Photocopy and Use Authorization

In presenting this thesis in partial fulfillment of the requirements for an advanced degree at Idaho State University, I agree that the Library shall make it freely available for inspection. I further state that permission for extensive copying of my thesis for scholarly purposes may be granted by the Dean of the Graduate School, Dean of my academic division, or by the University Librarian. It is understood that any copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Signature _____

Date _____

Relationships Between Norm-Referenced Test Scores and Narrative Language Sample Measures in School-Aged Children with Specific Language Impairment

by

Amy Rae Smith

A thesis

submitted in partial fulfillment

of the requirements for the degree of

Master of Science in the Department of Communication Sciences & Disorders

Idaho State University

Summer 2018

To the Graduate Faculty:

The members of the committee appointed to examine the thesis of Amy Rae Smith find it satisfactory and recommend that it be accepted.

Diane Ogiela, Ph.D., CCC-SLP
Major Advisor

Kristina Blaiser, Ph.D., CCC-SLP
Committee Member

Chung-Hau (Howard) Fan, Ph.D.
Graduate Faculty Representative

Nov 15, 2017

Amy Lester
Comm Sci Disorders/Deaf Educ
MS 8116

RE: regarding study number IRB-FY2018-122: Relationships Between Norm-Referenced Test Scores and Narrative Language Sample Measures in School-Aged Children with Specific Language Impairment

Dear Ms. Lester:

I agree that this study qualifies as exempt from review under the following guideline: Category 4. Research involving the collection or study of existing data, documents, records, pathological specimens, or diagnostic specimens, if these sources are publicly available or if the information is recorded by the investigator in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects.
This letter is your approval, please, keep this document in a safe place.

Notify the HSC of any adverse events. Serious, unexpected adverse events must be reported in writing within 10 business days.

You are granted permission to conduct your study effective immediately. The study is not subject to renewal.

Please note that any changes to the study as approved must be promptly reported and approved. Some changes may be approved by expedited review; others require full board review. Contact Tom Bailey (208-282-2179; fax 208-282-4723; email: humsubj@isu.edu) if you have any questions or require further information.

Sincerely,

Ralph Baergen, PhD, MPH, CIP
Human Subjects Chair

Acknowledgments

First and foremost I would like to thank my advisor, Dr. Diane Ogiela, for her guidance, continued inspiration, and knowledgeable contributions throughout my master's study and thesis research. I would also like to acknowledge the valued contributions from Rick Tivis, Jenny Simison, and Megan Mize to this project. Additionally, I extend my sincerest gratitude to my family and friends for their love and support, without whom this thesis would not have been possible. Finally, to those who are struggling to achieve your dreams, know that in the words of Mary Kay Ash, "Aerodynamically, the bumblebee shouldn't be able to fly, but the bumblebee doesn't know it so it goes on flying anyway." Be a bumblebee.

Table of Contents

List of Figures

List of Tables

Relationships Between Norm-Referenced Test Scores and Narrative Language Sample Measures in School-Aged Children with Specific Language Impairment

Thesis Abstract – Idaho State University (2018)

This study examined the relationships between norm-referenced test scores and language sample measures from oral narratives in 14 children with specific language impairment (SLI) and 14 age-matched peers with typical language (TL). Correlational analyses between the Expressive Language Index (ELI) of the Clinical Evaluation of Language Fundamentals, fourth edition and Oral Narration portion of the Test of Narrative Language (TNL) indicated two significant relationships between the norm-referenced test scores and narrative language sample measures for the TL group. However, no correlations were significant for the SLI group. This suggests that the evaluation tools are assessing different aspects of language for children with SLI. Additionally, effect sizes were consistently larger in the TL group than in the SLI group, suggesting less consistent performance across contexts for children with SLI. The results support the use of language sample analysis in conjunction with norm-referenced tests for accurate diagnosis of children with language deficits.

Key Words: specific language impairment, norm-referenced tests, language sample analysis, school-age children, language assessment, narrative, Systematic Analysis of Language Transcripts, Clinical Evaluation of Language Fundamentals, fourth edition, Test of Narrative Language

Chapter 1: Introduction

Identification of children for speech and language services is a multidimensional process. According to the American Speech-Language-Hearing Association (ASHA), if a child has a known disability, the potential adverse effects on speech and language must be assessed on an individual basis (ASHA, 2017). Therefore, a comprehensive evaluation, as required by ASHA (2017), includes a client case history/interview, review of cognitive, auditory, visual, and motor status, specific standardized and/or non-standardized measurements of speech, language, cognitive, and/or swallowing disorders, prognosis statement, standardized measures for speech, language, cognitive, and/or swallowing disorders which are ecologically valid and sensitive, and follow-up service recommendations. A complete evaluation helps to identify an accurate diagnosis as well as appropriate specific objectives and/or strategies for intervention.

Standardized assessment is most commonly utilized for diagnosing language impairment, due to the standardized procedures and measures required by many agencies (Spaulding, Plante, & Farinella, 2006). Norm-referenced, standardized tests allow for the results of the test to be interpreted in relation to the greater population (Spaulding, Szulga, & Figueroa, 2012). However, recent research indicates that there are significant weaknesses associated with relying primarily on standardized test scores. Betz, Eickhoff, and Sullivan (2013) determined that practicing speech language pathologists (SLPs) do not necessarily utilize tests based on their quality of psychometric properties. According to a survey of 364 SLPs, neither reliability nor validity significantly correlated with frequency of use. The only significant correlation measure of frequency of test selection was publication year, with more recent tests being used more frequently (Betz, Eickhoff, & Sullivan, 2013). Additionally, Spaulding, Plante, and Farinella (2006) identified diagnostic accuracy concerns in regard to test sensitivity and specificity levels.

These issues lead to the need for evidence-based practices, clinician education, and utilization of multiple tools for diagnosis and goal determination for children with language impairments.

Another more functional assessment of language is language sample analysis. Language samples allow for evaluation of language production in a variety of contexts, which can be analyzed with reference to expected levels of performance, past performance or other criterion-referenced measures. The "gold standard" for language assessment includes the use of multiple tools to measure various components of an individual's language and to identify an accurate profile of an individual's skills (ASHA, 2017).

The purpose of the current study is to examine the relationships between norm-referenced test performance and narrative language sample measures of oral narrative transcripts in order to better understand and identify the varied components of decontextualized and contextualized language skills of children with specific language impairment (SLI). Identification of school-aged children with SLI is multifaceted. It is based on research, individual factors, and an SLP's clinical expertise. Accurate evaluation for diagnosis of SLI is essential for the academic and social development of these children. Because of the heterogeneous profile of children with language disorders, resulting in varied performance across contexts, specific components of each assessment method may provide qualitatively different information. Further analysis of these potential differences is warranted.

Because different assessment methods provide insight into different aspects of linguistic skill, it is important for practicing SLPs to have accurate knowledge about the validity of what they are assessing with accurate interpretation of the results with regard to a child's language abilities. Due to the complex profiles of many children with language disorders, complete analysis of information gained through the assessment process is essential for accurate diagnosis

of language disorders (Betz et al., 2013; Ebert & Scott, 2014; Miller, Andriacchi, & Nockerts, 2016). Recognizing how children with language disorders are identified is the first step for understanding how each assessment component plays a role in the diagnosis and treatment of language impairments in children.

Chapter 2: Background

Norm-referenced tests and language samples are the primary methods of assessment used to evaluate the language skills of school-aged children (Ebert & Scott, 2014). Notable differences between these two methods require further analysis in order to properly interpret and compare the results. Information obtained from norm-referenced tests and from language sample analysis contributes to the comprehensive evaluation of a child's language abilities. Standardized test measures provide information regarding decontextualized language skills, an important aspect of academic and metalinguistic language (Ebert & Scott, 2014). In contrast, language sample analysis is a purposeful elicitation of language with contextualized support, important for academics, social interactions, and daily use of language (Pavelko, Owens, Ireland, & Hahs-Vaughn, 2016). According to the Individuals with Disability Education Act (IDEA; 2004), no single measure is appropriate to use when making a diagnosis; therefore, a comprehensive evaluation is required (ASHA, 2017). Due to the variability of various language skills in children with SLI, as well as other language disorders, it is important to have a good understanding of the type of information gained from various measures of language and the relationships between them.

Overview of Specific Language Impairment

When a child's primary disability is language-based with no known sensory, neurological, or developmental cause, these children are categorized as having a diagnosis of SLI. Leonard (2014) defines SLI as a multifaceted impairment of the use of language. Children with SLI develop language more slowly and require more assistance in the language learning process than typically developing children (Leonard, 2014). Diagnostic criteria are primarily exclusionary, but also require the documentation of significant deficits within language expression, reception, and construction. Due to the heterogeneous language profiles of children with SLI, many children are not diagnosed until school age, when language demands increase (Tomblin et al., 1997). Understanding of the variations in diagnostic procedures supports clinicians in the recognition of SLI characteristics and will further assist them in the diagnosis and development of appropriate objectives for language learning to gain the skills necessary for academics and socialization.

Discrimination of the SLI population is the first step to developing accurate methods of diagnosis and treatment (Leonard, 2014). Children with SLI are defined as having a significant deficit of language that is not explained by another disorder, intelligence, or hearing loss (Betz et al., 2013; Schwartz, 2009; Tomblin et al., 1997). Because there is not a specific known cause of SLI, children often go undiagnosed until increased language demands present during the school-aged years. An impairment of language is defined in many ways such that it allows for different interpretations, and is particularly dependent upon the type and context of the language requirements. Tomblin et al. (1997) completed an epidemiological study widely cited as placing the prevalence of SLI at 7% of the population at kindergarten. This percentage is greater than any other developmental disorder but is less understood due to the variability of performance

profiles (Betz et al., 2013). Language profiles of children with SLI are not distributed neatly into typical categories of language deficits (Leonard, 2014). Variability of language across the domains contributes to the challenge of identifying and diagnosing children with SLI. While there is no single clinical marker, specific morpho-syntactic aspects of language may notably help differentiate children with SLI.

Children with SLI demonstrate a broad range of language deficits that are highly variable between individuals. Leonard (2014) describes expressive language deficits throughout the domains of language including content, form, and use with varying degrees of impairment. Difficulties with semantics are observed as reduced vocabulary expression and restricted word use; morphology deficits are noted through inflectional and derivational morpheme errors and, more frequently, omissions of inflectional morphology; syntactic language deficits are demonstrated through restricted use of various sentence structures and limited sentence complexity. Any or all of these deficits impact the individual's ability to utilize language for social communication and academic learning (Leonard, 2014; Schwartz, 2009). A combined analysis of semantic, morphological, and syntactical components of language may be utilized as a clinical marker in language samples (Hoffman, 2009; Moyle, Krasinski, Weismer, & Gorman, 2011). Children with SLI also often demonstrate some degree of phonological memory impairment or restricted language understanding (Jackson, Leitao, & Claessen, 2016). Leonard (2014) argues that the language abilities of children with SLI lie along a continuum of language abilities that may not be easily differentiated from typically developing children. Since the language profiles of children with SLI are variable, different methods of assessment are likely to identify different areas of relative strengths and weaknesses.

Multiple diagnostic methods are used by SLPs to identify language impairments, but

language sample analysis and standardized tests are the two most commonly used (Ebert & Scott, 2014). However, only 66% of clinicians use language samples, often through means of conversation or observation, limiting the expressive language variability and complexity that can be analyzed in the samples (Pavelko et al., 2016). To target specific aspects of language in context, elicitation of language samples often requires structured materials and clinician prompts. The selection of appropriate assessment measures for diagnosis across contexts is the first step in identifying and treating children with SLI (Shahmanhmood, Jalaie, Soleymani, Haresabadi, & Nemati, 2016).

Norm-Referenced Tests

The use of norm-referenced, standardized test measures is the most common and widely accepted method for identifying language impairments and determining eligibility for services (Betz et al., 2013; Caesar & Kohler, 2009). Norm-reference tests are defined as formal assessments that measure select skills across varied ages and levels (Ebert & Scott, 2014). Outcomes of these tests result in standardized scores that can be utilized to compare an individual's performance to the general population. This allows an administrator to determine if a child's performance is within the normal range for their age (Spaulding, Szulga, & Figueroa, 2012). While these test scores are utilized as an important component of language impairment diagnosis, many tests indicate that they should not be used as a sole indicator (Betz et al., 2013).

The value of using norm-referenced tests is that they allow for comparison of results to a larger population (Ebert & Scott, 2014). Norm-referenced tests also allow for a sampling of language skills across different ages and language levels (Ebert & Scott, 2014). The design allows for a comparison of an individual's score to a norming sample in order to reflect how an individual performs on a particular task when compared to age-matched individuals (Spaulding

et al., 2012). Consistency of administration, objective analysis, and universal scoring measures contribute to the advantages of standardized test administration (Caesar & Kohler, 2009). While consistency of administration and scoring and standardization of test scores are valuable when diagnosing an individual, clinician knowledge, interpretation and client experience are likewise valuable. A single test may not accurately reflect the impact of an individual's language deficits. Additionally, scoring on a given test may not accurately reflect communicative performance across contexts.

While standardized tests are the most widely used form of assessment (Pavelko et al., 2016), serious concerns about clinical practices for test selection based on quality and psychometric properties have been identified by Betz and colleagues (2013). Reliability, validity, and diagnostic accuracy of norm-referenced tests are significant psychometric properties, which must be considered when using these tests to make diagnosis or when determining the degree of language impairment. Reliability is the degree to which the test results are accurate and consistent. Validity is whether or not the results of the test are an accurate measure of the target skill(s) being tested. Diagnostic accuracy involves the measures of sensitivity and specificity. Sensitivity refers to the number of children who are accurately identified as having a language impairment; specificity refers to the test's ability to identify typical children as having typical language abilities (Leonard, 2014; Spaulding et al., 2006). Professional standards consider a test with reliability, validity, and sensitivity/specificity measures of .80 or greater to have "good" psychometric properties. It is important to note that tests are not required to calculate or report these measures. (Betz et al., 2013). According to Betz et al. (2013) a clinician's likelihood to use any particular language test is not statistically correlated with the quality of the test. In fact, neither reliability, validity, nor test accuracy correlated with the frequency of test selection. If a

test does not report measures of sensitivity or specificity, have acceptable measures within those areas, or does not report other relevant psychometric properties, the test is not acceptable for classifying children with language impairment (Betz et al., 2013; Spaulding et al., 2006). Unfortunately, much of the time, SLPs are using these tests per state eligibility requirements or guidelines due to availability or familiarity, despite unacceptable or unreported psychometric properties (Betz et al., 2013). Best practice for making a diagnosis requires a comprehensive evaluation procedure that is accurate and sensitive to all aspects of relative strengths and weaknesses.

Spaulding et al. (2006) reviewed 43 language tests and determined that a large majority of norm-referenced tests were not designed to establish severity ratings and that they do not provide data for diagnosis or severity determination. Only nine test manuals provided information concerning sensitivity and specificity, and only four of those tests met the .80 acceptability criteria. Additionally, this level of sensitivity and specificity still allows for a 20% error rate, which is considerable and needs to be acknowledged and addressed. This can result in over and/or under diagnosis and the need for clinical expertise, alternative assessment and data collections methods, and education.

Another concern for the use of norm-referenced tests is that scoring requirements for determining eligibility for services have inconsistent guidelines, with various cutoff recommendations across many different states, contexts, and agencies.

> It is commonly assumed that children with language impairments can be identified because they will obtain low scores on tests of language. Indeed, school systems support this practice, frequently requiring children to score at the low end of a test's normative distribution to qualify for services. (Spaulding et al., 2006, p. 61)

As a result, children who score below an arbitrarily set standard score are diagnosed as having a language impairment (Shahmanhmood et al., 2016; Spaulding et al., 2012) and those who do not may not qualify for services. Contrary to this assumption used by schools, according to an evaluation of state education departments and test characteristics by Spaulding et al. (2012), the majority of tests lack empirical data, and "low score" criteria of the tests do not align with state cutoff-points due to inadequate sensitivity of many tests and variable guidelines. This indicates the importance of implementing evidence-based practice through educating clinicians on the need to choose tests based on the particular focus of language and their specific properties as well as supplementing with other assessment procedures (Spaulding et al., 2006).

When utilizing norm-referenced test results, the sampling population must be considered. Norm-referenced tests collect a normative sample that is supposed to be representative of the general population. Certain individual characteristics may not be represented within the reference population such as cultural and linguistic diversity (Ebert & Scott, 2014), in which case, utilizing these tests to evaluate an individual from a cultural or linguistic minority group would not be appropriate. Additionally, the selected reference group means and standard deviations are comparison scores, with the assumption that the scores from the group are in fact representative of the entire population. With this in mind, tests must not be used as a sole indicator of a language disorder (IDEA, 2004; Ireland, Hall-Mills, & Millikin, 2013; McCauley & Swisher, 1984; Spaulding et al., 2012).

Further considerations for the use of tests to identify language impairment include the decontextualized nature of the tests (Ebert & Scott, 2014), as well as adequate sampling of specific language skills. Language is used as a tool for social interactions and learning. Social and academic language skills are highly variable depending on the contextual support. Due to

such variability in linguistic level, there is a need to assess an individual's language abilities across contexts and settings. Assessment of decontextualized skills lacks ecological validity, which is the ability to generalize the results to a natural environment. Ecological validity is a measure of naturalistic expression, comprehensive abilities, and the application of language in real-life situations, including academic situations. Because tests are not a 'natural environment,' norm-referenced tests do not measure real-life, contextualized abilities. Test items assess specific skills through unnatural probes, and may only allow for expression of certain skills in limited opportunities at a surface level (Ebert & Scott, 2014). For example, the Word Structure subtest of the Clinical Evaluation of Language Fundamentals, fourth edition (CELF-4; Semel, Wiig, & Secord, 2003) measures the regular past-tense morpheme on only one single test item, limiting the number of expressive sampling opportunities. Many children with language impairments have difficulty with tense marking. If a child makes an error on this one item, the examiner cannot tell if the child's ability to use the morpheme is 0% or 90% because there was only one opportunity to use it. A school-age child with 0% accuracy on the past tense is highly likely to have a language impairment, while one with 90% accuracy is not. Thus, a child's potential difficulty with a known area of weakness for children with language impairment will be inadequately measured by this subtest.

Children may have difficulty with the artificial testing environment or the decontextualized nature of formal tests. Evaluation of decontextualized language skills through standardized testing can provide useful information regarding formal language skills. Although decontextualized characteristics of testing, may be comparable to some aspects of the academic setting, Ireland et al. (2013) indicates that according to the Connecticut State Department of Education, these tests do not "capture neither the complexities nor the subtle nuances of the

communication process" (p. 321). Higher levels of academic language involve more precise and complex vocabulary and syntax not commonly used in conversation (Barnes, Grifenhagen, & Dickinson, 2016). The shared context of a conversation supports interpretation of vague statements, whereas academic language in reading and writing must provide the explicit context to support the details for comprehension (Gee, 2014). Therefore, discourse such as narrative or expository language may be more appropriate for assessing academic language skills.

With the increased appreciation for the need for evidence-based practice, SLPs must consider the technical aspects of the tests in connection with appropriate assessment interpretations. Relationships between skills demonstrated in multiple contexts must be considered in order to make appropriate assumptions and provide optimal support for children with impaired language skills. Due to the concerns about diagnostic accuracy or insufficient psychometric properties of many tests and a lack of support for the use of arbitrary and unreliable cutoff scores, norm-referenced tests alone are inadequate for diagnosing language impairments in children.

Language Sample Analysis

Language sampling is a naturalistic, contextualized method of measuring a child's expressive language skills (Pavelko et al., 2016). Real-world situations and flexibility of language samples allow for a method of assessment that has strong ecological validity and may be used to obtain in-depth information of a child's realistic use and comprehension of language (Ebert & Scott, 2014). School-based SLPs report using language sample analysis as a part of their evaluation 67% of the time, with an emphasis on conversational elicitation procedures and less than a third using specific protocol for analysis (Pavelko et al., 2016). National surveys of SLPs' assessment methods report that time constraints and lack of knowledge may explain why

many clinicians do not utilize language samples as a means of assessment (Kemp & Klee, 1997; Pavelko et al., 2016). This fact is concerning as many professionals within the field of language assessment and development state that the use of language sample analysis is essential for proper assessment and monitoring of all children (e.g., Ebert & Scott, 2014; Miller et al., 2016; Petersen & Spencer, 2014).

Language samples can be obtained in a variety of ways. The most common methods of language sample elicitation include conversation, narrative or expository tasks, picture description, and/or observation (Pavelko et al., 2016). Different methods of elicitation may result in different demonstrations of strengths and weakness in relation to context or experiential support (Miller et al., 2016). According to Moyle, Karasinski, Weismer, and Gorman (2011) and Wetherall et al. (2007), conversational language samples play a significant role in measuring the pragmatic use of language, but the language elicited does not necessarily require advanced vocabulary or complex syntactic structures that are needed for academic language. Observation is also focused on the language use within a social context; however, unless purposeful attempts are made to prompt specific aspects of content or form, the expressive language will be simplistic in nature. Conversation and/or observation may be beneficial elicitation strategies for preschool-aged children. However, as children enter school-age and need more sophisticated and complex language skills, narrative or expository language samples are more appropriate methods for measuring language skills (Barnes et al., 2016; Epstein & Phillips, 2009; Scott & Windsor, 2000). Nippold et al. (2014) compared the expressive language of typical adolescents in conversation and narrative, and determined that narrative tasks elicited longer, more complex utterances. Moyle et al. (2011) also support the use of narrative tasks, stating that a more

demanding task for elicitation such as narrative, may help separate and identify children with language impairment when compared to their typically developing peers.

The use of oral storytelling has been determined to be a functional way to analyze more complex language in a connected expression of contextually supported ideas (Miller et al., 2016; Westerveld & Gillon, 2010). Oral narrative skills have also been shown to relate to social relationships and academic success (Crais & Lorch, 1994; Epstein & Phillips, 2000; Pavelko et al., 2016). When compared to typically developing peers, narrative expression proved to be a much more difficult task for children with language impairment, resulting in notable qualitative differences (Wetherell, Botting, & Conti-Ramsden, 2007). These differences include overall length, language complexity, level of independence, fluency, and total number of errors. Additional difficulties with oral narratives are evidenced in immature phonological processes, word naming, omissions, and limited complexity (Epstein & Phillips, 2009). Oral narratives can be used to analyze a variety of linguistic expression areas including content and form. Analysis of these narratives can be completed at two distinct levels: microstructure and macrostructure (Epstein & Phillips, 2009).

Macrostructure analysis examines the bigger picture of narrative organization and structure of story components, while *microstructure* analysis examines each utterance for grammar, vocabulary use, and complexity (Epstein & Phillips, 2009; Justice et al., 2006). According to previous studies, microstructure analysis has been determined to be a sensitive measure for narrative analysis of children with impaired language skills (Justice et al., 2006; Jerger & Thorne, 2016; Hoffman, 2009). Much effort has been put into determining which aspects of microstructure analysis identify children with SLI; however, no consensus has been

made due to variability within and across age groups. When analyzing language samples at the microstructure level, areas of productivity, complexity, and accuracy must be considered.

In the preschool years, verb tense and agreement morphology, along with utterance length, have been identified as sensitive measures for children with SLI (Bedore & Leonard, 1998; Guo & Schneider, 2016; Suoto, Leonard, & Deevy, 2014). As children develop appropriate morphology and increase their vocabulary during their primary school years, verb morphology or utterance length is not as useful as sole identification criteria (Moyle et al., 2011). Moyle et al. (2011) analyzed language transcripts of school-aged children and determined that a combination of verb and noun morphology along with MLU resulted in better discrimination between children with language impairment and typically developing language, when compared to either verb or noun morphology measures alone. These results indicate the need for a combination of more advanced analysis of language samples within appropriate context and complexity. Various measures for advanced analysis can be obtained through the Systematic Analysis of Language Transcripts (SALT) software (Miller & Iglesias, 2012), while also comparing the measures to developmental criteria and age-matched groups based on sample type. Productivity is measured by total number of words, number of different words, utterance length, and number of c-units, which are determined by a single main clause and any dependent clauses (Justice et al., 2006). Narrative complexity is measured by c-units with two or more clauses, number of coordinating or subordinating conjunctions (cohesive devices), and proportion of complex c-units (Justice et al., 2006). Accuracy is measured by completion of utterances, grammatical accuracy or tense agreement, and word omissions (Jerger & Thorne, 2016). With the variability of relative strengths and weaknesses, microstructure analysis measures help to

establish baseline performance in multiple aspects of language and can help identify functional therapy objectives.

Based on these measures, and the variability between individuals, no single measure appears to be adequate for identifying language impairments. Measures of complexity have been identified as promising for discriminating between groups of typically developing children and children with language impairment, but further research is warranted (Domsch et al., 2011; Hoffman, 2009; Moyle et al., 2011). Considerations of individualized skills and areas of identified deficits will establish a language profile that will assist in more accurate diagnosis and in establishing goals and supports needed for academic and social success.

Relationships Between Language Sample Analysis and Standardized Tests

Norm-referenced standardized tests and criterion-referenced language sample analysis are complementary measures that can assess a variety of language skills including phonological and lexical knowledge, semantics, morphology, syntax type and complexity, and pragmatics (Condouris, Meyer, & Tager-Flusberg, 2003). According to Condouris, Meyer, and Tager-Flusberg (2003), tests allow for normed comparisons across structured language components, while in contrast, language sample measures include a more contextually driven assessment through which pragmatics and discourse skills can be analyzed.

Although many clinicians and researchers recognize the complimentary value of language sample analysis and standardized tests, few studies have formally examined this relationship. Positive correlations have been found between language samples and tests performance within groups; however, classification agreement between typical and language impaired children is varied depending on the standardized cutoff score (Ebert & Scott, 2014; Manolitsi & Botting, 2011). Ebert and Scott (2014), and Ebert & Pham (2017) in a follow-up

study, found only moderate overlap between assessment tools, indicating that the assessment tools are not interchangeable due to diagnostic inconsistency. In two other studies, Condouris and colleagues (2003) and Manolitsti and Botting (2011), examined the associations between tests and language samples, with positive correlations; however, qualitative differences revealed linguistic variation based on the structure and influences within the testing situations.

Condouris and colleagues (2003) examined 44 children with autism and their performance on standardized and spontaneous speech measures. Their findings were that for children with autism, measures of standardized tests and spontaneous speech were significantly correlated, particularly with regard to lexical-semantics. The results suggest that these measures assess the same language abilities for children with autism (Condouris et al., 2003). However, this cannot be generalized to other populations of children or to skills other than lexical semantic skills.

Manolitsi and Botting (2011) looked at narrative language samples as a source of information in addition to standardized assessment across 26 Greek children with autism or SLI. Both the groups were found to have different performance across measures, resulting in qualitative differences reflected in the structure of the assessments. Based on their findings, Manolitsi and Botting concluded that, "the data suggests that narrative is a useful tool for revealing qualitative differences in language…since it provides information that is lost in more formalized testing" (p. 39). These findings support the purpose of this study as an indicator for the need to examine these patterns more closely.

Ebert and Scott (2014) completed a study that reviews multiple norm-referenced tests and language sample correlations, specifically within the school-age population. Assessment records were analyzed from a database of 73 school-age children on multi-level (word, sentence,

discourse) comparisons between a variety of language tests and editions (CELF-Core Language (Semel et al., 2003), Peabody Picture Vocabulary Test (PPVT; Dunn & Dunn, 2007), Gray Oral Reading Test—4th Edition (GORT; Wiederholt & Bryant, 2001) and microstructure components of narrative language sample from four different wordless picture books within a series. Eight SALT *z*-score measures were analyzed for partial correlations to control for age: MLU (in words), TNW, SI, NDW, omitted morphemes and words, and errors at the word and utterance-level. Results showed many more correlations between standardized test subtests and narrative sample measures that reached significance within the younger age group (6;0-8;11) as compared to the older age group (9;0-12;8). Some of the strongest correlations for the younger age group were MLU with PPVT, SI with GORT Fluency, NDW with CELF-Recalling Sentences, and Errors with CELF-Recalling Sentences. A second set of analyses examined the rate of agreement between assessment methods based on different standard deviation cutoff scores. Assessment tools were found to inconsistently identify children with a language impairment. With a -1.5 *SD* cutoff, the agreement was 42.8-75% across a variety of norm-referenced tests, while -1 *SD* cutoff identified more children with a language disorder, with 44.4-77.1% agreement. One tool did not consistently identify children more than another, highlighting the need for multiple tools. Variations between assessment results highlight diverse language ability categorization as a result of assessment accuracy and ecological validity (Ebert & Scott, 2014).

Ebert & Pham (2017) completed a follow-up study with similar methods as Ebert and Scott (2014) within a group of 51 Spanish/English bilingual children with language impairment. Analyses of the assessments considered age and native language. Within the younger age group (5;6-8;11) correlations between the methods has larger effect sizes, while minimal correlations were found for the older group of children (9;0-11;2). A relationship between the assessment

methods was found for English, indicating complementary assessment tools relating to academic language. These findings indicate significant implications for bilingual assessment and treatment.

These studies laid important groundwork for examining the relationships between norm-referenced assessments and narrative language samples that can be applied to specific populations using specific tests. A comprehensive profile of language can be obtained through interpretation of background information, test results, observations, and analysis of language use and understanding across contexts. When language is the primary impairment, a skilled SLP's, evidence-based, clinical analysis is essential to the diagnosis process (Betz et al., 2013). Diagnostic accuracy through appropriate use of language assessment tools allows for implementation of appropriate therapy goals in order to optimize a child's language skills for academic and social success.

Chapter 3: Study Purpose and Hypotheses

The primary purpose of this study is to examine the relationships between expressive language index (ELI) score subtests from the Clinical Evaluation of Language Fundamentals, fourth edition (CELF-4; Semel et al., 2003) and microstructure components of narrative language samples elicited using the Test of Narrative Language (TNL; Gillam & Pearson, 2004) in early school-age children with SLI and their peers with typical language (TL). The secondary purpose is to examine the relationships between the TNL Oral Narration (Oral_Narr) score and the microstructure components of narrative language samples elicited by that same test. The results will increase our understanding how language skills may be evaluated by either language sample analysis or norm-referenced measures. Furthermore, the results will also indicate whether the

relationships between the two assessment methods are the same for children with SLI and those with TL.

CELF-4 ELI subtests include: Word Structure (WS), Recalling Sentences (RS), and Formulated Sentences (FS). The microstructure measures are: Mean Length of Utterance in morphemes (MLU_m), Number of Different Words (NDW), Subordination Index (SI), and Morphosyntactic Accuracy Rate (MS_ACC) (accounting for production error substitutions and omissions at the morpheme and word levels). Three main hypotheses are proposed.

First, because several linguistic components of expressive tests of language are also included in language sample analysis, and because previous studies (Ebert & Scott, 2014; Ebert &Pham 2017; Condouris et al., 2003) found several significant correlations between norm-referenced tests and language sample measures, the first hypothesis is that scores from omnibus norm-referenced assessment of expressive language are expected to correlate with language sample analysis measures. This leads to specific predictions that the ELI composite score of the CELF-4 will correlate with various language sample measures from their narratives of the TNL. If these relationships are found to be significant, planned comparisons would be conducted to evaluate more specific predictions. The following specific relationships are predicted. WS is predicted to correlate with MLU_m due to the nature of the task and scoring guidelines with relation to morpheme (free and bound) production. WS is also predicted to correlate with MS_ACC rate based on the accuracy of morpheme productions. RS is predicted to correlate with MLU_m based on the child's ability to repeat appropriate morphemes, because a child's ability to repeat accurately is known to correlate with their ability to produce specific components of language (Klem et al., 2015). Additionally, RS is expected to correlate with MS_ACC because scoring of this task is based on number of errors and omissions. FS is

predicted to correlate with all language sample measures because this task most closely resembles spontaneous language production. FS relates to MLU_m because functional morphemes allows for a cohesive sentence production. NDW correlates because a higher variety of vocabulary is a resource for formulating sentences. SI is related because as the FS task progresses, the specific words given to use in sentences progress from concrete content words to abstract conjunctions, which require the use of dependent clauses and complex sentences. FS scoring is based on the child's ability to appropriately and accurately use these words in a sentence. And finally, FS is predicted to correlate with MS_ACC rate because scoring on the FS task is also based on morphological and syntactical production accuracy. The combined CELF-4 ELI is predicted to correlate with the overall MS_ACC rate of the narrative transcripts based on the predicted relations of this measure and all subtests of the CELF-4 ELI score.

Second, language sample measures obtained during administration of a norm-reference test of narrative skills should correlate with the standard score obtained on the test, as these components contribute to the scoring procedure. This hypothesis focuses on the information that is gained from the scoring of the TNL and from the analysis of the oral narrative samples gathered from it. All language sample measures are predicted to correlate with TNL Oral_Narr with varying degrees within groups, based on scoring procedures of the TNL, which includes the scoring of related components such as complete utterances with MLU_m, specific vocabulary with NDW, conjunctions and transition words with SI, grammatical and tense accuracy with MS_ACC.

At this age, children with TL have developed sophisticated, adult-like language and express themselves accurately, functionally, and consistently, while children with SLI are still struggling with the consistent, accurate use of various language forms and structures. Thus, the

final hypothesis is that the strength of the correlations will differ between the two groups. Specifically, these correlations are likely to be weaker or more variable for children with SLI when compared to children with TL. Children with SLI are likely to have more heterogeneous language skills due to different contexts, varying levels of language demands, and varied language profiles of children with SLI.

Chapter 4: Method

This study utilized an existing dataset collected by the Idaho State University's Child Language Lab for a larger study. Samples were collected by trained speech-language pathology students, supervised by the principal investigator, over a period of time from January 2013 to January 2018.

Participants

All of the participants in the study were monolingual English-speaking children who resided in Idaho, passed a hearing screening at 20dB in both ears, were observed to have normal oral-motor functioning, and scored within the normal range (above -2 *SD*) on the Test of Non-verbal Intelligence-Fourth Edition (TONI-4; Brown, Sherbenou, & Johnsen, 2010). The SLI group was comprised of 14 children (6 female, 8 male) between the ages of 6;2-8;9, with an average age of 7;3, who met the following criteria for eligibility: scored at or below 1.25 *SD* below the mean on the Expressive Language Index (ELI) of the CELF-4, no history of sensory impairment, or other developmental, genetic, or acquired disabilities. See Table 1 for a summary of the SLI participants' characteristics and test data. The control group was comprised of 14 age-matched (+/- 3 months) peers (7 female, 7 male), with typical language (TL), between the ages of 6;2-8;10, with an average age of 7;3, who met the following criteria for eligibility:

scored above 1.25 *SD* below the mean on the CELF-4 ELI, along with no history of sensory impairment, or other developmental, genetic, or acquired disability. See Table 2 for a summary of the TL participants' characteristics and test data. In addition to age, participants were matched within 1 *SD* on nonverbal intelligence (TONI-4) standard score.

Table 1

SLI Participant Characteristics

Participant	Age	Gender	TONI-4	CELF-4 ELI	TNL Oral_Narr
1	7;5	F	94	55	7
2	7;9	F	94	57	5
3	6;2	M	92	61	8
4	6;2	M	108	75	9
5	7;2	M	102	67	5
6	8;1	M	106	73	6
7	8;2	M	89	61	4
8	6;2	F	97	73	6
9	6;6	M	100	69	7
10	8;9	M	110	61	8
11	7;9	F	117	71	8
12	6;10	F	92	61	7
13	8;2	M	95	67	6
14	6;11	F	111	59	5

Note. Age reported as years;months. Gender reported as M = male, F = female. Test scores are reported as standard scores for TONI-4 and CELF-4 ELI. Test score are reported as scaled scores for TNL Oral_Narr. SLI = specific language impairment group. TONI-4 = Test of Nonverbal Intelligence, forth edition, CELF-4 = Clinical Evaluation of Language Fundamentals, forth edition, ELI = expressive language index, TNL Oral_Narr = Test of Narrative Language, Oral Narration scaled score

Table 2

TL Participant Characteristics

Participant	Age	Gender	TONI-4	CELF-4 ELI	TNL Oral_Narr
1	7;4	M	109	87	10
2	7;11	F	92	105	8
3	6;2	M	106	99	9
4	6;2	M	120	116	11
5	7;2	M	105	108	17
6	8;0	F	110	118	13
7	8;1	M	100	91	10
8	6;2	F	104	122	10
9	6;3	F	104	122	12
10	8;10	M	98	105	11
11	7;10	F	109	108	8
12	7;0	F	94	98	11
13	8;2	M	101	112	13
14	6;10	F	116	112	8

Note. Age reported as years;months. Gender reported as M = male, F = female. Test scores are reported as standard scores for TONI-4 and CELF-4 ELI. Test score are reported as scaled scores for TNL Oral_Narr. TL = typical language group. TONI-4 = Test of Nonverbal Intelligence, forth edition, CELF-4 = Clinical Evaluation of Language Fundamentals, fourth edition, ELI = expressive language index, TNL Oral_Narr = Test of Narrative Language, Oral Narration scaled score

Procedures

Trained speech-language pathology students working at the Idaho State University Child Language Lab administered multiple assessments across two sessions to all participants (TL and SLI). The CELF-4 was administered as a test for eligibility in the larger study and the TNL was administered for an analysis of oral narrative samples under consistent circumstances. Assessment sessions were audio recorded for scoring, transcription, and inter-rater reliability measures. CELF-4 core language subtests and the TNL were scored with regard to specific test manual procedures. The oral narrative tasks of TNL were transcribed and coded using SALT and lab-specific conventions by trained graduate students in speech-language pathology.

Scaled scores for both the SLI and TL children from the CELF-4 core language subtests included: Concepts & Following Directions, Word Structure (WS), Recalling Sentences (RS),

and Formulated Sentences (FS). The CELF-4 ELI score was derived from the WS, RS, and FS subtests. The CELF-4 ELI composite score was used as the comparison measure, rather than the individual subtests, to decrease the total number of correlations and correction factors.

The oral narrative components were elicited from the TNL expressive subtests of: McDonald's Retell (MR), Late for School Story (LSS), Aliens Story (AS). The narratives were transcribed into SALT and coded with standard SALT conventions and lab specific codes. Similar to the methods used in the Ebert and Scott (2014) study, the children's transcripts were analyzed for measures such as Mean Length of Utterance in morphemes (MLU_m), Number of Different Words (NDW), Subordination Index (SI), and Morphosyntactic Accuracy Rate (MS_ACC). Methods were similar to Ebert and Scott (2014); however, our study had a narrower age range and the assessments were the same across all participants so as to strengthen controls and decrease variability.

MLU_m is an index of expressive morphological productivity, calculated by the average number of morphemes per utterance. NDW is an index of expressive semantic productivity and diversity, calculated by totaling the number of different root words used throughout the narrative. SI is an index of expressive syntactic complexity, calculated by the ratio of clauses (main and subordinate) to total utterances. MS_Acc rate is an index of expressive morphosyntactic accuracy, calculated by categorization of verb and tense related morphemes (progressive -*ing*, irregular past-tense, regular past-tense -*ed*, third person regular -*s*, third person irregular, auxiliary/copula) and noun-related morphemes (plural -*s*, possessive *'s*, determiner, pronoun), and collapsed into a single category in order to decrease the overall number of comparisons. These measures were selected in order to analyze the microstructure aspects of productivity, complexity, and accuracy of language in oral narratives.

Reliability

Inter-rater reliability measures were conducted for test scoring and transcription coding procedures. Two participants from each group (14.3%) were randomly selected and independently analyzed for test scoring and transcript coding reliability by another trained graduate student in speech-language pathology. Inter-rater reliability resulted in 88% for item-by-item agreement of the CELF-4 ELI subtests and 86% for item-by-item agreement of the TNL Oral_Narr subtests. Reliability for narrative language transcript coding was 89%.

Data Analysis

Within group correlations for children with SLI and TL compared CELF-4 ELI composite scores to the narrative transcript measures to examine overlap between aspects of narrative analysis and the CELF-4 standard test score. These correlations included: MLU_m, NDW, SI and MS_Acc rate with composite CELF-4 ELI score. Within group comparisons between the TNL Oral_Narr and narrative transcript measures examined how transcript measures correlated with the TNL standardized test score from the same narratives that generated the transcript. Comparisons included: MLU_m, NDW, SI, and MS_ACC with TNL Oral_Narr. See Table 3 for the means and standard deviations of each measure by group.

The data was initially evaluated for normality in order to attempt Pearson correlations. However, analyses indicated that several variables were not normally distributed. Square root and log linear transformations were unsuccessful in normalizing the distributions for several variables within the TL group. Therefore, nonparametric Spearman correlations were applied in order to avoid violating the assumption of normality. Although, standard scores on the CELF-4 account for age differences across children, the language sample measures were still subject to age differences across children due to the use of raw data from transcripts. In order to control for

age in the language sample measures, partial correlations were performed between CELF-4 ELI and language sample measures. To account for the possibility of Type 1 error due to multiple comparisons, the false discover rate (FDR) procedure (Benjamini & Hochberg, 1995) using a rate of 0.1 was applied. Correlations were performed to examine whether the two types of assessment measures measured the same language constructs within groups. The effect sizes of various relationships were compared and qualitatively defined between the two groups.

Table 3

Means and Standard Deviations of Assessment Measures

Measures	CELF-4 ELI		TNL Oral_Narr		MLU_m		NDW		SI		MS_ACC	
Group	SLI	TL	SLI	TL	SLI	TL	SLI	TL	SLI	TL	SLI	TL
M	65.00	107.36	6.50	10.79	6.72	8.11	99.50	131.00	1.09	1.25	0.79	0.97
SD	6.52	10.78	1.45	2.46	1.10	1.54	35.40	39.10	0.16	0.17	0.13	0.03

Note. Test scores are reported as standard scores for ELI. Test score are reported as scaled scores for TNL Oral_Narr. *M* = mean. *SD* = standard deviation. SLI = specific language impairment group, TL = typical language group. WS = Word Structure subtest, RS = Recalling Sentences subtest, FS = Formulated Sentences subtest, CELF-4 ELI = Clinical Evaluation of Language Fundamentals, fourth edition, expressive language index, TNL Oral_Narr = Test of Narrative Language Oral Narration scaled score, MLU_m = mean length of utterance in morphemes, NDW = number of different words, SI = subordination index, MS_Acc = morphosyntactic accuracy

Chapter 5: Results

Surprisingly, few correlations were statistically significant. Those that were significant occurred only in the TL group. Results of the partial correlations between language sample measures with CELF-4 ELI and TNL Oral_Narr are displayed in Table 4. The effect sizes were interpreted following Cohen (1977) based on rho (ρ) values: 0.10-0.29 as small, 0.30-0.49 as medium, and ≥ 0.50 as large effect sizes. To avoid Type 1 errors, *p* values for each correlation were compared to a new critical value determined by the FDR correction procedure.

Results of partial correlations between CELF-4 ELI score and language sample measures, controlling for age, are displayed in the first row of data in Table 4 below. Within the SLI group,

none of the correlations were significant. Within the TL group, the correlation between CELF-4 ELI and MS_Acc (ρ=.63, *p*=.02) was significant before FDR correction and remained significant after the FDR correction procedure was applied.

Results of partial correlations between TNL Oral_Narr and language sample measures, controlling for age, are displayed in the second row of data in Table 4 below. Within the SLI group, none of the correlations were significant. Within the TL group, two correlations of TNL Oral_Narr with NDW (ρ=.82, *p*=.0007) and TNL Oral_Narr with SI (ρ=.57, *p*=.04) were significant before FDR corrections were applied. However, after FDR procedure was applied, only the correlation between TNL Oral_Narr and NDW remained significant.

Table 4

Effect Size Values (ρ) Within Groups

Measures	MLU_m		NDW		SI		MS_Acc	
Group	SLI	TL	SLI	TL	SLI	TL	SLI	TL
CELF-4 ELI	.0681	**.4429**	.0624	.0695	.1700	-.0090	**.3275**	**.6342****
TNL Oral_Narr	**.3435**	**.4242**	**.4479**	**.8165****	-.1092	**.5728***	**-.4345**	.1999

Note. ρ = rho, effect size. *p* = probability. MLUm = mean length of utterance in morphemes, NDW = number of different words, SI = subordination index, MS_Acc = morphosyntactic accuracy, CELF-4 ELI = Clinical Evaluation of Language Fundamentals, fourth edition, expressive language index, TNL Oral_Narr = Test of Narrative Language Oral Narration scaled score (TNL), SLI = specific language impairment group, TL = typical language group

* Significant (*p*< .05) before FDR correction. ** Significant after .1 FDR (Benjamini & Hochberg, 1995) correction Medium effect sizes are shown in boldface. Large effect sizes are shown in underlined boldface.

In addition to the two significant correlations for the TL group, there were considerable differences of effect size, reflected in the ρ between groups. See Figure 1 for effect size (ρ) values for each group for various relationships. In the SLI group, there were no large effect sizes. Medium effect sizes were found between correlations of CELF-4 ELI with MS_Acc

(ρ=.33), TNL Oral_Narr with MLU_m (ρ=.34), NDW (ρ=.45), and MS_Acc (ρ=-.43). The remaining effect sizes were small for the SLI group. In the TL group, large effect sizes were found for CELF-4 ELI with MS_Acc (ρ=.63), TNL Oral_Narr with NDW (ρ=.82), and TNL Oral_Narr with SI (ρ=.57). There were medium effect sizes for CELF-4 ELI with MLU_m (ρ=.44) and TNL Oral_Narr with MLU_m (ρ=.42). The remaining effect sizes for the TL group were small. Effect size differences in the groups indicate stronger relationships between the assessment measures for children with TL and weaker relationships between the measures for children with SLI. Differences in effect sizes will be identified and interpreted in the discussion.

Figure 1. Comparison of Language Measure Effect Sizes Between Groups

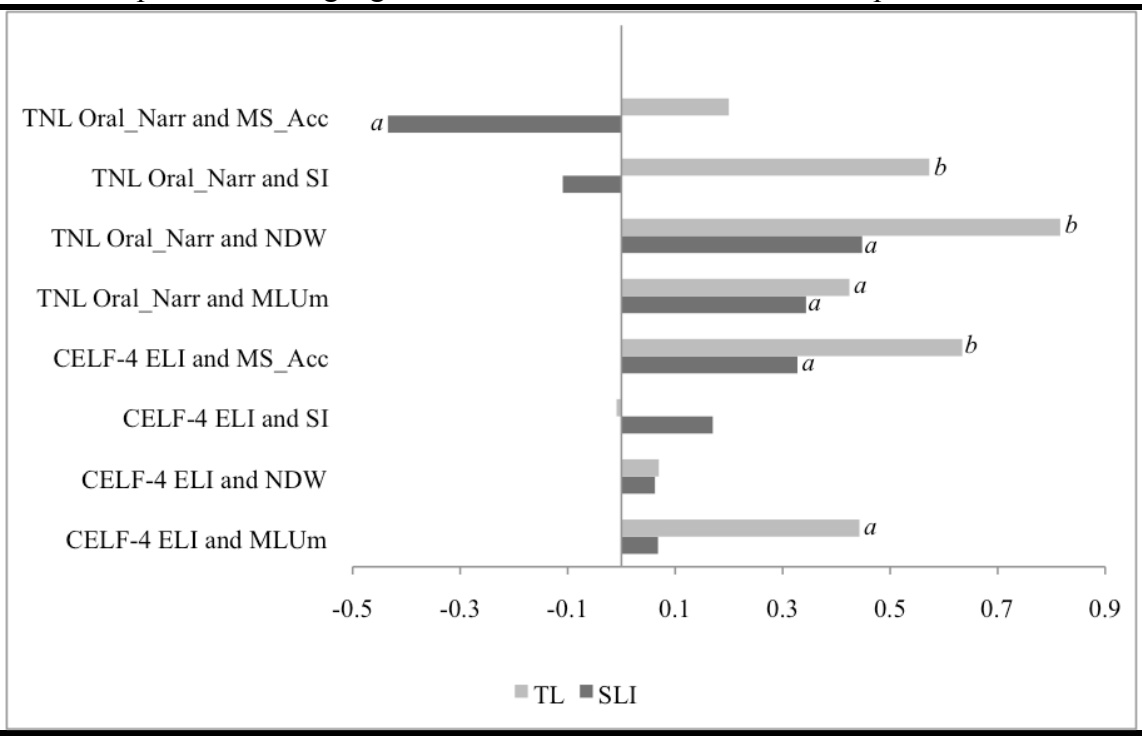

Figure 1. The figure displays the comparison of effect size (ρ) between groups.
SLI = specific language impairment group, TL = typical language group, CELF-4 ELI = Clinical Evaluation of Language Fundamentals, fourth edition, expressive language index, TNL Oral_Narr = Test of Narrative Language Oral Narration scaled score, MLUm = mean length of utterance in morphemes, NDW = number of different words, SI = subordination index, MS_Acc = morphosyntactic accuracy

a = medium effect size, ρ = 0.3-0.49; *b* = large effect size, ρ ≥ 0.5

Chapter 6: Discussion

Within group correlations between the CELF-4 ELI score and narrative analysis examined how language sample measures relate to the CELF-4 ELI standardized test scores. Correlations between the TNL Oral_Narr and the language sample measures examined the relationship between the analysis of the oral narrative and the TNL oral narrative standardized test scores that were based on the same narrative sample. Between group comparisons qualitatively examined differences in effect sizes for various relationships.

The first hypothesis that omnibus norm-referenced assessment of expressive language was expected to correlate with language sample analysis measures was only supported by a single measure within the TL group. The significant correlation between CELF-4 ELI and language sample measures suggests (MS_Acc) suggest that both the CELF-4 ELI subtests and the language sample measures evaluated similar aspects of morpho-syntax for the children with TL. It was initially anticipated that there would be a greater number of significant correlations in both groups between the CELF-4 ELI and the language sample measures that would allow for investigation of more fine-grained analysis of relationships between the oral narrative analysis and the specific CELF-4 expressive subtests. However, that was not the case. The lack of statistically significant correlations within the SLI group suggests that the scores and language sample measures appear to assess different aspect of language production.

The second hypothesis was that all narrative sample measures would correlate with norm-referenced tests of narrative language. Only a single significant correlation within the TL group supported this hypothesis. The correlation between TNL Oral_Narr and language sample measures (NDW) suggests that children with TL that have greater lexical diversity and have more in depth and complete expressive narratives as measured by higher TNL Oral_Narr. The

remaining predicted correlations were not found to be statistically significant in either group. This suggests that the TNL protocol designed to assess oral narratives does not analyze the microstructure components of the sample to the same extent that language transcript analysis through SALT software does. Discussion of the lack of relationships is discussed below.

These findings suggest that for children with SLI, the two standardized assessment scores of CELF-4 ELI and TNL Oral_Narr do not measure the same aspects of language that were measured by the language sample analysis measures, or that the same aspects are not measured at the same level of depth. These results highlight the importance of assessing children's language use in a variety of contexts based on their performance between these two assessment methods, especially for children suspected of having SLI. The lack of significant results supports the premise that the assessment methods are not equal or interchangeable measures of language performance for children with SLI, and that expressive language is highly variable across contexts.

Effect Size

Although there were limited findings of statistical significance, it is noteworthy that there were several effect sizes (ρ) that could be characterized as medium or large and that effect sizes differed substantially across groups. (See Figure 1). The differences of effect sizes for various relationships between the groups supports the third hypothesis that the strength of the relationships between norm-referenced test scores and language sample measures differ between the two groups. Specifically, these correlations were predicted to be weaker and more variable for children with SLI when compared to children with TL.

There where notable differences between the relationships in which TL had stronger effect sizes and SLI had weaker effect sizes. The stronger effect sizes within the TL group were

between CELF-4 ELI with MLU_m and MS_Acc and TNL Oral_Narr with NDW and SI. In the SLI group, only the relationship between TNL Oral_Narr and MS_Acc resulted in a larger effect size, and this was in the context of a negative relationship. The remaining relationships were not notable in differences. One possible explanation for these results is based on the knowledge that typical language development patterns show that school-aged children have generally mastered many of these components of language measured by norm-referenced assessments and language sample analysis; children with SLI have not yet reached this level of proficiency in their language. While children with TL are performing near ceiling levels as expected by this age, children with SLI perform more variably across measures given different contexts; therefore, resulting in weaker effect sizes. This is a reflection of inconsistent and inaccurate use of language across contexts for children with SLI. Another possibility for this difference is that, based on findings from Hoffman (2009), Leonard (2014), and Moyle and colleagues (2011), specific components of each assessment method may provide qualitatively different information based on specific strengths and weaknesses of the individual child. Performance across contexts of narrative samples and norm-referenced tests vary based on the specific components of language that are elicited. This indicates that these two assessment measures are not equal measures of semantics, morphology, or syntax given the different language requirements. This explanation suggests that different assessment contexts and demands impact language performance of children with SLI, while these factors impact children with TL to a lesser degree, resulting in larger effect sizes.

Relationship Between Norm-Referenced Tests and Language Samples

There are foundational differences in the assessment methods of norm-referenced testing of the CELF-4 and TNL when compared to narrative language sample analysis procedures,

which may partially account for the limited significant correlations, especially for the SLI group. A variety of factors must be considered. First, the purpose of each measure impacts the results and interpretation of those measures. Second, the type and level of sampling varies due to the opportunities for use and accuracy of different language components. Third, the level of support and contextualized nature of the measures may impact the overall performance. And fourth, patterns of performance must be examined with consideration of classification of children with and without SLI.

Interpretation of the results of norm-referenced tests may be applied differently than the analysis of language sample performance. Standard scores are often a required component of a comprehensive evaluation; however, many clinicians do not use language sample analysis or are prevented from using language sample analysis as an additional criterion-referenced measure during evaluation to help determine diagnosis and eligibility. Thus, this crucial and ecologically valid information can be missing from the decision-making process. Norm-referenced tests may be utilized to examine an individual's overall language performance in comparison to the norming population at the surface level, while language samples may provide a more in-depth analysis of within a functional context. For example, according to Semel, Wiig, and Secord (2003), the purpose of the CELF-4 is to determine eligibility of services, identify language strengths and weaknesses, and to provide an assessment method that is based on educational curriculum. The purpose also includes unbiased administration and scoring guidelines, to ensure consistent, reliable, and sensitive results for determining the presence or absence of a language disorder. While the CELF-4 may be an accurate measure for eligibility purposes based on reasonable sensitivity and specificity, it does not assess depth of language across contexts Furthermore, and although sensitivity and specificity may be at acceptable levels, they are not at

100% and clinicians need to be able evaluate those children who would otherwise be misidentified by norm-referenced tests (up to 20% of the time) through data-based approaches such as language sample analysis.

According to Gillam and Pearson (2004), the purpose of the TNL is to examine functional discourse through the ability to answer questions, retell stories, and create stories. Furthermore, the purpose of the test is to allow for normed narrative analysis without transcription so as to save time. While this format of testing allows for an assessment within the context of narrative comprehension and production, namely macrostructure analysis, which relates to social and academic development, the scoring procedures only scratch the surface with regard to depth of analysis for microstructure aspects of oral narratives as measured by scoring the use of specific semantic and syntactic structures and morphological accuracy measured by number of errors and tense consistency in the narratives. For example, if a child maintains tense throughout a story, they score a 2, but if they change tense once, they score a 1, and if they change tense two or more times, they score a 0. Similarly, if a child has no grammatical errors, they score a 2, but if they produce one or two grammatical errors, they score a 1, and if the produce three or more grammatical errors, they score a 0. In reality, some of the children in this study produced only three errors, while others produced 14 or more errors. This degree of such differences is not fully reflected in the scoring of the TNL. TNL scoring guidelines include aspects of narrative macrostructure, such as the use of character names, temporal relationships, conflict development, and organized story structure. These macrostructure elements more heavily influence the final score as compared to the microstructure elements. Narrative language sample analysis analyzes a sample of expressed language that is contextually based, and can be

examined in greater detail at the microstructure level for morphology, semantics, and syntax or at the macrostructure level for story grammar components, organization, and cohesion.

The issue of sampling, while not studied in depth in much in the research, is an important factor to consider with variable performance across measures. The depth and limitations of assessment can be examined with a case-by-case scenario. For example, within the WS subtest, a variety of morphemes are assessed one or two times, often within the same context, following an example and/or a prompt. In a language sample, a variety of morphemes are evaluated within multiple contexts and across utterances to allow for a more in depth analysis of morphology. Further consideration of expressed morphology is that the standardized tests control which morphemes are sampled, while language samples can only measure the morphemes for which obligatory contexts occur in the sample. Thus, norm-referenced tests often sample morphemes at the surface level, but across many more types, while language samples elicit a representative sample of those morphemes that a child does use or omit within a specific context at a deeper level. Additionally, the Word Classes and Expressive Vocabulary subtests of the CELF-4, which were not included in this study, assess semantics, but restrict the individual to the provided context, limiting the semantic variety. Language samples allow for greater freedom of expression, which in turn increases semantic variety. Furthermore, syntax is assessed within RS and FS, based on the child's ability to repeat certain sentence structures or produce sentence types, but it is also restricted by the provided context and presented word. Scoring of these subtests is partially dependent on morpho-syntactic accuracy; for example, if a child produces an accurate sentence with no errors, they score a 3, if a child produces one morphological or syntactic error, they score a 2, if they produce two-three errors, they score a 1, and if they

produce four or more errors, they score a 0. Errors are measured by whole or part-word omissions, substitutions, and word order.

Sampling issues within the TNL appear with reference to the scoring guidelines as well. As described above, certain components of morphology, semantics, and syntax are analyzed at a basic level, resulting in a score of 0, 1, or 2 based on the accuracy (microstructure) and use of narrative elements (macrostructure). These scoring procedures must be considered because children with SLI have variable degrees of accuracy that does not get reflected in the score. Because they may have many more errors than the scoring schema accounts for, the final score may not reflect their actual level of difficulty. According to Colozzo, Gillam, Wood, Schnell, and Johnston (2011), the more the child produces, the more information they are likely to provide; however, the more the child produces, the more errors they are likely to have. This may partially explain the negative relationship in the SLI group between the TNL Oral_Narr score and MS_Acc of the narrative (See Figure 1). The increased number or length of utterances, or amount of content in general, that the child produced, the more errors they had at the level of microstructure; however, it also appeared to result in a higher TNL Oral_Narr score. On the contrary, the shorter and less complex the narratives, the fewer errors the children with SLI produced at the level of microstructure, but this resulted in an overall lower TNL Oral_Narr score, due to very limited content. This relationship did not show up in the TL group, because their language performance across measures of narrative norm-referenced scores and analysis measures was more consistent.

Finally, narrative sample analysis through SALT can examine microstructure components at a deeper level, considering all measures of MLU, NDW, SI, and errors resulting at the morphological, word, or sentence level. At the same time, language samples allow for the

freedom of structure and expression, which may not set a child up to express a range of vocabulary, morphology, or complexity within their utterances. Looking at the range of semantic, morphology, and syntax use can best be done by employing both methods.

Another consideration of differences is within the levels of support and contextualization of the protocols of the assessments themselves. Expressive language components are restricted to the elicitation contexts within the CELF-4 and to some extent, the TNL as well. The WS subtest of the CELF-4 includes an indirect, delayed model, so as to elicit the specific morpheme intended. RS subtest, may actually assess working and phonological memory components along with expressive language abilities, and does not provide context for any of the produced sentences. The FS subtest, forces certain types of sentences and clauses to be produced in a limited context based on the presented word and picture, and does not allow for spontaneous production of sentence types. As identified by Barnes and colleagues (2016) and Gee (2014), structured language use is essential for higher-level academic language use and comprehension. Inability to produce or understand syntactically complex sentences may impact academic performance. While the specific structured sentence elicitation of FS may lack ecologic validity, performance on this subtest may indicate areas of strength and weakness relevant to academic language when compared to typical peers. Although, the TNL restricts the expressive context within the retell narrative and based on the provided picture(s) for context, the organization and variability of the expressed narrative is relatively free. This provides a somewhat spontaneous, contextualized, oral narrative that can then be examined through a norm-referenced score and language transcript analysis, as a measure of academic and social language abilities.

Overall, the patterns of performance across language domains may highlight meaningful differences that are identifiable for children with SLI when compared to peers with TL. These

patterns should be considered with respect to classification and treatment goals for children with SLI so as to optimize their therapy for academic and social success. Considering the lack of correlations between norm-referenced assessment scores and language sample analysis measures within the SLI group and differences of effect sizes between groups, heterogeneity is highlighted across assessment methods and between groups.

Related Research

The results of the current study differ in several ways from Ebert and Scott's (2014) study findings, despite similar methods of analysis. Ebert and Scott's retrospective study of narrative language sample and norm-referenced assessment comparison for school-aged children (73 participants), found many significant correlations ($p < 0.10$ after FDR correction) that reached significance within the younger age group (6;0-8;11), that did not hold for the older age group (9;0-12;8). Resulting correlations between narrative sample measures and CELF for the younger group subtests included: MLU with WS and FS, SI with RS, NDW with WS, RS, and FS, and Errors with WS, RS, and FS. The present study found a single significant correlation between CELF-4 ELI and MS_Acc in the TL group, which is similar to their reported correlations of errors with specific ELI subtests. Several factors may have contributed to the different results in Ebert and Scott and the present study. Ebert and Scott used a variety of test instruments and different wordless picture books for narrative sample elicitation. A total of four different standardized tests (11 subtests) and four different wordless picture books were utilized in the study, rather than a consistent smaller set of assessments. This results in the use of different norm groups used for comparison and a variety of elicitation contexts for narratives. Procedural methods of the present study added tighter controls through the method that children were all given the same standardized tests and the same standardized procedure for narrative

sample elicitation. Ebert and Scott included participants who had been referred for speech-language evaluation within the past 10 years and who had completed both a narrative language sample and a score on a norm-referenced language test during the evaluation and had no evidence of intellectual disability. This study controlled group participation through eligibility criteria for the SLI group and then had a control group of children with TL matched on age and nonverbal cognitive scores. Group criterion was similar to their younger age group on the basis of age and nonverbal intelligence. However, their statistical correlations included children with and without language impairment, while our correlations were done within distinct groups. This likely influenced the distribution of results, as the children without language impairment may have made a substantial contribution to the correlations while the contributions of the children with language impairments may have made lesser contributions.

Despite the differences in the actual correlations obtained in the two studies, both sets of results suggest that the use of both standardized and narrative language sample procedures are valuable components to a comprehensive evaluation so as to assess across contexts and as a multi-modal classification of language impairment. Ebert and Scott (2014) also examined classification agreement and the role of age. Our study did not examine these aspects due to a smaller age range of participants, ensuring tighter controls for age. Future analysis of the data, including a larger sample size and increased age range, could potentially examine classification rates between the CELF-4 and language sample measures along with the effect of language variability across the ages.

A follow-up study by Ebert & Pham (2017) found that for younger school-aged, bilingual children with language impairment, correlations between language sample and standardized test scores had larger effect sizes than for an older group of bilingual children with language

impairment. These findings were specific to bilingual assessment. Other studies that consider language sample analysis and standardized test performance found that for children with autism, standardized test scores and spontaneous speech were significantly correlated, with regard to lexical-semantics (Condouris et al., 2003). The present study made more use of measures of morphology and syntax than lexical semantics. However, this study did not include a control group, and the findings cannot be generalized to populations of children with language difficulties other than autism. Manolitsi and Botting (2011) found that groups of children with autism and SLI were found to have qualitative differences across measures, concluding that information from language sample analysis provides information that is not examined in formalized testing methods. Those findings support the results of this study that correlations between language samples analysis and standardized test scores lacked significance. Differences in the findings of similar studies suggest that specific procedural methods have a strong impact on the results, indicating the need for tighter controls within future studies. However, they both support the use of narrative samples as an assessment tool that may provide additional diagnostic and clinically relevant information not obtained from standardized language assessment.

Limitations

The number of participants limited the power of this study. While Ebert & Scott (2014) had a total of 73 qualifying participants, this study had 14 per group. Internal validity is strengthened by tight participant and procedural controls. However, this limits the generalizability of the results to different assessment methods and other populations of children. Additionally, this study did not include semantic errors, due to difficulty with calculating rates of semantic errors and the overall limited presence of semantic errors. Errors of prepositions, double tense marking, and idiosyncratic substitutions were noted within the SLI group; therefore,

future research may examine these types of errors based on potential impact of academic, semantic comprehension and use.

Clinical Implications

Based on requirements for a comprehensive assessment outlined by ASHA and IDEA 2004, language sample analysis may provide an additional and vital assessment procedure that compliments norm-referenced assessment procedures and adds valuable information to the analysis of a child's language profile. A deeper sampling, along with relatively unrestricted contextualized expression of language, allows for an analysis of language across contexts that relates to academic and social performance. Norm-referenced assessment is also an essential component as outline by state eligibility guidelines for services as well as for comparisons to peers.

Norm-referenced assessment of the CELF-4 results in a standard score, which can be used as a general comparison to expected age-matched performance of peers if the child is represented in the group. This score is valuable for eligibility for services, required by most school districts and states. However, the cut off scores based on number of standard deviations from the mean, are arbitrary, not test specific and not necessarily tied to clinical significance (Spaulding, et al. 2006). Additionally, the CELF-4 reports psychometric properties that allow for the results of the tests to have acceptable measures of validity, reliability, and accurate (sensitivity and specificity) differentiation between language impaired and typical language.

TNL assessment is a valuable standardized measure of narrative comprehension and expression. While assessment procedures do not require transcription and analysis, expediting the assessment process, valuable information from the oral narratives could be obtained if it is transcribed and analyzed. The TNL scoring procedure includes both microstructure and

macrostructure components, with an emphasis of story grammar components. Narrative macrostructure components were not measured in this study, but may hold valuable information about differentiating children with language impairment. The examiner manual of the TNL states that, "no single instrument is adequate to identify a language disorder" and that "the TNL cannot differentiate between a language differences and a language disorder," but results may describe narrative performance, which adds valuable qualitative information to child's language abilities profile (p. 9, Gillam and Pearson, 2004). Furthermore, issues of cultural and dialectical biases must be considered when administering such standardized tests. Factors such as linguistic or cultural diversity and exposure to narrative structure may impact a child's ability to comprehend or express narratives.

Language sample analysis can fill in the gaps of missing information from a child's language profile, and may provide a more in-depth sampling of a child's expressive language abilities within certain areas when analyzed at domain-specific levels for morphology, semantics, and syntax. The ability to express cohesive, organized language samples in an entertaining manner may also reflect aspects of pragmatic language and discourse style. Language transcript analysis results in extremely valuable information that leads to the development of more effective and efficient treatment.

Chapter 7: Conclusion

Children with SLI often remain undiagnosed until school-age, at which time academic language is a necessity for learning and keeping up with peers. These children are missing out on valuable time during their younger development for treatment and prevention of negative impacts of SLI later in life.

The results of this study support the idea that language sample analysis, in addition to standardized tests, is an essential component of a comprehensive evaluation for identifying children with language impairment. Accurate interpretations of test scores and language samples together allow clinicians to classify children as performing within expectations or as having some degree of language impairment. A complete evaluation results in an appropriate implementation of a treatment plan to optimize a child's language skills. The dynamic nature of language must be considered on an individual basis.

While much research has been done concerning identification of children with SLI, no single measure has been determined to be sufficient. Developing narrow classification profile of language deficits may not be possible due to the heterogeneous profile of children with SLI. A combination of information gained through standardized tests and language sample analysis creates a broader understanding of language deficits. Having multiple sources of data regarding a child's language characteristics will increase the identification of children with SLI. This will support optimal, comprehensive assessment methods, so as to develop a profile of strengths and weaknesses for efficacious intervention.

References

American Speech-Language-Hearing Association (2017). *Assessment and evaluation of speech-language disorders in schools.* Retrieved from http://www.asha.org/SLP/Assessment-and-Evaluation-of-Speech-Language-Disorders-in-Schools/

Barnes, E. M., Grifenhagen, J. F., & Dickinson, D. K. (2016). Academic language in early childhood classrooms. *The Reading Teacher*, *70*(1), 39-48. doi:10.1002/trtr.1463

Bedore, L. M., & Leonard, L. B. (1998). Specific language impairment and grammatical morphology: A discriminant function analysis. *Journal of Speech, Language, and Hearing Research*, *41*(5), 1185-1192. doi:10.1044/jslhr.4105.1185

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, 57*, 289–300.

Betz, S. K., Eickhoff, J. R., & Sullivan, S. F. (2013). Factors influencing the selection of standardized tests for the diagnosis of specific language impairment. *Language, Speech, and Hearing Services in Schools*, *44*(2), 133-146. doi:10.1044/0161-1461(2012/12-0093)

Brown, L., Sherbenou, R., J., & Johnsen, S. K. (2010). *Test of Nonverbal Intelligence – Fourth Edition*. Austin, TX: Pearson.

Caesar, L. G., & Kohler, P. D. (2009). Tools clinicians use: A survey of language assessment procedures used by school-based speech-language pathologists. *Communication Disorders Quarterly*, *30*(4), 226-236. doi:10.1177/1525740108326334

Colozzo, P., Gillam, R. B., Wood, M., Schnell, R. D., & Johnston, J. R. (2011). Content and form in the narratives of children with specific language impairment. *Journal of Speech,*

Language, and Hearing Research, *54*(6), 1609-1627. doi:10.1044/1092-4388(2011/10-0247)

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. Academic Press: New York, NY.

Condouris, K., Meyer, E., & Tager-Flusberg, H. (2003). The relationship between standardized measures of language and measures of spontaneous speech in children with autism. *American Journal Of Speech-Language Pathology*, *12*(3), 349-358. doi:1058-0360/03/1203-0349

Crais, E. R., & Lorch, N. (1994). Oral narratives in school-age children. *Topics in Language Disorders*, *14*(3), 13-28. doi:10.1097/00011363-199405000-00004

Domsch, C., Richels, C., Saldana, M., Coleman, C., Wimberly, C., & Maxwell, L. (2012). Narrative skill and syntactic complexity in school‐age children with and without late language emergence. *International Journal of Language & Communication Disorders*, *47*(2), 197-207. doi:10.1111/j.1460-6984.2011.00095.x

Dunn, L. M., & Dunn, D. M. (2007). *Peabody Picture Vocabulary Test—Fourth Edition*. Minneapolis, MN: Pearson Assessments.

Ebert, K. D., & Pham, G. (2017). Synthesizing information from language samples and standardized tests in school-age bilingual assessment. *Language, Speech, and Hearing Services in Schools*, *48*(1), 42-55. doi:10.1044/2016_LSHSS-16-0007

Ebert, K. D., & Scott, C. M. (2014). Relationships between narrative language samples and norm-referenced test scores in language assessments of school-age children. *Language, Speech, and Hearing Services in Schools*, *45*(4), 337-350. doi:10.1044/2014_LSHSS-14-0034

Epstein, S. A., & Phillips, J. (2009). Storytelling skills of children with specific language impairment. *Child Language Teaching and Therapy*, *25*(3), 285-300. doi:10.1177/0265659009339819

Gee, J. P. (2014). Decontextualized language: A problem, not a solution. *International Multilingual Research Journal*, *8*(1), 9-23. doi:10.1080/19313152.2014.852424

Gillam, R. B., & Pearson, N. A. (2004). *Test of Narrative Language*. Austin, TX: Pro-Ed.

Guo, L. Y., & Schneider, P. (2016). Differentiating school-aged children with and without language impairment using tense and grammaticality measures from a narrative task. *Journal of Speech, Language, and Hearing Research*, *59*(2), 317-329. doi:10.1044/2015_JSLHR-L-15-0066

Hoffman, L. M. (2009). The utility of school-age narrative microstructure indices: INMIS and the proportion of restricted utterances. *Language, Speech & Hearing Services in Schools*, *40*(4), 365-375. doi:10.1044/0161-1461(2009/08-0017)

Individual with Disabilities Education Act of 2004, 20 U.S.C. § 1400 et seq.

Ireland, M., Hall-Mills, S., & Millikin, C. (2013). Appropriate implementation of severity ratings, regulations, and state guidance: A response to "Using norm-referenced rests to determine severity of language impairment in children: Disconnect between US policy makers and test developers" by Spaulding, Szulga, & Figueria (2012). *Language, Speech & Hearing Services in Schools*, *44*(3), 320-323. doi:10.1044/0161-1461(2012/12-0048)

Jackson, E., Leitao, S., & Claessen, M. (2016). The relationship between phonological short-term memory, receptive vocabulary, and fast mapping in children with specific language impairment. *International Journal Of Language & Communication Disorders*, *51*(1), 61-73. doi:10.1111/1460-6984.12185

Justice, L. M., Bowles, R. P., Kaderavek, J. N., Ukrainetz, T. A., Eisenberg, S. L., & Gillam, R. B. (2006). The index of narrative microstructure: A clinical tool for analyzing school-age children's narrative performances. *American Journal of Speech-Language Pathology*, *15*(2), 177-191. doi:10.1044/1058-0360(2006/017)

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*(2), 161-176. doi:10.1177/026565909701300204

Klem, M., Melby‐Lervåg, M., Hagtvet, B., Lyster, S. A. H., Gustafsson, J. E., & Hulme, C. (2015). Sentence repetition is a measure of children's language skills rather than working memory limitations. *Developmental Science*, *18*(1), 146-154. doi:10.1111/desc.12202

Leonard, L. (2014). *Children with specific language impairment.* Massachusetts Institute of Technology: Cambridge, MA.

Manolitsi, M., & Botting, N. (2011). Language abilities in children with autism and language impairment: Using narrative as a additional source of clinical information. *Child Language Teaching & Therapy*, *27*(1), 39-55. doi:10.1177/0265659010369991

McCauley, R. J., & Swisher, L. (1984). Use and misuse of norm-referenced tests in clinical assessment: A hypothetical case. *Journal of Speech and Hearing Disorders*, *49*(4), 338-348. doi:10.1044/jshd.4904.338

Miller, J. F., Andriacchi, K., & Nockerts, A. (2016). Using language sample analysis to assess spoken language production in adolescents. *Language, Speech & Hearing Services In Schools*, *47*(2), 99-112. doi:10.1044/2015_LSHSS-15-0051

Miller, J., & Iglesias, A. (2012). Systematic Analysis of Language Transcripts (SALT), Research Version 2012 [Computer Software]. Middleton, WI: SALT Software, LLC.

Moyle, M. J., Karasinski, C., Weismer, S. E., & Gorman, B. K. (2011). Grammatical morphology in school-age children with and without language impairment: A discriminant function analysis. *Language, Speech, and Hearing Services in Schools*, *42*(4), 550-560. doi:10.1044/0161-1461(2011/10-0029)

Nippold, M. A., Frantz-Kaspar, M. W., Cramond, P. M., Kirk, C., Hayward-Mayhew, C., & MacKinnon, M. (2014). Conversational and narrative speaking in adolescents: Examining the use of complex syntax. *Journal of Speech, Language, and Hearing Research*, *57*(3), 876-886. doi:10.1044/1092-4388(2013/13-0097)

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

Petersen, D., & Spencer, T. D. (2014). Narrative assessment and intervention: A clinical tutorial on extending explicit language instruction and progress monitoring to all students. *Perspectives on Communication Disorders and Sciences in Culturally and Linguistically Diverse Populations*, *21*(1), 5-21. doi:10.1044/cds21.1.5

Schwartz, R. G. (2009). *Handbook of child language disorders*. New York, NY, US: Psychology Press.

Scott, C. M., & Windsor, J. (2000). General language performance measures in spoken and written narrative and expository discourse of school-age children with language learning disabilities. *Journal of Speech, Language, and Hearing Research*, *43*(2), 324-339. doi:10.1044/jslhr.4302.324

Semel, E., Wiig, E., & Secord, W. (2003). *Clinical Evaluation of Language Fundamentals—*

Fourth Edition. San Antonio, TX: The Psychological Corporation.

Shahmahmood, T. M., Jalaie, S., Soleymani, Z., Haresabadi, F., & Nemati, P. (2016). A systematic review on diagnostic procedures for specific language impairment: The sensitivity and specificity issues. *Journal of Research in Medical Sciences*, *21*(5), 1-16. doi:10.4103/1735-1995.189648

Spaulding, T. J., Plante, E., & Farinella, K. A. (2006). Eligibility criteria for language impairments: The low end of normal always appropriate?. *Language, Speech, and Hearing Services in Schools*, *37*(1), 61-72. doi:10.1044/0161-1461(2006/007)

Spaulding, T. J., Szulga, M. S., & Figueroa, C. (2012). Using norm-referenced tests to determine severity of language impairment in children: Disconnect between US policy makers and test developers. *Language, Speech, and Hearing Services in Schools*, *43*(2), 176-190. doi:10.1044/0161-1461(2011/10-0103)

Souto, S. M., Leonard, L. B., & Deevy, P. (2014). Identifying risk for specific language impairment with narrow and global measures of grammar. *Clinical Linguistics & Phonetics*, *28*(10), 741-756. doi:10.3109/02699206.2014.893372

Tomblin, J. B., Records, N. L., Buckwalter, P., Zhang, X., Smith, E., & O'Brien, M. (1997). Prevalence of specific language impairment in kindergarten children. *Journal of Speech, Language, and Hearing Research*, *40*(6), 1245-1260. doi:10.1044/jslhr.4006.1245

Westerveld, M., & Gillon, G. (2010). Oral narrative context effects on poor readers' spoken language performance: Story retelling, story generation, and personal narratives. *International Journal Of Speech-Language Pathology*, *12*(2), 132-141. doi:10.3109/17549500903414440

Wiederholt, J. L., & Bryant, B. R. (2001). *Gray Oral Reading Tests—Fourth Edition*. Austin,

TX: Pro-Ed.

Wetherell, D., Botting, N., & Conti‐Ramsden, G. (2007). Narrative in adolescent specific language impairment (SLI): A comparison with peers across two different narrative genres. *International Journal of Language & Communication Disorders*, *42*(5), 583-605. doi:10.1080/13682820601056228

Appendix

Appendix A. Summary of Levels of Evidence of Language Sample Analysis and Norm-Referenced Assessment Studies in Speech-Language Pathology

Study	Participants	Method	Results	Level of Evidence
Bedore & Leonard, 1998	38 children; 19 with SLI, 19 with TL; ages 3;7-5;9	Spontaneous speech samples were transcribed and coded for verb morphology, noun morphology, and MLU.	Verb morphology resulted in fair sensitivity for SLI and very good sensitivity for TL. MLU was sensitive for SLI, but specificity was lower for TL. Verb composites and MLU were most sensitive for classifying SLI vs. TL.	IV – Case-Control Study
Betz, Eickhoff, & Sullivan, 2013	364 SLPs	SLPs completed a survey regarding standardized tests used to diagnose SLI; 55 test manuals were reviewed	The most frequently utilized tests (CELF-4, PLS-4, PPVT-4) correlated with publication year, and not quality of psychometric properties.	V – Cross-Sectional Survey
Condouris, Meyer, & Tager-Flusberg, 2003	44 children with autism; ages 4-14	Scores from CELF-P/CELF-3, PPVT-3, EVT were compared to spontaneous language sample measures (MLU, IPSyn, NDWR) derived from a play-based setting.	2 of the spontaneous language sample measures, NDWR and MLU, correlated with the standardized tests within the same domain. For children with autism performance across standardized tests and spontaneous language is relatively consistent.	IV – Cohort Study
Ebert & Pham, 2017	51 bilingual children with primary language impairment; ages groups 5;6-8;11, 9;0-11;2	Narrative language sample measures derived from wordless picture book and raw scores from standardized tests in English and Spanish were examined for correlations. Sample measures included MLU_W, NDW, WPM, and grammatical accuracy of utterances.	Nine significant correlations were found for the younger group of children, and only one correlation was found for the older group, between raw scores from standardized tests and language sample measures. No complete overlap areas were found. Correlations decrease with age.	IV – Cohort Study
Ebert & Scott, 2014	73 school-age children; age groups 6;0-8;11, 9;0-12;8	Language sample measures collected from a wordless picture book narrative and 23 different standardized tests were examined for correlations. Sample measures included MLU_W, TNW, SI, NDW, omissions/errors at word, morpheme, and utterance level.	Correlation results indicated that for younger children, seven significant correlations were found, and for older children, four correlations were significant. Further analysis worked to classify children within normal limits and those with a language disorder. The rate of agreement between the measures was inconsistent (37-77%) based on different *SD* cutoff score.	IV – Cohort Study
Guo & Schneider, 2016	61 6-year-olds (50 Tl, 11 LI), 67 8-year-olds (50 TL, 17LI)	Narrative samples were collected from black and white pictures sequences (3 levels of complexities);	All measures were sensitive for diagnosis of LI at age 6, but only percent grammatical C-units had acceptable diagnostic accuracy at	IV – Case-Control Study

Guo & Schneider, 2016	61 6-year-olds (50 Tl, 11 LI), 67 8-year-olds (50 TL, 17LI)	Narrative samples were collected from black and white pictures sequences (3 levels of complexities); coded for verb morphology, errors, and percent grammatical C-units	All measures were sensitive for diagnosis of LI at age 6, but only percent grammatical C-units had acceptable diagnostic accuracy at age 8.	IV – Case-Control Study
Justice, Bowles, Kaderavek, Ukrainetz, Eisenberg, & Gillam, 2006	250 children, ages 5-12	Microstructure indices were drawn from the narrative samples of children. Samples were segmented into T-units and coded and analyzed for determining a comprehensive set.	Productivity and complexity were moderately related measures of microstructure. Productivity is measured by word output, diversity, and t-units. Syntactic levels of t-units measure complexity. Formulas for calculating performance were provided (INMIS score), a clinical tool for analyzing the microstructure of language samples.	IV – Cohort Study
Manolitsi & Botting, 2011	26 Greek children; 13 with ASD (age 4;2-13;0), and 13 with SLI (age 5;0-13;0)	Standardized measures of structural and pragmatic language were compared to micro- and macro-skills of structured narrative re-tell task across groups of children with ASD and SLI.	Children with ASD had lower receptive scores, but relatively equivalent expressive language score compared to children with SLI. Children with ASD performed significantly lower on expressive narrative tasks, with more divergent characteristics of strength and weakness. Comparisons of narrative measures to language samples showed no relationships within the SLI group, indicating different skills sets are measured.	IV – Cohort Study
Moyle, Karasinski, Weismer, & Gorman, 2011	50 school-aged children with SLI (age 5;5-9;8); 50 age-matched children with TL (age 6;0-9;9)	Conversational language samples (15-min) were collected via question-answer prompts. Samples were transcribed in SALT and analyzed for target morphemes, omissions, and errors. Verb-tense, noun morpheme, and MLU_m composites were calculated.	Scores for children with SLI were significantly lower for verb morpheme, noun morpheme, and MLU_m. Different between groups were significant, indicating that these variables can be used for classification of children with SLI vs. TL at 80% accuracy when combined.	IV – Case-Control Study
Spaulding, Plante, & Farinella, 2006	43 commercially available tests of child language	Tests were reviewed for their purposes of identifying language impairment.	Reviews of commercial language tests failed to meet the assumption that children with language impairment will routinely score at the low end of normal distribution of tests. For a majority of the tests, children with language impairment scored within 1.5 *SD*, with scores within 1 *SD* for 27% of the tests. Test	V – Case Report

Article Details
Number of Studies Included: 10
Years Included: 1998-2017

Levels of Evidence of Intervention Studies Suggested for Use in Communication Sciences and Disorder

Levels	Description
I	Systematic reviews and meta-analyses of randomized clinical trials and other well designed studies.
II	Double-blinded, prospective, randomized, controlled clinical trials.
III	Nonrandomized intervention studies.
IV	Nonintervention studies: • Cohort studies • Case-control studies • Cross-sectional surveys
V	Case reports
VI	Expert opinion of respected authorities.

Adapted from Cox, 2005

Understanding Research and Evidence-Based Practice in Communication Disorders: A Primer for Students and Practitioners William O. Haynes, *Auburn University* Carole E. Johnson, *Auburn University.*