1995). However, this position is based on the practice of designing computer software to execute existing manual analysis procedures. The ability of computers to precisely replicate tools created for manual use is presently limited. Fully automated programs permit the user to input an uncoded transcript, and the software codes each utterance and computes the results (Long, 1991). Such software is well within the scope of current technology. To achieve acceptable levels of accuracy and efficiency, however, fully automated programs must represent independent indices designed specifically for automated analysis.

Clearly there is a need for an automated index that carries out the same function as DSS. The index should serve as more than a simple imitation of manual methods. Rather, such a program should accomplish the same goals as DSS but constitute a new, distinct instrument. Modifications to the prescribed procedures of manual DSS can be made to accommodate the constraints of automation, while maintaining the integral components of grammatical analysis. As with all independent measures, automated indices must be psychometrically evaluated to establish compliance with standards of acceptable clinical testing (American Psychological Association, 1985; Worthen, White, Fan, & Sudweeks, 1999). In addition, separate normative data must be collected for the index, independent of data compiled in the original DSS literature.

An analog of DSS grew out of initial attempts to refine existing versions of automated DSS. Recognizing that some elements of DSS couldn’t be automated (e.g.

sentence points, attempt marks) and other elements were functionally unnecessary (e.g. using only complete utterances), Channell (2000) developed a new measure based on the principles of DSS but with modifications to the original procedure. The result is an independent index called Scoring Sentences Developmentally (SSD).

The present study looks at the SSD and examines how well it correlates with manual DSS and MLU. The analog was assessed to determine its ability to obtain a detailed, quantified, and scored evaluation of grammatical structures comparable to results obtained with manual DSS and MLU procedures. Such a comparison provides information regarding the effectiveness and value of the analog. The correlational analysis of this study represents only the first step in developing and evaluating a fully automated index of grammatical complexity. Future research is necessary to investigate the psychometric validity and reliability of the index and to establish an independent compilation of normative data.

Review of Literature

Standards for Evaluating Assessment Instruments

The use of norm-referenced and standardized tests is widespread in educational, psychological, and clinical settings. Criteria have been established to evaluate psychometric measures used in assessment procedures (American Psychological Association, 1985). Validity and reliability have been identified as the primary standards that must be met in all clinical tests before operational use. Validity refers to the appropriateness and usefulness of inferences drawn from a test. Construct validity focuses on the ability of the test to measure the characteristic of interest. Content validity demonstrates the degree to which individual items or components of the test represent the domain of content. Criterion-related validity refers to the relationship between tests scores and some predetermined external criterion. Reliability is defined as the extent to which the test is free from errors of measurement. Four types of reliability are generally considered, including test-retest, parallel form, internal consistency, and interrater reliability (Worthen et al., 1999).

Psychometric standards of testing have been applied to tests assessing language disorders. McCauley and Swisher (1984a) asserted the importance of using appropriate norm-referenced tests to separate disordered from non-disordered language. Thirty norm-referenced language and articulation tests designed for use with preschool children were evaluated on the basis of 10 psychometric criteria. Results of the examination indicated that fewer than 20% of the reviewed tests met 5 of the 10 criteria and 50% of the tests met two or fewer criteria. Criteria requiring empirical evidence of validity and reliability were met least often, indicating that these tests failed to demonstrate many of the psychometric characteristics required of well-designed norm-referenced tests.

A companion article by McCauley and Swisher (1984b) acknowledged the flaws and misuses of norm-referenced tests while still asserting the value and necessity of such tests when used properly. Using a hypothetical client, the authors addressed four common errors associated with norm-referenced testing and provided guidelines to avoid potential problems. Although McCauley and Swisher maintained their support of norm-referenced testing, they conceded that the tendency for norm-referenced tests to provide incomplete or misleading information requires greater reliance on the use of language sample analysis and development of criterion-referenced tests.

Muma (1998) contended that McCauley and Swisher (1984a) misrepresented his views regarding the usefulness of psychometric testing in the problem – no problem issue. Muma reaffirmed the role of norm-referenced tests in identifying language disorders but criticized the heavy reliance on psychometric normative testing for overall language assessment. Citing construct validity as the crucial standard for any test, Muma stated that many tests widely used in clinical practice lack this type of validity. Further, Muma questioned the practice of using norm-referenced testing in which "contrived activities are imposed on an individual in a priori procedures" (p. 179) rather than allowing for descriptions of spontaneous intentional language within a natural context. Muma advocated the use of descriptive procedures, such as language sampling, to overcome this issue. Psychometric standards have traditionally not been applied to language sampling procedures since few procedures are norm-referenced and sample collection techniques are not standardized. Muma notes, however, that descriptive assessment is "well grounded on philosophical view and theoretical perspectives thereby having construct validity" (pp. 177-178), thus yielding strong psychometric support to language sample analysis.

Language Sample Analysis

Language production in its many manifestations is the most seriously impaired process among children with language disorders (Miller, 1991). The clinical value of language sampling in the assessment of child language has long been established (Bloom & Lahey, 1978; Gallagher, 1983; Hux et al., 1993; Klee, 1985; Lee, 1974). The primary purposes of language sample analysis are to characterize the nature of a child's linguistic system, both individually and in relation to same-age peers, and to develop and evaluate appropriate goals for intervention (Klee & Paul, 1981). A variety of analysis procedures and instruments have been developed. Menyuk (1964) broadly classified these approaches as descriptions of sentence length, examinations of sentence structure complexity, and proportions of usage of different sentence structures at various age levels. Miller (1981) differentiated procedures on the basis of whether they quantify structural and semantic development to evaluate developmental status of a child or identify structural or semantic problems within a child's system.

Prevalence of Language Sampling

Muma, Pierce, and Muma (1983) surveyed the philosophical orientation and the assessment and intervention procedures advocated by speech-language pathology training programs. Open-response surveys were completed by 76 training programs recognized by American Speech and Hearing Association. Of the 76 respondents, 71 reported using language sampling and analysis techniques. Thirty-seven respondents specifically mentioned the use of DSS. Results indicated that language sampling procedures were most frequently used with young children. Muma et al. concluded that practices reported speech-language pathology training programs reflect a recognition of the importance of language-based assessment and intervention.

Hux et al. (1993) examined the language sampling practices of school-based speech-language pathologists across nine states. The study included responses to 51 questions addressing the background, attitudes and sampling and analysis procedures used by 239 speech-language pathologists. Although time constraints, lack of skills, and diminished resources are common difficulties associated with language sampling, results of the survey revealed that respondents routinely use language sampling practices in assessment and treatment of school-aged children. The majority of respondents (60%)

obtained samples of 51 to 100 utterances in length. Fifty-one percent of respondents reported collecting samples during one setting only. Respondents also showed a clear preference for non-standardized procedures of analysis. Respondents indicating a preference for standardized procedures identified DSS as the only method used with regularity. The majority of respondents judged language sampling as a reliable and useful means of distinguishing between students with normal and disordered language. Hux et al. reported that although 82% of respondents indicated language sampling was not mandated by local or state agencies, speech-language pathologists regularly implemented such practices as part of assessment.   Hux et al. cited the infrequency of language sampling for adolescent, culturally diverse, or mildly impaired populations, and the tendency of clinicians to rely on self-designed methods rather than standardized procedures with proven validity and reliability as areas of concern.

Kemp and Klee (1997) followed up with a similar survey to assess the generalizabilty of the Hux et al. (1993) findings and to judge the extent to which changes in the workplace had impacted clinical use of language sampling. Kemp and Klee surveyed 253 speech-language pathologists employed in preschool positions across 45 states regarding language sampling practices. Eight-five percent of respondents reported using language sample analysis in the assessment of language impairment in preschool children. Of clinicians using language sample analysis, 92% reported using it for diagnosis, 44% for screening, 77% for intervention, and 64% for post intervention. Clinicians not using language sampling reported lack of time (86%), lack of computer resources (40%), lack of training and expertise (16% each), and financial constraints (15%) as reasons for not using analysis procedures. Almost half of the respondents preferred collecting samples based on the number of utterances rather than length of time. Nearly half of the respondents also indicated a preference for non-standardized procedures of analysis. Of the standardized procedures noted, DSS (35%) and Lahey's (1988) Content/Form/Use (29%) were most often cited. Only 8% reported using a computer program for language sample analysis. Kemp and Klee observed that most clinicians endorsed language sample analysis as important in the assessment process but found that the time, effort, and skills required often make the practice difficult. Kemp and Klee concluded that clinical practice must find ways to accommodate the demands placed on clinicians by developing assistive technology to aid in the transcription and analysis of language samples.

Simple Count Analyses

Type/Token Ratio. Simple frequency counts have been used to quantify semantic aspects of language such as lexical diversity (Miller, 1981; Richards, 1986). Templin (1957) studied 480 children and devised the Type/Token Ratio (TTR) as a means of weighing the number of different words produced in a 50-utterance sample against the total number of words produced. Templin found a ratio of 1:2 (.50) to be consistent across age, sex, and socio-economic status. Miller (1981) viewed TTR as a valuable clinical tool for baseline assessment due to its consistency. Traditionally, a low TTR has been used as a warning for possible restrictions on the range of vocabulary used by a child in his or her syntactic repertoire (Fletcher, 1985). Richards (1987) argued, however, that TTR reveals more about the number of tokens in the sample rather than the actual range of vocabulary usage. He suggested that without adequate sample sizes and established norms, the clinical use of TTR is unreliable. In addition, Bennett-Kastor

(1988) noted that TTR is sensitive to context constraints and should not be used the sole measure.

Mean Length of Utterance. The use of MLU as a measure syntactic complexity in child language is a long-standing practice. Brown (1973) popularized the use of MLU based on morpheme count as a simple index of grammatical development. He asserted that as a child's grammar increases in complexity through the acquisition of additional morphemes and structures, there is a corresponding increase in utterance length. Brown identified 14 categories of grammatical morphemes and established a set of guidelines for counting the number of morphemes in each utterance. Brown described five stages of development defined by intervals on the continuum of MLU scores, contending that specific aspects of syntactic development correlate with the number of morphemes used. Brown found that MLU was strongly correlated to chronological age and proposed that was predictive of the acquisition of morphemes assigned to each stage of development.

Subsequent studies substantiated the high positive correlation between chronological age and MLU (de Villiers & de Villiers 1973; Miller & Chapman, 1981; Miller 1991). The correlation between MLU and the acquisition of grammatical morphemes has also been verified (de Villiers & de Villiers, 1973; Klee & Fitzgerald, 1985; Rondal et al., 1986). However, several limitations and problems with MLU have also been identified. Chabon, Kent-Udolf, and Egolf (1982) found that MLU scores were unreliable for children beyond Brown's Stage V of development. Other findings challenge the validity of MLU beyond Stage II, at values of approximately 2.0 to 3.0 (Bennet-Kastor, 1988; Klee et al., 1989; Rondal et al., 1986).

Perhaps even more significant is the question of whether or not MLU is a valid measure of syntactic complexity at all. Klee and Fitzgerald (1985) examined the MLU scores and grammatical complexity of language samples obtained from18 children. Although the acquisition of grammatical morphemes did correlate with increases in MLU, changes in syntactic structure and diversity were not reflected. Klee and Fitzgerald concluded that MLU is not a good indicator of grammatical development in terms of syntactic construction. Perhaps MLU is not a sensitive measure of any linguistic construct other than utterance length itself (Klee et al., 1989). Miller (1991) also acknowledged that older children could increase the complexity of the system without increasing utterance length.

Language Assessment, Remediation, and Screening Procedure (LARSP)

Crystal, Fletcher, and Garman (1989) developed a qualitative procedure for grammatical analysis called LARSP. The descriptive framework of LARSP is based on seven stages of grammatical acquisition through which children pass. A 30-minute language sample is collected and analyzed on the word, phrase, clause, and sentence level. The frequency count of various structures at each level is tallied on a profile chart. A pattern of syntax is established by comparing several samples in order to establish an expected pattern (Crystal, 1982). Klee and Paul (1981) noted that LARSP yields an age score by giving some indication of acceptable variation around a general developmental stage. However, the measure has not been standardized and provides only raw data without conventions for summarization and interpretation.

Index of Productive Syntax (IPSyn)

The Index of Productive Syntax (IPSyn) was developed by Scarborough (1990) as an easily obtained summary scale of grammatical complexity to be used for the study of individual differences in language acquisition. A primary goal of the index is to provide numerical scores suitable for statistical analysis and standardization. IPSyn measures the emergence of syntactic and morphological structures in productive language. Scarborough developed IPSyn using 75 samples obtained longitudinally from 15 children. The first 100 successive, intelligible utterances in each sample were coded for 56 grammatical forms to develop the IPSyn score sheet. Data from the score sheet was used to derive a final IPSyn score. A comparison of mean IPSyn and MLU values at each age revealed that IPSyn is a reliable age-sensitive summary of grammatical complexity. Scarborough cautioned, however, that the index does not provide detailed diagnostic information about a child's mastery of specific structures and rules. Scarborough concluded that IPSyn is most suitable as a tool for comparing or matching subjects in research groups.

IPSyn has been applied in a variety of uses by researchers. In a comparative study involving autistic Down syndrome, and normal children, Tager-Flusberg and Calkins (1990) used IPSyn to investigate whether imitation is more advanced than spontaneous language. IPSyn was used to evaluate the grammatical content of the imitative and spontaneous corpora. An additional study of autistic and Down syndrome children used IPSyn as one of the comparative measures of language acquisition and development (Tager-Flusberg et al., 1990). Scarborough, Rescorla, Tager-Flusberg, Fowler, and Sudhalter (1991) examined the relationship between utterance length and grammatical complexity in normal and language-disordered children. IPSyn was used as the measure of syntactic and morphological proficiency and correlated to MLU scores for each group. Scarborough et al. found excellent agreement between IPSyn and MLU scores for children from 2 to 4 years old.

Developmental Sentence Scoring (DSS)

Development of DSS. Developmental Sentence Analysis (Lee & Cantor, 1971; Lee, 1974) was developed as a standardized method for making a quantified evaluation of a child's use of standard grammatical rules during spontaneous speech. The procedure involves two components: Developmental Sentence Types (DST) and Developmental Sentence Scoring (DSS). The DST chart is used to classify pre-sentence utterances containing only partial subject-verb grammatical structure, including single words, two-word combinations, and multiword constructions forming incomplete sentences. DSS is used for samples containing a majority of complete sentences comprised of a subject and a verb. The first version of DSS (Lee & Canter, 1971) introduced a developmental sequence of grammatical forms assigned a weighted score in eight categories. The DSS analysis scores a sample of 50 complete (noun and verb in subject-predicate form) sentences. Generally, the last 50 utterances from the sample are selected. Point values are assigned to grammatical forms in the eight categories. Incomplete and incorrect structures receive an "attempt" mark, but no score is given. An additional point is added to each sentence that meets all adult standard rules. A final DSS score is obtained by adding the total sentence scores from the sample and dividing by 50. Percentiles of DSS scores of 160 normally developing children from 3;0 to 6;11 were presented.

A subsequent publication by Lee (1974) presented the finalized version of DSS, including a re-weighted scoring procedure and detailed statistical data for 40 children from 2;0 to 2;11. The re-weighted procedure was also performed on the original 160 samples, bringing the total to 200 children from 2;0 to 6;11. The reassignment of weights of the structures at developmental intervals allowed for comparisons not only within grammatical categories, but across categories as well. Lee suggested that the DSS of an individual child could be compared with normative data collected for normally developing children of the same chronological age. A child's DSS performance can also be judged against the mean of a lower age group in order to estimate the degree of language delay in months. An additional function of DSS is to plot a child's scores over time in order to measure the rate of progress during language intervention. Lee acknowledged, however, that diagnosis should never be made on the basis of DSS scores alone, nor should a child's DSS score be used to make broad assumptions about his language development.

DSS Validity. Leonard (1972) offered a comprehensive description of deviant language, which included the use of DSS in the comparison of children with deviant and normal language. Leonard compared samples from nine children with normal language to nine matched children with deviant language. Leonard's findings indicated that differences between deviant and normal speakers were not qualitative, but rather, quantitative in terms of frequency of usage of deviant forms and structures. Leonard concluded that DSS is a useful measure of syntactic development "equipped with an abundance of empirical support" (p. 428) and may be the most effective means to distinguish between deviant language requiring clinical attention and more minor language delays.

A series of investigations of the validity of the DSS procedure were performed by Koenigsknecht (1974) as part of the finalized version of DSS. Koenigsknecht reported that the validity of DSS scores was "indicated by significant differences produced among successive age groups of normally developing children" (p. 223). A cross-sectional study of 200 children ages 2;0 to 6;11 revealed significant differences in syntactic structures and consistent increases in DSS scores between all successive age levels. Results confirmed the grammatical hierarchy and weighting system of the final DSS procedure (Lee, 1974).

The issue of language delay versus language deviance was further explored using DSS. Rondal (1978) analyzed samples from14 normal and 14 MLU-matched children with Down syndrome. DSS results revealed that children with language impairments due to Down syndrome tended to demonstrate less syntactic sophistication than their normal peers. Findings indicated quantitative differences in the frequency of use of syntactic structures between the two groups, substantiating Leonard's (1972) conclusion that DSS is sensitive to the distinction between language deviance and delay.

This notion was further supported by Liles and Watt (1984) in a study comparing 12 males judged to have communication impairment and 12 MLU-matched males with normal linguistic performance. A 100-utterance sample from each child was collected and analyzed using DSS. Although overall DSS scores between the two groups were not significantly different, a multiple discriminate analysis showed that individual differences within nine variables (the eight grammatical categories and the number of sentence points) were significant when operating together. Liles and Watt found that seven of the variables (excluding indefinite pronouns and Wh-questions), when considered together, contributed significantly to the ability of DSS to discriminate between normal and communicatively impaired children.

DSS Reliability. Koenigsknecht's (1974) examination of the final version of DSS also included three aspects of reliability: stimulus material differences, temporal reliability, and sentence sequence effects. The preschool children in the original DSS research were used as subjects in all three probes. The use of different stimulus materials resulted in changes in four individual categories (indefinite pronouns, personal pronouns, secondary verbs, and interrogative reversals), but overall DSS scores were not significantly affected. A longitudinal analysis of temporal reliability involved four repeated applications in a two-week period, three repeated applications at four-month intervals, and rank ordering of the DSS scores across six applications. Significant increases in overall DSS scores were noted across all applications in the two-week and four month intervals. However, the changes were in harmony with developmental patterns and increases were consistent among subjects. In order to analyze the sentence sequence effects, the first 25 sentences in each sample were compared with the last 25. Analysis of 60 samples yielded no statistically significant difference in overall or individual category DSS scores. Koenigsknecht concluded that results from the three probes support the stability and reliability of the DSS procedure.

Recognizing that the reliability of a measure increases as the sample size increases, Johnson and Tomblin (1975) sought to estimate the reliability of DSS using the recommended 50-utterance sample size. Twenty-five sentences were randomly selected from 50-sentence samples obtained from 50 children between the ages of 4;8 and 5;8. Sentences were analyzed according to DSS procedures to obtain overall and component scores. Using an analysis of variance approach, the reliability of DSS was estimated for sample sizes of five to 250. As predicted, the reliability for all scores increased with larger sample sizes. Reliability of total DSS scores for 50 sentences was reported to be only 0.75. Johnson and Tomblin suggested that a larger sample, perhaps as high as175 sentences, is required to obtain acceptable levels of reliability. The authors acknowledged the difficulty of collecting samples of such size and therefore concluded that DSS should not be used to discriminate disordered from normal language. Rather, it should be used only to identify specific areas of syntactic concern in individual cases.

Applications of DSS. DSS has been used for a variety of research purposes. Blaxley, Clinker, and Warr-Leeper (1983) used DSS to assess the accuracy of two screening tools for language impairment, while Johnston and Kamhi (1984) applied the DSS procedure in their investigation of the syntactic and semantic patterns in children with language impairment. Klee (1985) pointed out the usefulness of DSS in establishing linguistic baselines for deriving intervention goals. Variations of the DSS procedure have also been adapted for use with different populations, including Spanish-speaking children (Toronto, 1976), older children up to age 9;11 (Stephens, Dallman, & Montgomery, 1988), and speakers of Black English (Nelson & Hyter, 1990).

The value of DSS has been proven during more than 20 years of clinical use. Lively (1984) observed that DSS is a popular and widely used method in evaluating the syntactic and morphological development of children. Lively noted that deriving full clinical benefit from DSS is dependent on the correct use and application of the procedure. She identified common scoring errors and emphasized the importance of proper education and training of clinicians. In addition, Lively reiterated Lee’s (1974) caution against using DSS as the sole means of evaluation. Despite its shortcomings, DSS has weathered criticism and maintained its place in clinical practice (Hughes et al., 1992). Surveys have revealed that DSS is the most widely used form of standardized analysis used by speech language pathologists practicing language sampling (Hux et al., 1993; Kemp & Klee, 1997).

Automated Language Sample Analysis

The development of computer technology has provided researchers and clinicians with new means of decreasing demands of time and resources required for language sample analysis. Several programs have been developed to perform analysis of text files, including *Automated* LARSP (Bishop, 1984), *Systematic Analysis of Language Transcripts* (Miller & Chapman, 1990), and *Computerized Profiling* (Long & Fey, 1993). Long (1991), acknowledging time as the most valuable commodity for a clinician, examined the contribution of computers in promoting efficiency and simplifying the process of clinical language analysis. Computers can provide assistance in the collection and analysis of the sample and the interpretation of the data. Long outlined the necessary steps performed in all analysis programs: (a) utterances are coded by the clinician to identify grammatical or phonological structures, (b) the program recognizes, analyzes, and tabulates information in the sample, and (c) results of the analysis are presented for interpretation. Long cautioned it remains the responsibility of the clinician to derive information from the data and make assessment decisions.

The public school system has been a particular target for implementing computer-assisted language sampling (Miller, Freiberg, Rolland, & Reeves, 1992). Miller et al. identified obstacles toward widespread language sampling in schools, including the lack of consistent transcription formats and standardized analysis procedures, and the lack of normative databases of measures from typically developing children for comparative purposes. Miller et al. suggested that automated analysis procedures can assist in overcoming these problems.

Several programs have attempted to use computer technology to perform DSS analysis. Klee and Sahlie (1986) reviewed the first computer-assisted DSS software, a program developed by Hixson in 1983. *Computerized DSS* was designed to reduce the time needed for analysis by automatically tallying the points manually assigned by a clinician. An Attempt Score and an Error Score are also computed for comparison against the standardized normative data. Klee and Sahlie addressed two specific weaknesses of the program. First, ambiguous lexical items are not recognized by the program and accurate analysis is dependent on the precision of the manual transcription. Second, several errors and omissions, including discrepancies with the original DSS chart, were noted in the output from the computer application.

Later computer programs were developed to perform fully automated language sample analyses, including DSS. These programs require a specific format for transcriptions, but clinician pre-coding for DSS is not necessary. CLAN is part of the Child Language Data Exchanges System (MacWhinney, 1991), a software package and database available on the Internet. CLAN performs over 20 language sample analysis procedures, including DSS, MLU, and simple frequency counts. Formal research on the accuracy and efficiency of CLAN DSS analysis has not been published.

Computerized Profiling (Long & Fey, 1988, 1993) is another automated application created to foster greater clinical use of language sampling by alleviating some of the accompanying time demands. The program includes six modules: the CORPUS module for formatting the transcript and five analysis modules, including automated LARSP and DSS. In order for the DSS analysis to be performed, the transcript must first be run through the LARSP module. In a review of the LARSP module of CP, Klee and Sahlie (1987) found the program to be easy to learn. However, the reviewers found that the software generated errors requiring correction by the user, largely negating the timesaving advantage. The review did not include an evaluation of the DSS module. Baker-Van Den Goorbergh (1994) made similar criticism of the LARSP module of CP, claiming that it incorrectly analyzed most of the utterances input by the reviewers.

Long and Fey (1995) responded to the criticisms delineated by Baker-Van Den Goorbergh, stating that the findings were inaccurate and undocumented beyond the author's personal experience. Long and Fey maintained that Baker-Van Den Goorbergh's description of data analysis neglected key modules of the programs, rendering her evaluations incomplete. Further, Long and Fey argued that although automated coding procedures do generate mistakes, these potential errors do not reverse the overall benefits of using computer programs. The clinician still reviews the output and maintains control over the final analysis, while retaining the advantage of increased speed and efficiency. A later review (Gregg & Andrews, 1995) substantiated this position. In an examination of the efficiency and accuracy of the DSS module, Gregg and Andrews noted that the accuracy of the DSS analysis is dependent on the accuracy of the LARSP output. Therefore, as with the LARSP module, the DSS analysis must be reviewed by the clinician. The authors proposed that although corrections require additional time, clinicians with a knowledge of LARSP and DSS who use these modules regularly can complete the corrections in less time than required for manually analysis.

An unpublished master's thesis by Boyce (1995) investigated the accuracy of automated DSS analysis performed by CP and CLAN software. The first 200 utterances of 75 language samples from the CHILDES archive were analyzed using standard DSS procedures. Automated analysis was performed on the same samples using both CP and CLAN. Findings indicated that accuracy varied from 0% to 94% among the individual categories and between the two programs. Boyce suggested that the high variability in both programs warrants further research and refinement before the software can perform fully automated language sample analysis.

In addition to decreasing the time and energy required to perform actual language sample analysis, computers have also been used to lessen the time required to train clinicians in DSS analysis. Hughes, Fey, Kertoy, and Nelson (1994) developed a computer-assisted instruction program to for learning DSS. Fifty-five graduate students from three universities participated in a study of the DSS tutorial. All subjects received an introductory lecture and a pre-test, followed by 8 weeks of training. Twenty-six students received traditional classroom-based instruction, while twenty-nine used the computer-assisted tutorial. Results indicated that students in both groups achieved comparable levels of proficiency for clinical use of DSS. The computer-assisted program, however, required significantly less time for both instructors and students. Hughes et al. concluded that computer-assisted instruction is valuable in "enhancing the efficiency and effectiveness of instruction in the analysis of children's language samples" (p. 94).

Method

Participants

In this study, three subsets of previously collected language samples were used. The total corpus used consists of 80 samples containing approximately 18,400 utterances. Samples were obtained from 50 typically developing children and 30 children with language impairment. A total of 14,117 DSS-analyzable utterances were extracted from the entire corpus.

Reno Samples. Thirty samples collected by Fujiki, Brinton, and Sonnenberg (1990) in Reno, Nevada were used. Approximately 8,700 utterances were obtained from 30 samples. A total of 6,889 utterances were extracted for analysis. The participants included 10 children with language impairment (LI), 10 language matched children (LA), and 10 chronological age matched children (CA). The LI children ranged in age from 7;6 to 11;1 years and were all receiving language intervention by a school-based speech-language pathologist. All LI children exhibited comprehension and production deficits, scoring at least one standard deviation below the mean on two formal tests. Each LI child was matched to a LA child, ranging from 5;6 to 8;4 years, on the basis of a language age score within 6 months of the impaired child performance on the Utah Test of Language Development (Mecham, Jex, Jones, 1967). Each LI child was also matched to a CA child (within 4 months of age of the LI match) from the same elementary school. The CA group ranged in age from 7;6 to 11;2 years.

Jordan Samples. Twenty samples containing approximately 3,700 utterances from children with LI were collected from Jordan School District in Utah (Collingridge, 1998). A total of 2,394 utterances were extracted for analysis. The participants consisted of 11 female and 9 male English-speaking children between the six and ten years of age. All children were considered by a speech-language pathologist to have language impairment. All children were required to have at least 80% intelligibility and adequate language skills to actively participate in conversation. At the time the samples were collected, all 20 children were receiving pull-out intervention or services in self-contained communication or learning disorders classrooms.

Wymount Samples. Channell and Johnson (1999) used 30 previously collected samples of typically developing children. Approximately 6,000 utterances were obtained during naturalistic interactions between each child and one of three graduate students enrolled in a master's program in speech-language pathology. A total of 4,835 utterances were extracted for analysis. All subjects were native English speakers residing in Provo, Utah with no history of language or hearing impairment. The children ranged in age from 2;6 to 7;11, with 3 children in each six-month interval.

DSS Analysis

Manual DSS analysis followed established procedural guidelines (Lee, 1974). Only samples in which at least fifty percent of utterances were complete (i.e. utterances containing a subject and a predicate) were included in the corpus. A total of at least 50 utterances were analyzed from each sample; however, one sample (Jordan sample #7) was later found to contain only 48 analyzable utterances. The utterances were formatted using the following standards: (a) mazes, repetitions, revisions, and interjections were placed in parentheses and not analyzed, (b) punctuation was used at the end of each utterance, and (c) only proper nouns and the pronoun *I* were capitalized. Grammatical forms from the eight standard DSS categories were scored in each utterance. An additional Sentence Point was awarded to sentences meeting all adult standard rules. Attempt marks receiving no score were assigned to structures not meeting the requirements of adult Standard English. A mean sentence score was derived by totaling the individual sentence scores and dividing by the total number of utterances analyzed.

I performed manual DSS analysis on all samples included in the corpus. Interrater reliability was established by having a second clinician analyze 10% of the total samples. Agreement was required for both grammatical categorization and developmental complexity. Results were correlated to my analyses and found to be in 97% agreement.

MLU Analysis

Manual MLU analysis was based on the morpheme-count procedure described by Brown (1973). Utterances in a sample meeting the following criteria were used for analysis: (a) only fully transcribed utterances were used, (b) only the most complete form of a repeated word was counted, (c) fillers such as *um* or *oh* were omitted, (d) all compound words, proper names, and ritualized reduplications were counted as single words, (e) irregular past tense verbs were counted as one morpheme, (f) diminutive forms were counted as one morpheme, and (g) all auxiliaries and catenatives were counted as one morpheme. In addition, only utterances meeting the qualifications for DSS analysis were included in the MLU analysis. An MLU score was obtained for each sample by averaging the individual morpheme count for all analyzed utterances.

I calculated the MLU on all samples included in the corpus. Interrater reliability was established by having a second clinician analyze 500 utterances randomly selected from the set of samples; our MLU counts agreed on 98% of these utterances.

SSD Software Analysis

Automated analysis of the samples was performed using the SSD software. The software analyzes the grammatical forms in utterances extracted from naturalistic samples of children's expressive language and computes a score based essentially on the mean frequencies of the same items scored by DSS.

Purpose of SSD. The SSD index is designed to be a norm-referenced measure comparable to DSS, IPSyn, and MLU. As with DSS and MLU analysis, SSD analysis requires that utterances be formatted using standardized guidelines. However, unlike automated versions of existing measures, SSD is entirely automated and does not require any manual pre-coding.

File Format. The software employs the same file format used in Computerized Profiling (Long, Fey, & Channell, 2000). The format includes the following guidelines: (a) conventional English spelling is used; however, semi-auxiliaries (e.g. gonna) can be transcribed as spoken, (b) only one utterance per line, (c) all utterances are in lower case except for proper nouns, (d) any revisions, repetitions, and interjections are placed in parentheses, and (e) any entire utterance to be skipped is prefaced by a non-alphanumeric character.

File Processing. The program consists of two modules. Utterances are input into the first module where they are grammatically tagged using a tagging scheme adapted from the LARSP approach of Crystal et al. (1989). Each word in the utterance receives an appropriate grammatical tag such as: *he* <PP *has* <V.z *a* <D *fever* <N. The grammatical tags are then used to generate a sentence syntactical development analysis (SSD) patterned after DSS (Lee, 1971). The software can process approximately 100 utterances per second. Data obtained from the utterance-by-utterance analysis is used in a second module to generate a total index score.

Procedure

A manual utterance-by-utterance analysis was performed on each sample in the corpus to obtain a DSS score and a MLU score. Each sample was formatted according to guidelines for Computerized Profiling, with the following additional levels of coding: (a) a level beginning with #d containing manual DSS codes, and (b) a level beginning with #m containing manual MLU totals. Each sample was coded in the following format:

> *I like to color too.*
> #d p1 m1 s5 +
> #m 5

Each utterance was then run through the automated software to obtain an SSD score. The SSD analysis generates two additional levels of coding, grammatical tagging (#g) and SSD (#s). Output for each utterance is coded in the following format:

> *I like to color too.*
> #g I <PP like <V to <TO color <V too <AV . <.
> #s p1 m1 s5
> #d p1 m1 s5 +
> #m 5

Each sample was run through the second module to obtain total scores for the three indices, SSD, DSS, and MLU.

Pearson's *r* correlations were performed on the three data points, SSD, DSS, and MLU scores extracted from each sample. Correlations were tabulated between SSD and DSS, SSD and MLU, and DSS and MLU for each corpus.

Results

Reno Corpus

The SSD, DSS, and MLU scores for each sample in the Reno corpus were calculated and are presented in Table 1. Results in Table 1 show that SSD scores ranged from 4.86 to 12.36 with an average of 8.75 (*SD* = 1.86). DSS scores ranged from 4.25 to 13.33 with an average of 9.42 (*SD* = 2.26). MLU scores ranged from 5.47 to 9.92 with an average of 7.77 (*SD* = 1.16).

Pearson's *r* correlations among these scores revealed SSD and DSS to be highly

Table 1

Descriptive Statistics on the Reno Samples

Child	N Utterances	SSD	DSS	MLU
R1	279	8.97	10.07	8.70
R2	210	9.51	10.41	7.52
R3	130	7.62	8.00	7.43
R4	284	8.61	9.29	7.39
R5	136	6.44	6.51	7.84
R6	188	12.12	13.07	9.44
R7	187	7.73	8.49	7.30
R8	249	11.96	12.80	9.57
R9	166	8.32	8.72	8.05
R10	273	8.15	9.03	7.4
R11	78	5.97	4.90	6.33
R12	307	9.38	10.28	8.68
R13	331	9.88	11.18	7.94
R14	203	10.03	10.71	8.69
R15	186	7.81	8.97	6.68
R16	138	7.42	8.04	6.56
R17	297	9.19	10.20	7.23
R19	239	11.02	11.93	9.15
R20	193	6.74	7.20	6.24
R21	337	7.58	8.65	6.40
R22	239	8.36	9.21	6.97
R23	398	8.85	9.89	7.23
R24	290	9.86	10.69	8.36
R25	301	7.40	8.08	7.02
R26	193	10.31	11.40	9.18
R27	247	7.78	8.65	8.09
R28	214	12.36	13.33	9.92
R29	146	6.79	6.23	6.62
R30	118	4.86	4.25	5.47

correlated (*r* = .98). Both measures were also correlated with MLU, finding SSD correlated with MLU at *r* = .89 and DSS correlated with MLU at *r* = .86. All three correlations were statistically significant (*p* < .0001), suggesting only a slight probability that such similarities are a result of chance.

The three measures were also separately analyzed for each of the three subgroups in the Reno corpus. The means and standard deviations for each measure are presented in Table 2. The means of the CA group were higher than those and the LA group, and the means of the LA group were higher than the LI group on all three measures. However, it can be seen that the standard deviations of the group scores are larger than the differences between the group means. These scores were compared using one-way analysis of variance tests; no significant differences between the means were observed.

Jordan Corpus

The SSD, DSS, and MLU scores for each sample in the Jordan corpus are presented in Table 3. It can be seen in Table 3 that SSD scores ranged from 4.31 to 9.17 with an average of 7.15 (*SD* = 1.27). DSS scores ranged from 4.72 to 10.12 with an average of 7.75 (*SD* = 1.48). MLU scores ranged from 4.60 to 7.97 with an average of 6.43 (*SD* = 0.89).

Pearson's correlations among these scores showed SSD and DSS to be highly correlated, *r* = .92 (*p* < .0001). Both measures were also correlated with MLU, finding SSD correlated with MLU at *r* = .69 (*p* = .0005) and DSS correlated with MLU at *r* = .62 (*p* = .0028).

Wymount Corpus

The SSD, DSS, and MLU scores for samples in the Wymount corpus are

Table 2

Descriptive Statistics on the Reno Subgroups

	SSD		DSS		MLU	
Group	*M*	*SD*	*M*	*SD*	*M*	*SD*
CA	8.94	1.83	9.64	2.05	8.07	0.86
LA	8.88	1.81	9.60	2.34	7.71	1.26
LI	8.42	2.08	9.04	2.57	7.52	1.35

Table 3

Descriptive Statistics on the Jordan Samples

Child	N Utterances	SSD	DSS	MLU
J1	150	8.06	8.32	7.32
J2	129	8.27	8.87	6.43
J3	97	6.07	6.16	6.28
J4	128	8.84	10.12	7.42
J5	99	6.98	7.75	6.96
J6	137	6.93	7.33	6.33
J7	48	5.42	6.06	5.42
J8	105	7.99	9.14	6.41
J9	180	7.81	8.49	6.58
J10	121	6.44	7.72	5.69
J11	98	4.96	5.80	5.15
J12	134	7.42	8.73	6.67
J13	86	4.31	4.72	5.22
J14	179	7.80	9.16	7.13
J15	86	7.36	7.19	7.97
J16	142	7.08	6.95	7.51
J17	186	9.17	9.45	7.10
J18	109	8.51	9.08	6.49
J19	105	6.39	5.59	4.60
J20	75	7.24	8.35	5.95

presented in Table 4. SSD scores ranged from 4.73 to 13.09 with an average of 8.35 (*SD* = 2.11). DSS scores ranged from 4.34 to 14.60 with an average of 9.26 (*SD* = 2.34). MLU scores ranged from 4.28 to 10.61 with an average of 6.62 (*SD* = 1.63).

Pearson's correlations among these scores showed SSD and DSS to be highly correlated (r = .98). Both measures were also correlated with MLU, finding SSD correlated with MLU at r = .94 and DSS correlated with MLU at r = .91. All three correlations were statistically significant ($p < .0001$).

Given the wide age range of children in the Wymount corpus (2;6 to 7;11), some of the correlation among measures may be simply a result of the correlation that each measure shared with age. Partial correlations were therefore used to examine correlation among measures independent of the measures' correlation with age. The correlation between SSD and DSS remained strong (r = .91, $p < .0001$). However, the correlation of SSD with MLU decreased (r = .61, p = .0002) and the correlation of DSS with MLU changed direction and no longer reached statistical significance (r = -.28, $p > .05$).

Discussion

A comparison of manual DSS and MLU procedures with the automated SSD analog resulted in significant correlations among the measures. The DSS and SSD scores were highly correlated in all three corpora, as well as the subgroups of normal children and children with language impairments in the Reno corpus. It should be noted, however, that SSD scores tended to be slightly lower than DSS scores (typically about a 0.5 point difference). These differences in the absolute magnitude of the scores can be attributed to the fact that the computational rules of the two indices are different. In this study, no attempts were made to identify exact scoring differences within each utterance, thus

Table 4

Descriptive Statistics on the Wymount Samples

Child	N Utterances	SSD	DSS	MLU
W1	145	6.22	6.95	4.97
W2	199	12.20	13.70	9.19
W3	163	8.20	9.20	6.34
W4	188	7.36	8.28	6.52
W5	164	8.02	9.38	6.51
W6	142	6.40	6.55	4.73
W7	132	6.20	6.62	5.17
W8	139	7.25	8.59	5.36
W9	158	6.44	7.82	5.82
W10	164	8.23	9.27	5.79
W11	197	13.09	14.60	10.17
W12	191	11.46	12.05	10.02
W13	67	9.00	10.07	6.88
W14	140	10.54	11.49	7.83
W15	163	6.58	7.79	5.34
W16	187	9.35	9.94	7.44
W17	161	10.31	11.05	6.73
W18	164	8.20	9.51	6.45
W19	149	7.42	8.74	5.76
W20	101	4.73	4.34	4.84
W21	150	9.24	9.75	6.75
W22	164	6.42	6.68	5.57
W23	182	8.77	9.71	7.05
W24	148	8.95	10.24	6.38
W25	196	12.02	12.95	10.61
W26	166	6.33	7.74	5.25
W27	117	5.79	6.27	4.28
W28	155	7.39	8.24	6.25
W29	178	10.65	12.12	8.11
W30	183	7.87	8.09	6.60

variations in the treatment of specific grammatical categories resulting different total scores between SSD and DSS have not been identified.

Correlations with DSS and SSD to MLU were moderately high, but not as high as those between DSS and SSD. These results are not unexpected, as the procedures for DSS and SSD share more similarities with one another than either measure does with MLU. In addition, the majority of samples included in the three corpora were collected from older school-aged children. There is evidence suggesting that beyond age three (typically MLU values of 3.0 to 4.0 in normally developing children) MLU is not a valid predictor grammatical complexity (Klee & Fitzgerald, 1985; Klee et al., 1989; Rondal et al., 1986). The present study does not consider syntactic complexity; rather, it simply looks at numeric score correlations between the three measures.

Although correlations of DSS and SSD with MLU were only moderate, even these correlations are higher than levels obtained in previous studies comparing various measures purporting to assess a specific language domain. For example, Channell and Peek (1989) compared four similar measures of vocabulary ability in preschool children and found only moderate associations, suggesting a significant lack of agreement among the measures. In a separate study of four grammatic completion measures, Channell and Ford (1991) found moderate to high correlations, with results slightly lower than those obtained in this study. A comparison of existing research to the current findings suggests that the three measures examined are at least as comparable to one another, if not more so, as analogous measures in other domains.

It should be noted that these findings are subject to the limitations of the present study. The school-aged children in the three corpora were typically older than the children included in the original DSS research (Lee, 1974), introducing the possibility of age-related variability in the results. In addition, the design of this study does not control for any differences among the three groups of samples. There are differences in sample size and collection procedures among the three corpora. For example, the Jordan samples are significantly shorter than the samples in the other two groups, which may account for the lower correlations obtained for the Jordan corpus.

The high correlation between SSD and DSS is a promising indicator that the software analog parallels DSS in scope and function, suggesting that SSD could eventually be used clinically to replace manual DSS. However, the correlational analysis performed constitutes only a preliminary exploration of the utility of the SSD software. At the current time it would be premature to apply SSD clinically. Additional research is needed to investigate the psychometric characteristics of the new measure. Due to the similarity between the two measures, it is possible that some of the critiques against DSS may hold up against SSD as well. Criticism regarding sample size, sampling variability, temporal stability, and the validity of the developmental sequence have been raised against DSS (Bennett-Kastor, 1988; Bloom and Lahey, 1978; Johnson & Tomblin, 1975; Klee & Sahlie, 1986). These issues must be investigated relative to SSD as well. The test-retest reliability of SSD must be studied, particularly as a function of sample size. Some studies show DSS to be sensitive to differences between disordered and non-disordered language (Hughes et al., 1992; Lee, 1974; Leonard, 1972; Liles & Watt, 1984). Since SSD correlates so highly to DSS, it is reasonable to suggest that it would be at least as useful as DSS in this regard. However, further investigation of the ability of SSD to discriminate between normal and disordered language is warranted. Finally, since the computational rules of SSD differ from those outlined in DSS, the normative data compiled by Lee (1994) cannot be applied to SSD with validity. New normative data must be collected specific to the SSD software.

Language sample analysis has long been recognized as an important tool in the clinical assessment of children's productive language. Although the value of language sampling is widely accepted, the actual implementation of analysis procedures is far less prevalent. Issues such as inter-scorer reliability, clinician training, and time and resource demands tend to limit the practical value of existing manual procedures. The use of computer technology can reduce or eliminate some of the difficulties associated with manual language sampling. Long (1991) outlined several advantages of computer-assisted analysis, including increased speed and accuracy of quantification and analysis, long-term storage of transcripts, and multiple analyses of a single transcript. Current findings show these advantages holding true for the SSD software application.

Unlike MLU or DSS, SSD requires that a sample be transcribed into computer format. However, the time required for input is substantially offset by the benefits the software ultimately offers. SSD generates rapid, fully automated quantification of grammatical development, decreasing the time demands placed on clinicians. The automated nature of the measure has the added benefit of consistency of analysis across clinicians, removing problems of inter-scorer reliability. In addition, the computer formatting utilized by the SSD software provides easy, convenient storage and retrieval of large transcripts. Samples can be used for more descriptive analysis after being run through the software. Previously collected and analyzed samples can also be reprocessed using future versions of the software for the purpose of comparison across time. For example, a baseline sample collected from a child can be compared to a more recent sample to measure progress over time, a practice that cannot be validly performed with different versions of manual measures and tests.

As future research is conducted and the program is refined, SSD has the potential to provide a much-needed alternative to existing measures of grammatical development such as DSS. In addition to providing greater speed and accuracy, the fully automated nature of the program eliminates the need for extensive procedural training of clinicians. Rather, clinician skills can be utilized for more descriptive analysis and interpretations of results produced by automated analysis. The potential advantages of SSD could provide an incentive for clinicians to incorporate language sampling into the comprehensive evaluation of the productive language development of children, thus enhancing the quality of clinical assessment.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, D. C.: American Psychological Association.

Baker-Van Den Goorbergh, L. (1994). Computers and language analysis: theory and practice. *Child Language Teaching and Therapy, 10*, 329-348.

Baker-Van Den Goorbergh, L., & Baker, K. (1991). *Computerized Language Error Analysis Report (CLEAR)*. Kibworth, Leics: FAR Communications.

Bennett-Kastor, T. (1988). *Analyzing children's language: Methods and theories.* New York: Basil Blackwell.

Bishop, D. V. M. (1984). Automated LARSP; computer-assisted grammatical analysis. *British Journal of Disorders of Communication, 19,* 78-87.

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

Blaxley, L., Clinker, M., & Warr-Leeper, G. (1983). Two language screening tests compared with Developmental Sentence Scoring. *Language, Speech, and Hearing Services in Schools, 14,* 38-46.

Boyce, L. L. (1995). *Accuracy of automated Developmental Sentence Scoring.* Unpublished master's thesis, Brigham Young University, Provo, UT.

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 and Hearing Research, 25*, 124-128.

Channell, R. W., & Ford, C. T. (1991). Four grammatic completion measures of language ability. *Language, Speech, and Hearing Services in Schools, 22,* 211-218.

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

Channell, R. W., & Peek, M. S. (1989). Four measures of vocabulary ability compared in older preschool children. *Language, Speech, and Hearing Services in Schools, 20*, 407-420.

Collingridge, J. D. (1998). *Comparison of DSS Scores from On-Line and Subsequent Language Sample Transcriptions.* Unpublished master's thesis, Brigham Young University, Provo, UT.

Crystal, D. (1982). *Profiling Linguistic Disability*. London: Edward Arnold.

Crystal, D., Fletcher, P., & Garman, M. (1989). *The Grammatical Analysis of Language Disability: a Procedure for Assessment and Remediation* (2nd ed.). London: Cole and Whurr.

de Villiers, J. G., & de Villiers, P. A. (1973). A cross-sectional study of the acquisition of grammatical morphemes in child speech. *Journal of Psycholinguistic Research, 2*(3), 267-278.

Fletcher, P. (1985). *A child's learning of English*. Oxford: Blackwell.

Fujiki, M., Brinton, B., & Sonnenberg, E. A. (1990). Repair of overlapping speech in the conversations of specifically language-impaired and normally developing children. *Applied Psycholinguistics, 11,* 201-215.

Gregg, E. M., & Andrews, V. (1995). Review of Computerized Profiling. *Child*

Language Teaching and Therapy, 11, 209-216.

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.

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.

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.

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.

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.

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

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.

Klee, T., & Paul, R. (1981). A comparison of six structural analysis procedures: A case study. In J. F. Miller (Ed.), *Assessing language production in children (pp. 73-110).* Austin, TX: PRO-ED.

Klee, T., & Sahlie, E. (1986). Review of DSS computer program. *Child Language Teaching and Therapy, 2*, 97-100.

Klee, T., & Sahlie, E. (1987). Review of Computerized Profiling. *Child Language Teaching and Therapy, 3*, 87-93.

Klee, T., Schaffer, M., May, S., Membrino, I., & Mougey, K. (1989). A comparison of the age-MLU relation in normal and specifically language-impaired preschool children. *Journal of Speech and Hearing Disorders, 54*, 226-233.

Koenigsknecht, R. A. (1974). Statistical information on development sentence analysis. In L. L. Lee *Developmental sentence analysis: A grammatical assessment procedure for speech and language clinicians*. (pp. 222-268). Evanston, IL: Northwestern University Press.

Lahey, M. (1988). *Language Disorders and Language Development.* Needham, MA: Macmillan Publishing Company.

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

Lee, L. L., & Canter, S. M. (1971). Developmental Sentence Scoring: A clinical procedure for estimating syntactic development in children's spontaneous speech. *Journal of Speech and Hearing Disorders, 36,* 315-340.

Leonard, L. B. (1972). What is deviant language? *Journal of Speech and Hearing Disorders, 37(4),* 427-446.

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

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

Long, S. H. (1991). Integrating microcomputer applications into speech and language assessment. *Topics in Language Disorders*, *11(2)*, 1-17.

Long, S. H., & Fey, M. E. (1993). *Computerized Profiling.* The Psychological Corporation.

Long, S. H., & Fey, M. E. (1995). Clearing the air: A comment on Baker-Van Den Goorbergh (1994). *Child Language Teaching and Therapy, 11,* 185-192.

MacWhinney, B. (1991). *The CHILDES project: Tools for analyzing talk.* Hillsdale, NJ: Lawrence Erlbaum Associates.

McCauley, R. J., & Swisher, L. (1984a). Psychometric review of language and articulation tests for preschool children. *Journal of Speech and Hearing Disorders, 49,* 34-42.

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

Mecham, M., Jex, J. L., & Jones, J. D. (1967). *Utah Test of Language Development*. Salt Lake City, UT: Communication Research Associates.

Menyuk, P. (1964). Comparison of grammar of children with functionally deviant and normal speech. *Journal of Speech and Hearing Research, 7,* 109-121.

Miller, J. F. (1981). *Assessing language production in children*. Austin, TX: PRO-ED.

Miller, J. F. (1991). *Research on child language disorders: A decade in progress.*

Austin, TX: PRO-ED.

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

Miller, J. F., & Chapman, R. S. (1990). *Systematic analysis of language transcripts (SALT) Version 1.3 (MS-DOS).* [Computer program]. Madison, WI: Language Analysis Laboratory. Waisman Center on mental Retardation and Human Development.

Muma, J. R. (1998). *Effective Speech-Language Pathology: A cognitive socialization approach.* Mahwah, NJ: Lawrence Erlbaum Associates.

Muma, J. R., Pierce, S., & Muma, D. (1983). Language training in ASHA: A survey of substantive issues. *Asha, 35,* 35-40.

Nelson, N., & Hyter, Y. (1990, November). How to use Black English Sentence Scoring for non-biased assessment. Short course presented at the convention of the American Speech-Language Hearing Association, Seattle, WA.

Richards, B. (1987). Type/Token Rations: What do they really tell us? *Journal of Child Language, 14*, 201-209.

Rondal, J. A. (1978). Developmental sentence scoring procedure and the delay-difference question in language development of Downs Syndrome children. *Mental Retardation, 16,* 169-171.

Rondal, J. A., Ghiotto, M., Bredart, S., & Bachelet, J. (1986). Age-relation, reliability, and grammatical validity of measures of utterance length. *Child Language, 14,* 433-446.

Scarborough, H. S. (1990). Index of Productive Syntax. *Applied Psycholinguistics*, *11*, 1-22.

Scarborough, H. S., Rescorla, L., Tager-Flusberg, H., Fowler, A. E., & Sudhalter, V. (1991). The relation of utterance length to grammatical complexity in normal and language-disordered groups. *Applied Psycholinguistics, 12*, 23-45.

Stephens, I., Dallman, W., & Montgomery, A. (1988, November). Developmental sentence scoring through age nine. Paper presented at the annual convention of the American Speech-Language Hearing Association, Boston, MA.

Tager-Flusberg, H., & Calkins, S. (1990). Does imitation facilitate the acquisition of grammar? Evidence from a study of autistic, Down's syndrome and normal children. *Journal of Child Language*, *17*, 591-606.

Tager-Flusberg, H., Calkins, S., Nolin, T., Baumberger, T., Anderson, M., & Chadwich-Dias, A. (1990). A longitudinal study of language acquisition in autistic and Down syndrome children. *Journal of Autism and Developmental Disorders, 20*, 1-21.

Templin, M. C. (1957). *Certain language skills in children: Their development and interrelationships*. Minneapolis: The University of Minnesota Press.

Toronto, A. S. (1976). Developmental assessment of Spanish grammar. *Journal of Speech and Hearing Disorders, 41,* 150-171.