2015

Evaluation of Teacher Ratings to Improve Child Language Screenings in Speech-Language Pathology

Kyomi Dana Gregory  
*Louisiana State University and Agricultural and Mechanical College*

Follow this and additional works at: [https://digitalcommons.lsu.edu/gradschool_dissertations](https://digitalcommons.lsu.edu/gradschool_dissertations)

Part of the Communication Sciences and Disorders Commons

**Recommended Citation**

[https://digitalcommons.lsu.edu/gradschool_dissertations/355](https://digitalcommons.lsu.edu/gradschool_dissertations/355)

This Dissertation is brought to you for free and open access by the Graduate School at LSU Digital Commons. It has been accepted for inclusion in LSU Doctoral Dissertations by an authorized graduate school editor of LSU Digital Commons. For more information, please contact gradetd@lsu.edu.
EVALUATION OF TEACHER RATINGS TO IMPROVE CHILD LANGUAGE SCREENINGS IN SPEECH-LANGUAGE PATHOLOGY

A Dissertation

Submitted to the Graduate Faculty of the Louisiana State University and Agricultural and Mechanical College in partial fulfillment of the requirements for the degree of Doctor of Philosophy

in

The Department of Communication Sciences and Disorders

by

Kyomi Dana Gregory
B.A., State University of New York at New Paltz, 2005
M.A., State University of New York at Buffalo, 2007
December 2015
This dissertation is dedicated to my mother, Cheryl Gregory who always told me “dare to be you.” You always told me that I would pursue my PhD even before I thought that this dream could be made a reality.
ACKNOWLEDGEMENTS

First I thank God for providing me with this opportunity. I know without God and strong faith that completion of the PhD journey would not be possible. Every step taken has been based on the guidance and strength of God lighting my path. My journey from Brooklyn, NY brought me all the way to Baton Rouge, LA with God showing me the way. I am blessed that you placed this dream in my heart and I am so grateful that I stayed the course on this journey.

To my mom, Cheryl Gregory, thank you for being my number one cheerleader. You have inspired me to reach my full potential and encouraged me along every journey that I have taken. I watched you sacrifice in order for me to succeed.

To my sister, Tamiko, thank you. You were there to listen when I wanted to pack up and come back to New York City. You have always been able to provide a word of encouragement. You let me know that my ability to tackle challenges inspired you and encouraged me to keep pushing forward.

Thank you to my dissertation committee: Dr. Janna Oetting, Dr. Janet Norris, Dr. Cassandra Chaney, and Dr. Mark Benfield. Dr. Oetting, thank you for being an excellent researcher and writer. You showed me that hard work and commitment always pays off. You have definitely challenged me to go further than I ever expected. Dr. Norris, thank you for showing me your passion for the profession of speech-language pathology and always stressing the importance of clinical implications in the research process. Dr. Chaney, thank you for being a breath of fresh air during the PhD process. I love your passion for what you do and how you are able to show that enthusiasm to students in the classroom.

I would like to thank Dr. Janet McDonald and Dr. Eugene Kennedy for helping me with some of the statistical analyses in this dissertation. I appreciate both of you for taking time out of your busy schedule to help me.
To my LDD lab mates and cohort, thank you for the fun times in the office and lab. Without everyone’s hard work this dissertation would not be possible. You all worked tirelessly administering testing and scoring data. I appreciate the opportunity to work with each of you.

To my mentors: Wendy Alleyne, Dr. Gwyneth Lloyd and Dr. Rosemary Lubinski, thank you. Without the three of you I would not be writing a dissertation. Wendy Alleyne you have mentored me since junior high school and have been with me every step along my journey. I appreciate your countless calls and prayers telling me to stay the course. Dr. Lloyd, thank you for pushing me from a BA to an MA and now to the pursuit of a PhD. The day you decided to mentor me, I had no idea you decided to make a lifetime commitment. Lastly, Dr. Lubinski thank you for personally taking time out of your schedule to let me know that I should pursue my PhD at the MA level. Your words of encouragement stayed with me for four years until I finally made the decision to go back to school.

To my cousin, Dr. Simone Gill-Alvarez, thank you. I was lucky to have someone in academia to be honest about the process. You prepared me in so many ways and set the example that it is possible to stay the course and complete a PhD. I appreciate you staying in touch with me throughout the process and providing advice. I am proud to have you as an example of what it means to be an exemplary scholar in academia.

Thank you to my speech-language pathology cheerleaders: Britt Morris, Raquel Payne, Cyreeta Dobbs, and Johanna Diaz. Each of you have spoken with me for hours and encouraged me along the journey. Your encouragement has made this dream a reality.
# TABLE OF CONTENTS

ACKNOWLEDGEMENTS ............................................................................................................................... iii

LIST OF ACRONYMS ........................................................................................................................................ vii

ABSTRACT .................................................................................................................................................... viii

CHAPTER ONE: INTRODUCTION ................................................................................................................. 1

CHAPTER TWO: REVIEW OF LITERATURE ...................................................................................................... 6
  Studies of Teaer Ratings ............................................................................................................................... 6
  Measures of Convergent Validity ................................................................................................................ 15
  The Diagnostic Evaluation of Language Variation- Screen (DELV-ST) .................................................... 16
  The Dynamic Indicators of Basic Early Learning Skills (DIBELS) Next .................................................. 18
  The Peabody Picture Vocabulary Test ..................................................................................................... 21
  Measures of Predictive Validity ................................................................................................................ 25
  The Diagnostic Evaluation of Language Variation- Norm Referenced (DELV-NR) ................................. 25
Purpose of the Current Study .......................................................................................................................... 26
Research Questions ...................................................................................................................................... 27

CHAPTER THREE: METHODS ..................................................................................................................... 28
  Participants ................................................................................................................................................. 28
  The Classroom Teachers .......................................................................................................................... 28
  The Child Participants .............................................................................................................................. 29
  Teacher Ratings ........................................................................................................................................ 33
  Tools Used to Evaluate Convergent Validity of Teacher Ratings ............................................................ 35
  Tools Used to Evaluate Predictive Validity of Teacher Ratings ................................................................ 37
  Procedures ................................................................................................................................................ 38
  Reliability .................................................................................................................................................. 38

CHAPTER FOUR: RESULTS .......................................................................................................................... 39
  Convergent Validity of Teacher Ratings .................................................................................................... 39
  Predictive Validity of Teacher Ratings: The DELV-NR ........................................................................... 50
  Combining Screeners to Improve Diagnostic Sensitivity and Specificity ................................................ 53
**LIST OF ACRONYMS**

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAE</td>
<td>African-American English</td>
</tr>
<tr>
<td>CCC-2</td>
<td>Children’s Communication Checklist-2</td>
</tr>
<tr>
<td>DELV-ST</td>
<td><em>Diagnostic Evaluation of Language Variation: Screening Test</em></td>
</tr>
<tr>
<td>DELV-NR</td>
<td><em>Diagnostic Evaluation of Language Variation Norm-Referenced Test</em></td>
</tr>
<tr>
<td>DIBELS</td>
<td><em>Dynamic Indicators of Basic Early Literacy Skills</em></td>
</tr>
<tr>
<td>GFTA-2</td>
<td><em>Goldman Fristoe Test of Articulation-2</em></td>
</tr>
<tr>
<td>MAE</td>
<td>Mainstream American English</td>
</tr>
<tr>
<td>PPVT-4</td>
<td><em>Peabody Picture Vocabulary Test- Fourth Edition</em></td>
</tr>
<tr>
<td>SLI</td>
<td>Specific Language Impairment</td>
</tr>
<tr>
<td>SLP</td>
<td>Speech Language Pathologist</td>
</tr>
<tr>
<td>SWE</td>
<td>Southern White English</td>
</tr>
<tr>
<td>TD</td>
<td>Typically Developing</td>
</tr>
<tr>
<td>TROLL</td>
<td><em>Teacher Rating of Oral Language and Literacy</em></td>
</tr>
</tbody>
</table>
ABSTRACT

The purpose of this study was to examine the validity of teacher ratings for screening children’s language skills. Teacher ratings were measured through the use of two tools, the Children’s Communication Checklist-2 (CCC-2; Bishop, 2006) and the Teacher Rating of Oral Language and Literacy (TROLL; Dickinson, McCabe, & Sprague, 2001). The data for this study were from 77 kindergarteners who lived in rural Louisiana and spoke a non-mainstream dialect of English; 51 were classified as typically developing and 26 as presenting with Specific Language Impairment. Convergent validity was examined by comparing the two teacher rating tools to each other and to three language and literacy screeners. Predictive validity was examined by comparing the teacher rating tools to the Diagnostic Evaluation of Language Variation: Norm Referenced (Seymour, Roeper, & de Villiers, 2005).

Although some results supported the convergent and predictive validity of the teacher ratings as measured by the two tools, neither tool nor the other screeners accurately identified a sufficient number of children with SLI. When empirically-derived cut scores for each screener were utilized, accuracy of the teacher ratings improved, and a discriminant function analysis with empirically-derived cut scores selected the Peabody Picture Vocabulary Test 4 (PPVT-4; Dunn & Dunn, 2007) and TROLL as the best tools to utilize in combination for screening purposes. This combination of tools accounted for 73% of the variation in the children’s DELV-NR scores. However, the PPVT-4 by itself was just as accurate as the PPVT-4 and TROLL together, with 95% of the children accurately screened (Sensitivity = .88; Specificity = .98).

For screening purposes, these findings support the use of teacher ratings as measured by the TROLL when empirically-derived cut scores are used and when the TROLL is combined with the PPVT-4.
CHAPTER ONE: INTRODUCTION

All children deserve an appropriate education that is able to meet their individual learning needs and prepares them for a successful future (National Education Association, 2007). Appropriate educational services include the provision of special education services for children with language and learning disabilities. Identifying children with language learning difficulties and intervening during the early school years can proactively address academic failure (Anderson et al., 2003; ASHA, 2000). Identifying and serving this group of children during the early school years has also been linked to increased graduation rates and enhanced productivity in life (Strickland & Riley-Ayers, 2006).

Within schools, the Speech-Language Pathologist (SLP) plays an integral role within the special education evaluation team. SLPs serve children with a variety of communication disorders, and this includes children with Specific Language Impairment (SLI). Children with SLI have receptive and expressive language deficits that are not attributed to cognitive deficits, speech disorders, autism, acquired brain damage, or hearing loss (Leonard, 2014; Stark & Tallal, 1981). Results from a large epidemiology study estimates the prevalence of SLI to be 7% in kindergarten, although the prevalence rate varies slightly by the race and ethnicity (White = 7%, Hispanic = 8%, African-American 11%) of the child (Tomblin et al., 1997).

By definition, children with SLI present weak language skills, and weak language skills have been linked to poor social competence and an increased risk for teasing, fewer friendships, and lower popularity with peers (Conti-Ramsden & Botting, 2004). Children with a history of speech and language impairment also frequently present later reading difficulties (Catts et al., 1999; Durand et al., 2013; Rescorla, 2002; Snowling et al., 2000; Stackhouse & Wells, 1997; Tallal et al., 1997). For example, in a study by Catts et al. (2006), 78% of children with poor language comprehension in third grade continued to show clinically significant impairments
when they were 13 to 14 years of age, and an additional 13% continued to show milder
difficulties. Also, a twenty-year follow-up study of children with speech-language impairments
showed poorer outcomes in communication, academics, educational attainment, and
occupational status in comparison to their peers without early language impairments (Johnson et
al., 2010). Finally, below grade level scores in reading or math has been shown to increase a
child’s risk for dropping out of school by 134% (Sparks, 2013), and 23% of students with speech
and language delays do not graduate from high school (U.S. Department of Education, 2005-
2006). For these many reasons, it is imperative to identify children with SLI during the early
school years because targeting these children for early intervening services may improve long
term reading outcomes, academic achievement, and quality of life.

The first step in determining a child’s eligibility for special education services is the
SLP’s use of a screener to determine if further testing is warranted. Teachers’ involvement in
the SLP’s screening process is often advocated to improve the accuracy of the screening (Fujiki
& Brinton, 1984; Marsh et al., 2006; Patterson & Wright, 1990; Whitworth et al., 1993).
Teachers observe children in their classrooms individually and as collective groups during a
wide range of academic and non-academic activities. Teacher ratings are also efficient in terms
of time and cost.

Teacher referral has been shown to be a strong predictor of eligibility for the broad
category of special education services. Studies show that 73% to 90% of students referred by
classroom teachers for evaluations based on academic performance are found eligible for special
education services (Algozzine, Christenson, & Ysseldyke, 1982; Gerber & Semmel, 1984;
Gottlieb, Alter, Gotlieb, &Wishner, 1994; Harry & Klinger, 2006; Pugach, 1985; Ysseldyke,
2001). However, Gottlieb and Wishner (1993) found that more than 60% of referrals made by
classroom teachers included children in the lowest quartiles of their classes academically (as cited in, Gottlieb, Alter, Gottlieb, & Wishner, 1994). Given this, teachers may miss children with less severe communication disorders, such as children with SLI, who need a referral for special education services (see also Voigt & Accardo, 2015 for similar difficulties that pediatricians have identifying children with language impairment). Evidence to support this assumption includes Tomblin and colleague’s (1997) finding that 71% of kindergarteners who present with SLI are not referred by parents or teachers for speech and language services (Tomblin et al., 1997).

Teachers may fail to notice the language and literacy weaknesses of children with SLI, especially when they come from culturally and linguistically diverse backgrounds. This is especially true for children who are speakers of various types of non-mainstream American English dialects, because clinicians have difficulties distinguishing dialect differences from language disorders (Seymour, Bland-Stewart, & Green, 1998). To date, no studies in the field of speech-language pathology have focused on teachers’ ratings of non-mainstream American English-speaking children’s language abilities. There is a need for this type of research, and kindergarten is an ideal grade at which to conduct this work. Kindergarten is a critical year for children because it is often their first encounter with school. For children who speak a non-mainstream American English dialect, kindergarten may also reflect their first encounter with a general American English dialect and the social norms and customs of individuals who speak general American English (Craig, Kolenic, & Hensel, 2013).

To address this need, this dissertation focused on teacher ratings of the language skills of kindergarteners who spoke one of two non-mainstream American English dialects, African-American English (AAE) and Southern White English (SWE). Specifically, this dissertation
examined the validity of teacher ratings. Teacher ratings were operationally defined as scores on two teacher-rating scales: the Teacher Rating of Oral Language and Literacy (TROLL; Dickinson, McCabe, & Sprague, 2001) and the Children’s Communication Checklist-2 (CCC-2; Bishop, 2006). Convergent validity was examined by comparing the two teacher-rating scales to each other and to three language and literacy screeners: the Diagnostic Evaluation of Language Variation: Screening Test (DELV-ST; Seymour, Roeper, & de Villiers, 2003a), the Dynamic Indicators of Basic Early Literacy Skills Next (DIBELS Next; Good, Gruba, & Kaminski, 2009), and the Peabody Picture Vocabulary Test 4 (PPVT-4; Dunn & Dunn, 2007). Predictive validity was examined by comparing the two teacher rating scales to the Diagnostic Evaluation of Language Variation: Norm Referenced (DELV-NR; Seymour, Roeper, & de Villiers, 2005).

The data for this study were selected from a large study that examined the grammar systems of kindergarten students (Oetting, McDonald, Hegarty, 2009-2015). Over 650 children contributed data to this dataset, and teachers of 77 of these participants completed the teacher rating forms. The 77 children demonstrate a range of language skills, with 51 classified as typically developing and 26 as SLI. All of these children were classified as speaking a non-mainstream dialect of American English, and all of these children also completed the standardized tests that were used to evaluate the teacher ratings.

As background for this study, the literature review is organized into three major sections. The first section focuses on studies that have examined the usefulness of teacher rating tools. The second section focuses on the tools used to examine convergent validity of the teacher ratings. The third section focuses on the tool used to examine the predictive validity of the teacher ratings. The chapter ends with the research questions that guided the study. Throughout this paper, a number of acronyms are utilized. To facilitate the readability of the paper, the
acronyms and their meanings are listed in the beginning of the document. They are also defined upon first mention within the document.
CHAPTER TWO: REVIEW OF LITERATURE

Studies of Teacher Ratings

At least eleven studies have examined the usefulness of teacher ratings to evaluate the speech and language skills of children. These studies are important to review, because they provided the literature base, which guided the present study. As will be evident, the eleven studies have varied in their methods and participant pools, and only two have included children from culturally and linguistically diverse backgrounds. Although findings from some of these studies support the use of teacher ratings, the level of support varies as a function of the methods and analyses conducted. Below, findings from each of these studies are detailed (see also the Appendix A for a summary of these findings in tabular form).

Botting, Conti-Ramsden, and Crutchley’s (1997) study included teacher and SLP judgments of 242 children, aged 6-8 years, who had been previously identified as having speech and language impairments. A total of 963 judgments were made about the children’s abilities in four areas: articulation, phonology, semantics/pragmatics, and syntax/morphology. According to Botting et al. (1997), these four areas were chosen to describe each child’s strengths and weaknesses. Interviews were used to collect the judgments of the teachers and SLPs. The interviews were informal in nature and utilized a semi-structured format. The following language assessments were also administered to the children: the Goldman-Fristoe Test of Articulation (Goldman & Fristoe, 1986), subtests of naming vocabulary and word naming from the British Ability Scales (Elliot, 1983), the Illinois Test of Psycholinguistic Ability (Kirk, McCarthy & Kirk, 1968), the Test for Reception of Grammar (Bishop, 1982), and The Renfrew Bus Story (Renfrew, 1991). Identification of impairment on the tests was determined using a 25th percentile cut-off. Children who scored at or below the cut-off were classified as impaired and
those who scored above this cut-off were classified as typical. Results indicated a 66% agreement level between the teacher and SLPs’ opinions of the children and the standardized tests in the domains of articulation, phonology, and syntax/morphology. However, the agreement level was low (41%) for the domain of semantics/pragmatics.

Bishop et al. (2006) also conducted a twin study of children, aged 6 years, to measure the effectiveness of parent and teacher ratings of communication skills using a dataset of twins. Thirty-two pairs of twins were classified as language impaired because at least one twin had low language skills, and 66 pairs of twins were classified as low risk because neither presented with a language difficulty. Parents and teachers completed the Children’s Communication Checklist (CCC; Bishop, 1998). The teacher’s overall ratings were based on the General Communication Composite (GCC) of the CCC, which is formed by summing all of the communication scales. The teachers’ overall rating scores were statistically higher for the low-risk children ($M = 216.1$, $SD = 11.00$) than for the children with language impairments ($M = 203.4$, $SD = 17.13$), and the effect size of the difference was large ($d = .82$). The parents’ overall rating scores were also statistically higher for the low-risk children ($M = 212.6$, $SD = 10.43$) than for the children with language impairments ($M = 201.0$, $SD = 16.10$), and the effect size of the difference was large ($d = .79$). These findings indicate that both teacher and parent ratings were effective at differentiating children with language impairment from children classified as low risk.

Cabell et al. (2009) examined teacher ratings of 209 children, aged 3-5 years. Forty-four teachers participated. The teacher scale focused on ratings of children’s emergent literacy skills based on an abbreviated, 12-item version of the Pre-Literacy Rating Scale from the Clinical Evaluation of Language Fundamentals Preschool – Second Edition (Wiig, Secord, & Semel, 2004). The children also completed four language assessments: the Preschool Word and Print
Awareness (Justice & Ezell, 2001), the Phonological Awareness Literacy Screening for Preschool (Invernizzi, Sullivan, Meier, & Swank, 2004), an upper-case alphabet recognition task, and a name-writing task. The overall accuracy of the teachers’ identification of language impairment was 79% (teachers accurately identified 165 children out of 209), with low sensitivity (27/52 = 52%) and higher specificity (138/157 = 88%). Sensitivity is the accuracy level of identifying children who are impaired as impaired, and specificity is the accuracy of identifying children who are typically developing as typical. According to Plante and Vance (1994), most researchers agree that values equal or greater than .90 are good and values between .80 and .89 are fair. These results indicate that the teachers were unable to accurately identify children with impairments as impaired and only fairly accurate in their ability to classify children without impairments as typical.

Jessup et al. (2008) found a similar finding when they examined the accuracy of teacher identification of speech and language impairment utilizing the Kindergarten Development Check (Office for Educational Review, 2003). This study was conducted in Tasmania, Australia. This study did not report the number of teachers who participated, but teacher ratings were collected on 286 kindergarteners, aged 4-5 years. The overall accuracy of the teacher’s identification of language impairment using the checklist was 71% (teachers accurately identified 203 children out of 286); however, this tool showed very low sensitivity (14/91 = 15%) and high specificity (189/195 = 97%). As was found by Cabell et al. (2009), the teachers were unable to adequately identify children with language impairments.

Williams (2006) examined the ability of teachers to identify language impairment in children who were enrolled in kindergarten or first grade. This study included 29 teachers across five schools in Western Australia. Unlike the previous studies reviewed, the teachers in this
study received a full day of in-service to build their knowledge of children with language impairment. Following the in-service, teachers were asked to identify children whose language skills were poor and not receiving services. Teachers were also asked to identify children who they thought presented typically developing language skills. Results indicated that the teachers identified 86% of the children with language impairment as impaired; however, they identified only 68% of the children who presented with typically developing language as typical. In other words, the training led to fair levels of sensitivity and low levels of specificity for the teachers’ ratings. This pattern of findings is the opposite of what was found by Cabell et al. (2009) and Jessup et al. (2008).

Gilmore and Vance (2007) explored teacher ratings of 52 children, aged 4-5 years, who failed at least one of two formal assessments of verbal comprehension and speech discrimination, the XAB task of the *Speech Input Processing in Children Battery* (Vance, Rosen, & Coleman, 2009) and the Linguistic Concepts subtest of the *Clinical Evaluation of Language Fundamentals- Preschool* (Wiig, Secord, & Semel, 2000). An 8-item questionnaire was used to investigate the teachers’ perceptions of the children’s attentive listening and verbal comprehension abilities. Results showed that the teachers’ ratings of the children’s verbal comprehension skills were positively correlated ($r = .31$) to the children’s verbal comprehension test scores and negatively correlated ($r = -.32$) to the children’s speech discrimination in noise. As can be seen by the $r$ values, while both correlations were statistically significant, the relationship between the teacher’s ratings and the children’s test scores were low in magnitude.

Boynton-Hauwewas and Stone’s (2000) study focused on 86 children, aged 5-7 years. Two groups of children participated: a group identified as having SLI and a group classified as typically developing. Teacher ratings were collected, and children completed two language tests:
the Expressive Language subtest of the *Curtiss-Yamada Comprehensive Language Evaluation* (CYCLE; Curtiss & Yamada, 1988) and the *Test for Auditory Comprehension of Language, Revised Edition* (TACL, Haynes, 1990). For the children with SLI, the teacher ratings were correlated to the children’s scores on the CYCLE ($r = .55$) and TACL ($r = .35$). Interestingly, however, the teachers’ ratings did not correlate significantly with the language test scores of the children with typical language on the CYCLE ($r = .12$) or TACL ($r = .06$).

Finally, Bedore, Pena, Joyner, and Macken’s (2001) study included 549 children, aged 4-11 years, who were selected from a pool of children participating in the norming study of the *Bilingual English Spanish Assessment* (BESA; Pena, Gutierrez-Clellen, Iglesias, Goldstein, & Bedore, In preparation). Bedore et al. (2001) collected parent and teacher questionnaires, and they administered subtests of the BESA to the children. Teachers provided information about the children’s relative use of Spanish/English in the classroom and their language proficiency on a scale of 0 to 4. Results showed that the teachers’ ratings had a low positive correlation ($r = .23, p < 0.001$) to the morphosyntax subtest of the BESA for all children, regardless of the children’s dominance in English or Spanish.

In summary, eight of eleven studies focusing on teacher ratings have been reviewed. One study focused on the agreement level between teacher and SLP judgment (Botting et al., 1997). This study found strong agreement between the teacher and SLPs opinion. Four studies focused on group differences and correlations, which led to findings that also support the use of teacher ratings in research and clinical practice (Bedore et al., 2001; Bishop et al., 2006; Boynton-Hauerwas & Stone, 2000; Gilmore & Vance, 2007). Three studies examined the diagnostic sensitivity and specificity of teacher ratings, with Cabell et al. (2009) and Jessup et al. (2008) reporting low sensitivity and fair to high specificity, and Williams (2006) reporting fair
sensitivity and low specificity. Findings from these three studies are less supportive of teacher ratings for screening than findings from the other studies.

Two teacher-rating tools that were examined in this dissertation were the CCC-2 and the TROLL. These tools were chosen based on the skill areas covered, the number of items, and past research related to these tools and SLI populations. The three studies reviewed next included these teacher rating scales within their analyses. Bishop and McDonald (2009) and Timler (2014) examined the ability of the CCC-2 to classify children who were at risk for language impairment, and Rodriguez and Guiberson (2011) examined the usefulness of the TROLL.

The CCC-2 is a tool designed to assess children’s communication skills in the areas of pragmatics, syntax, morphology, semantics and speech for children, aged 4-16 years. The manual for the CCC-2 reports that a General Communication Composite (GCC) score of 85 or lower is associated with .70 sensitivity (i.e., 70% of children who had SLI were identified as SLI) and .85 specificity (85% of children with typical development were identified as typical). The diagnostic accuracy of the tool is also reported in the manual through Positive Predictive Power (PPP) values and Negative Predictive Power (NPP) values. PPP is the likelihood that a person with positive test results actually has the disorder. NPP is the likelihood that someone with a negative test result does not have the disorder. Unlike sensitivity and specificity values, PPP and NPP values include consideration of the base rate of the condition in the population. Using a cut-off score of 85 and with a 50% prevalence rate in the population, the CCC-2 yields PPP and NPP values of .83 and .74, respectively. A PPP or NPP value that is above .90 is considered ideal. As will be evident in the next section of my review, these PPP and NPP scores,
while not ideal, are not necessarily lower than those that would be generated from many other assessment tools in the field.

Bishop and McDonald (2009) conducted a study with 245 children, aged 9-10 years. All of the participants were twins who came from the Twins Early Development Study (Trouton, Spinath, & Plomin, 2002). Children were classified into three groups: Not Referred (N = 193), Clinically Referred (N = 52), and the Other group (N = 20). The clinically referred group were referred to an SLP for assessment or intervention. The Other group had a statement of special educational need, but with no indication of a speech or language impairment. Parents completed the CCC-2, and the children completed several tests: the block design and vocabulary subtest of the Weschler Abbreviated Scale of Intelligence (Weschler, 1999), the understanding directions subtest of the Woodcock-Johnson III (Woodcock, McGrew, & Mather, 2001), the Expression, Reception and Recall of Narrative Instrument (ERRNI; Bishop, 2004), a developmental neurophysiological assessment, a sentence repetition task (NEPSY; Korkman, Kirk, & Kemp, 1998), a non-word repetition task, an oromotor sequence task, the Test of Word Reading Efficiency (Torgesen, Wagner, & Rashotte, 1999), and the Neale Analysis of Reading Ability-Second Revised British Edition (Neale, 1997).

Results were that the ERRNI, the NEPSY, and two indices from the CCC-2, the general communication composite (GCC) and the social interaction deviance composite (SIDC), yielded high sensitivity (Se = .91) but low specificity (Sp = .46) for the not referred and clinically referred group. From these findings, the authors argued that the SLP screening process should include combining measures to determine who should be referred for additional testing. However, it is important to note that the parent report indices from the CCC-2 did as well or better than the psychometric tests in distinguishing groups.
Timler (2014) examined the CCC-2 to classify children at risk for language impairment in children, aged 5 to 8 years. Forty-four children participated in the study, 32 children with Attention Deficit Hyperactivity Disorder (ADHD) and 12 children who were typically developing. The typically developing group was recruited from a community setting, and the children with ADHD were recruited from the university research clinic. Parents of the children completed the CCC-2. The children also completed *The Clinical Evaluation of Language Fundamentals, Fourth Edition* (Semel, Wiig, & Secord, 2003) and the *Test of Narrative Language* (Gillam & Pearson, 2004). Using a cut-off score of 85 on the CCC-2, sensitivity and specificity of the parent ratings were 1.0 and .85, respectively. This study and the one by Bishop and McDonald (2009) provide support for the use of the CCC-2 as a screening tool. However, one limitation was the recruitment of participants whose children were already classified as impaired and/or referred for services. As a result, parents of these participants were already concerned with one or more of their children’s communication needs, and this concern had been validated before they were asked to complete the CCC-2. This may have led to parents being more likely to identify their children’s weaknesses on the CCC-2 rating scale.

The TROLL is another teacher rating scale that can be found in the literature. It was developed to guide teachers’ observations of children’s language and learning skills. The TROLL includes oral language, reading, and writing scales. The TROLL developers reported that total scores were moderately correlated to children’s scores on three other measures of language, including the *Peabody Picture Vocabulary Test-III* (Dunn & Dunn, 1997) raw score (*r* = .47; *p* < .001), *The Emergent Literacy Profile* (Dickinson & Chaney, 1997) total score (*r* = .43; *p* < .01), and *the Early Phonemic Awareness Profile* (Dickinson & Chaney, 1997) total score (*r* = .45; *p* < .001). The TROLL also has been administered to 973 children in the context of research.
examining early literacy development, and data from 534 children was used to examine the tool’s internal consistency (Dickinson, McCabe, & Sprague, 2003). Cronbach alpha estimates of internal consistency for this tool ranged from .77 to .92 for separate subscales, and these values indicate strong internal consistency.

Rodriguez and Guiberson’s (2011) study included 353 English-speaking, Spanish-speaking, and English/Spanish bilingual children, aged 4 years. All of the children were developing language typically. The TROLL was used to collect the teachers’ ratings, and the Preschool Language Scale-4 (PLS; Zimmerman, Steiner, & Pond, 1992) was used to directly measure the children’s language abilities. Results for the TROLL varied by the child’s group. For the English-speaking children, results showed that the TROLL was significantly correlated with the PLS expressive ($r = .22, p = .001$) and the receptive ($r = .20, p = .002$) subtest scores. For the Spanish-speaking children, the TROLL was correlated with only the PLS receptive ($r = .22, p = .002$) subtest scores. For the bilingual children, the TROLL was not correlated with either the expressive ($r = .07, p = .444$) or receptive ($r = .05, p = .585$) language subscale of the PLS-4. Also, the English-speaking children scored higher on the TROLL in comparison to their Spanish-speaking and bilingual peers. Means and standard deviations on the TROLL total scores were: English-speaking children ($M = 54.93; SD = 12.86$), Spanish-speaking children ($M = 47.88; SD = 9.62$), and bilingual children ($M = 50.19; SD = 12.99$).

In addition, Rodriguez and Guiberson (2011) examined fail rates for the TROLL among the participants. A total score of 43 corresponded to the 10th percentile for the 4-year-olds in the norming sample that was collected by the test developers (Dickinson et al. 2001; p. 4). Approximately 20% of the English-speaking children, 35% of the Spanish-speaking children, and 32% of the bilingual children received a failing score of 43 or below. These results reflect
higher fail rates on the TROLL for children from bilingual and Spanish-speaking backgrounds, but manageable fail rates for children who speak English.

In summary, Bishop and McDonald (2009) reported high sensitivity and low specificity for the CCC-2, while Timler (2014) reported high sensitivity and specificity. Findings from Rodriguez and Guiberson (2011) indicated that the TROLL teachers’ ratings for English-speaking children were correlated to another language measure and led to manageable fail rates even though findings were less positive for Spanish-speaking and bilingual children. Given these findings, one can conclude that there is some support for further study of the CCC-2 and the TROLL as tools that can be used to collect useful information from teachers for screening purposes. In comparison to other teacher rating tools that have been examined in studies, both the CCC-2 and TROLL are readily available to clinicians and researchers. Both tools also include more items than the other tools reviewed. The CCC-2 focuses on all aspects of language development while the TROLL focuses on language, reading, and writing. In addition, the CCC-2 has been examined for diagnostic sensitivity and specificity, while the TROLL has not.

**Measures of Convergent Validity**

According to Thorndike (1997) convergent validity is demonstrated when “different ways of measuring the same trait yield high correlations” (p.161). Convergent validity is established if two similar constructs correspond with one another. This section focused on the screening tools (i.e. the DELV-ST, the DIBELS Next, and the PPVT-4) that were utilized in this dissertation to examine the convergent validity of the teacher rating tools. As will be evident from the literature review, these screening tools, like the teacher questionnaires, are not without limitations. Nevertheless, few screening tools exist within the field and of those available, the DELV-ST
DIBELS Next, and PPVT-4 are the most widely used and included within previous research studies.

**The Diagnostic Evaluation of Language Variation- Screener (DELV-ST).** The DELV series was developed as a dialect neutral child language measure (Seymour et al; 2003a, 2005). The DELV test series targets language structures that are spoken by both speakers of non-mainstream dialects of American English and speakers of General American English. Children from African American backgrounds were also oversampled in the norming population for the DELV-ST. The DELV test series includes two tools: the DELV-ST and the DELV-NR. Initially, a criterion-referenced version of the test (Seymour et al., 2003b) was published but it was replaced by the DELV-NR.

The DELV-ST is a 32-item screener that assesses two aspects of a child’s language system: the Degree of Language Variation and the Degree of Risk for a language disorder. The Degree of Language Variation section identifies the amount of a child’s Non-Mainstream American English usage. Children are asked to complete 15 items that target specific phonemes and grammar structures. Based on the children’s responses to these items, their dialects are classified as Mainstream American English (MAE), some variation from MAE, or strong variation from MAE. The Degree of Risk section includes 17 items and evaluates a child’s understanding of complex Wh-questions, the production of dialect-neutral grammar structures, and the ability to repeat nonwords. Based on the responses to these items, children are classified as lowest risk, low to medium, medium to high risk, or highest risk for a language disorder. Scores from the screening items for Degree of Risk can be used to decide if a comprehensive diagnostic is warranted.
Studies of the DELV-ST have been conducted by Oetting et al. (2009) and Moland (2011). Both studies focused on the usefulness of the DELV-ST in isolation and in comparison to other tools. Oetting et al. (2009) examined the pass/fail rates of the DELV-ST using data from 121 children, aged 4 years. Seventy-three children attended Head Start, a publicly funded preschool that is designed for children from low-income backgrounds, and 48 children attended private schools. Results showed that children identified as presenting strong variation from MAE varied by the school setting (Head Start = 62% vs. Private = 21%). In addition the children differed in their risk for impairment by school setting (Head Start = 44% vs. Private = 15%). The correlations between the children’s dialect ratings and risk for impairment were statistically significant, but were relatively low ($r = .27$). This finding indicates that the children’s dialect ratings accounted for a small percentage (7%) of the variance within the children’s risk scores. This finding in isolation does not refute the use of the DELV-ST for children who speak a non-mainstream dialect of American English dialect, but it does question the dialect neutral nature of the items included on the screener.

Finally, Moland (2011) evaluated 73 African-American (AA) children who spoke AAE, aged 4 to 5 years, who attended publically funded Pre-K and Head Start programs. The children were administered the DELV-ST, the *Fluharty Preschool Speech and Language Screen Test-Second Edition* (Fluharty-2; Fluharty, 2001), and the *Washington-Craig Language Screener* (WCLS; Craig & Washington, 2000, 2002; Craig, Washington, & Thompson-Porter, 1998; Washington & Craig, 2004). Fail rates were highest for the Fluharty-2 (57%), lower for the DELV-ST (52%), and lowest for the WCLS (46%). A total of 18 children (25%) passed all three screening tools, and 21 children (29%) failed all of the screenings. In other words, the screeners led to identical clinical outcomes for 54% of the children. For the remaining 46%, they either
failed only one screening tool (24%) or two screening tools (22%), which indicates some inconsistency across the screeners.

To further explore the inconsistency across the three screeners, Moland examined the role that the test order played in the results. Of the 33 children who failed one to two screeners, 14 children (42%) failed the first screener administered, 16 children (49%) failed the second screener, and 3 children (9%) failed the third screener. These findings suggest that the children’s performance may have improved with practice. Fortunately, practice effects would have influenced the fail rates of all three screeners in the same way because the order of the screeners was counterbalanced. For the purpose of the current study, Moland’s findings suggest that the DELV-ST was comparable to the other screeners. The findings from these two studies suggest that the DELV-ST is at least equivalent to others that could be selected to examine the convergent validity of teacher ratings.

The Dynamic Indicators of Basic Early Learning Skills (DIBELS) Next. The DIBELS Next (which is a revised version of the original DIBELS) was used to evaluate the convergent validity of the teacher ratings in this dissertation study. The DIBELS Next is appropriate for children in kindergarten through sixth grade. It is administered three times throughout the school year: at the beginning of the school year, between months one through three, in the middle, between months four through six, and at the end of the school year, between months seven through nine. The DIBELS and DIBELS Next are Curriculum-Based Measurement (CBM) tools. CBM was developed as a formative assessment for adapting teaching to meet students’ needs (Deno, 1985; Kaminski & Cummings, 2007). CBM tools provide additional information to teachers to establish student goals and monitor reading
progress (Fuchs & Fuchs, 1986; Taylor, Roehrig, Soden Hensler, Connor, & Schatschneider, 2010).

The DIBELS and DIBELS Next tools provide criterion-referenced benchmark scores that represent adequate reading progress. If a student attains a positive benchmark score, then the student is likely to achieve positive reading outcomes in the future (Dewey, Kaminski, & Good, 2014). The DIBELS is utilized for screening and progress monitoring to identify students who require additional academic support. Multiple studies have investigated the reliability, validity, and usefulness of the DIBELS and the DIBELS Next (Cummings, Park, & Bauer Schaper, 2013). Four studies of the DIBELS and/or DIBELS Next were reviewed. The studies are by Kaminski and Good (1996), Elliot, Lee, and Tollefson (2001), Shaw and Shaw (2002), and Good, Powell-Smith, Murdoch, and Latimer (2011).

Kaminski and Good (1996) conducted a study of 38 kindergarten and 40 first grade children from general education classrooms. The children were administered three DIBELS measures (twice during the school year): phonemic segmentation fluency, letter naming fluency, and picture naming fluency. In addition, several criterion measures were administered: a CBM reading probe as described by Marston (1989), McCarthy Scales of Children’s Abilities (McCarthy, 1972), The Metropolitan Readiness Test (Nurss & McGauvran, 1986), The Stanford Diagnostic Reading Test (Karlen & Gardner, 1985), The Rhode Island Pupil Identification Scale (Novak, Bonaventura, & Merenda, 1973), and a teacher rating scale. The teacher rating scale included a 5-point Likert scale for various aspects of reading: reading readiness/achievement, rate of progress in reading readiness/achievement, and level of risk for development of later reading difficulties. For the kindergarten children, moderate to high positive correlations were found for all DIBELS measures with all criterion measures ($r$ values ranged from .43 to .90, $p <$
.01). Fewer significant positive correlations were seen between the DIBELS scores and the criterion measures for the first grade children.

Elliot, Lee, and Tollefson (2001) conducted a study of 75 kindergarten children (63% White and 37% non-White) from four classrooms in three elementary schools. The children were repeatedly administered a modified version of the DIBELS. This version included Letter Naming Fluency, Sound Naming Fluency, Initial Phoneme Ability, and Phoneme Segmentation Ability. The testing occurred for nine-weeks in two-week intervals. Three types of reliability were measured: inter-rater reliability, test-retest reliability, and alternate form reliability. Interrater reliability for individual subtests ranged from 82% to 94%. Test-retest reliability for each subtest was between 74% and 93%. Finally, equivalent forms for reliability for the individual subtests ranged from 64% to 91%.

In addition, this study examined the concurrent validity of the DIBELS scores over repeated administrations as measured against other criterion measurements. The measures were: the Kaufman Brief Intelligence Test (K-BIT; Kaufman & Kaufman, 1990), the Test of Phonological Awareness (TOPA; Torgeson & Bryant, 1994), the Developing Skills Checklist (DSC; CTB Macmillan/McGrawHill, 1990), the Woodcock-Johnson Psychoeducational Achievement Battery-Revised (WJ-R; Woodcock & Johnson, 1989, 1990), and an informal teacher pre-reading rating questionnaire. Correlations between the modified DIBELS and the K-BIT ranged from .36 to .59. Correlations between the modified DIBELS, the TOPA ($r = .69$), the DSC ($r = .74$), and the teacher’s pre-reading rating questionnaire ($r = .67$) ranged from .67 to .74. Correlations between the DIBELS and the WJ-R ranged from .62 to .81.

Shaw and Shaw (2002) examined the convergent validity of the DIBELS in relation to the reading portion of the Colorado State Assessment Program (CSAP). Fifty-two participants
in the third grade took the DIBELS Oral Reading Fluency (ORF) subtest three times throughout the school year as well as the CSAP in the spring. Correlations ranged from .73 (in the fall/winter) to .80 (in the spring). Based on these findings, the authors argued that the DIBELS springs score is a strong predictor of spring CSAP scores. In addition, when the CSAP was utilized as an outcome measure, the DIBELS ORF (administered in the Spring) with a cut score of 90 presented with a sensitivity and specificity of .73 and .91, respectively. Recall that values above .90 are considered good and those above .80 are considered fair. While not ideal, similar levels of sensitivity and specificity have been reported for different teacher rating scales.

Finally, Good et al. (2011) examined the extent to which the DIBELS Next measures predicted outcomes on the Group Reading Assessment and Diagnostic Evaluation (GRADE). The study included 3,816 participants in kindergarten to sixth grade from 13 schools in five school districts within five US regions. Results were that the DIBELS Composite Score (administered in the Fall) in kindergarten was moderately correlated \( r = .50 \) with the GRADE total test scores at the end of the school year.

Findings from the DIBELS studies reviewed showed moderate to high correlations with other criterion-referenced measures of children’s reading abilities. The DIBELS measures were also reported to be highly reliable. This information supports the use of DIBELS for kindergarteners to determine who is and is not at-risk for reading failure. For these reasons, DIBELS is an appropriate tool for evaluating the convergent validity of teacher ratings.

**The Peabody Picture Vocabulary Test.** The PPVT remains the most widely used test of receptive vocabulary abilities in the field of speech-language pathology (Rice & Watkins, 1996; Eickhoff, Betz, & Ristow, 2010). This tool has also been recommended as part of a language screener (Moland, 2011; Washington & Craig, 2004). However, since its first
publication in 1959, this tool has faced criticism as being inappropriate for children and adults from culturally and linguistically diverse backgrounds (Stockman, 2000). Specifically, the first edition (1959) did not include participants from minority cultures within its normative samples. The revised edition (PPVT-R) was published in 1981 and contained changes to the normative sample, which included eliminating items that were culturally offensive or biased. Despite the revisions to the PPVT-R, this version was still considered inappropriate for use with culturally and linguistically diverse populations (Argulewicz & Abel, 1984; Halpin, Simpson, & Martin, 1990; Washington & Craig, 1992). The third edition of the PPVT was published in 1997. Stockman (2000) noted differences between the PPVT-R and PPVT-3 (Dunn & Dunn, 1997) that included the number of words sampled, the number of age levels sampled, and none of the visual stimuli utilized to elicit words. In addition, the normative sample of the PPVT-3 included a greater number of children from minority backgrounds (34% versus 15%) and a greater number of children with different types of disabilities (10% versus none) than the PPVT-R.

Five studies summarized in this section have examined the PPVT-3. These studies were conducted by Washington and Craig (1999), Champion, Hyter, McCabe, and Bland-Stewart (2003), Webb, Cohen, and Schwaneflugel (2008), Platt (2010), and Gray, Plante, Vance, and Henrichsen (1999). Washington and Craig (1999) conducted a study in which the PPVT-3 was administered to 59 African American children, aged 3-4 years, who were enrolled in Head Start in Detroit. Results showed that the preschoolers scored significantly higher on the PPVT-3 ($M = 91; SD = 11$) in comparison to the PPVT-R ($M = 80.2; SD = 19.1$) (Washington & Craig, 1992). The performance of the participants also resulted in a normal distribution of PPVT-3 scores. Differences in performance based on gender and income were not seen. Although the level of education of the primary caregiver was significantly related to the children’s scores, effects of
maternal education on children’s language abilities were found for both minority and majority cultures. From these findings, the authors concluded that the PPVT-3 was a culturally fair and appropriate language measure for African American children.

Champion et al. (2003) conducted a study of the PPVT-3 with 49 African American preschoolers enrolled in Head Start, aged 3-5 years. Results indicated that AA children performed at 1 SD below the mean ($M = 86.84; SD = 10.96$) on the PPVT-3, which was significantly different from the normative sample for the test ($M = 100; SD = 15$). However, only these items (porcupine, camper, furry) were missed by 50% or more of the children. As a result, adjusting individual scores by omitting these specific items had little impact on a child’s score. In other words, an item analysis did not reveal a bias in the PPVT-3 test. Given this, the authors supported the use of the PPVT-3 for African-American children enrolled in Head Start programs.

Webb et al. (2008) studied 210 prekindergarten children (AA= 121, White = 89), aged 4 to 5 years. Children were administered the PPVT-3. The researchers conducted differential item functioning (DIF) analysis, which allows researchers to identify items on tests that have unequal scores between groups (i.e., gender, race). Results of the DIF analysis indicated that only one item on the PPVT was detected that could be attributed to the presence of the children’s race. Examination of this item indicated that it favored African American children. As a result, differences in scores between the two groups on the PPVT-3 could not be attributed to a race bias. Based on this finding, the authors concluded that the PPVT-3 is a culturally fair assessment tool.

Platt (2010) examined the appropriateness of the PPVT-3 for children, aged 4 years, who were classified as either primary speakers of English (EL1) ($N = 42$) or English Language
Learners (ELL) (N = 73). The children lived in Ontario, Canada, and children classified as ELL spoke diverse first languages including Korean, Urdu, Farsi, and Mandarin. Significant differences were found between the two groups in scores on the PPVT-3, with the EL1 (M = 108.90; SD = 15.75) group performing significantly higher than the ELL (M = 92.79; SD = 16.82) group. A Differential Item Functioning (DIF) analysis was conducted to assess how individual test items functioned for the two language groups. Results of the DIF analysis indicated that the differences between the language groups were not related to item functioning. Overall, the authors concluded that the PPVT-3 is sensitive to the vocabulary gap that exists between EL1 and ELL children in preschool.

Finally, Gray et al. (1999) examined the use of four vocabulary tests- the PPVT-3, the Receptive One-Word Vocabulary Test (Gardner, 1985), the Expressive Vocabulary Test (Williams, 1997), and the Expressive One-Word Vocabulary Test-Revised (Gardner, 1990)- to screen and identify children with SLI. The study was conducted with 62 children (SLI = 31, typically developing = 31), aged 4-5 years. The PPVT-3 scores for the children with SLI (M = 97; SD = 14) were lower than the typically developing children (M = 112; SD = 12), with a relatively large effect size (d = .89). Correlations between the PPVT-3 and the other tests ranged from .44 - .78 for children classified as having SLI and from .45 - .77 for the typically developing group.

The studies in this section indicate that the PPVT-3 is an appropriate test for culturally and linguistically diverse populations. Given this, the tool is appropriate to examine the convergent validity of teacher rating scales.
Measures of Predictive Validity

Predictive validity refers to the degree to which a score can predict other measures of the same construct. Studies of predictive validity often ask whether scores from a measure or test accurately predict the results of a measure or test that is given at a later time. This section focuses on the DELV-NR which was utilized in this dissertation to examine the predictive validity of the teacher rating tools.

The Diagnostic Evaluation of Language Variation- Norm Referenced (DELV-NR).

The normative sample of the DELV-NR includes data on 1,258 children (51% male and 49% female) who ranged in age from 4 to 12 years and were from four regions of the United States: the South, Northeast, North Central, and the West (Seymour, Roeper, & de Villiers, 2005). Of the children within the normative sample, 37% were previously identified as General American English speakers and 63% were previously identified as speakers of a non-mainstream dialect of American English speakers.

The manual for the DELV-NR reports that a DELV-NR Norm Referenced Composite score that is -1 standard deviation below the mean is associated with .95 sensitivity (i.e. 95% of children who had SLI were correctly identified as SLI) and .93 specificity (i.e. 93% of children with typical development were identified as typical). Studies focusing on the DELV-NR have been conducted by Pearson, de Villiers, Magaziner, Perisho, and Sutherland (2005) and Pearson, Jackson, and Wu (2014).

Pearson et al. (2005) examined the concurrent validity of the DELV-NR with children’s language samples. The study included 78 children (58 typically developing and 20 LI), aged 4-9 years. The authors examined the following language sample measures: total number of utterances, number of different words, mean length of utterances in words and morphemes,
sentence structure, a syntactic complexity profile, and an extended discourse pragmatics composite. Results showed a significant relationship between DELV-NR total scores and measures generated from the language samples that ranged from .29 to .66. Using a cut-off score of 1.5 standard deviations below the mean, the authors also showed that the children who passed the DELV-NR scored significantly higher than children who failed the DELV-NR on all of the language sample measures.

A later study by Pearson, Jackson, and Wu (2014) further examined the data from the Pearson et al. (2005) study of the DELV-NR. Results of this study indicate that accuracy of the DELV-NR in comparison to language sample profiles was greater than that of prior diagnoses. Estimates for the sensitivity and specificity for the DELV-NR were .85 and .93, respectively. These results indicated fair sensitivity and high specificity of the DELV-NR.

The DELV-NR studies support the use of the tool for children from culturally diverse backgrounds. The test also shows good sensitivity and specificity as reported in the test manual and as examined by Pearson et al. (2005; 2014). These results indicate that the DELV-NR is a viable tool to evaluate the predictive validity of teacher ratings.

**Purpose of the Current Study**

The purpose of this study was to examine the validity of teachers’ ratings of children’s language skills. To do this, teacher ratings were measured through the use of two tools, the CCC-2 and the TROLL. These tools were examined for their convergent and predictive validity with direct measures of children’s language and literacy skills.
Research Questions

1. Are the children’s scores on the CCC-2 and TROLL affected by their clinical status and by other sociodemographic variables (nonmainstream dialect use, school, gender) in a way that is consistent with each other and with the DELV-ST, DIBELS, and the PPVT-4?

2. Are the children’s scores on the CCC-2 and TROLL correlated with each other and to the children’s scores on the DELV-ST, DIBELS, and the PPVT-4?

3. What percentage of children pass and fail the CCC-2 and TROLL relative to each other and to those who pass and fail the DELV-ST, the DIBELS, and the PPVT-4?

4. Are the children’s scores on the CCC-2 and TROLL affected by their clinical status and by other sociodemographic variables (nonmainstream dialect use, school, gender) in a way that is consistent with the effects of these variables on the DELV-NR?

5. Are the children’s scores on the CCC-2 and TROLL correlated with the children’s scores on the DELV-NR?

6. What percentages of children pass and fail the CCC-2 and TROLL relative to those who pass and fail the DELV-NR?

7. If low predictive validity of the teacher ratings are revealed, does predictive validity increase if empirically derived cut-off scores are established and the teacher ratings are combined with the other screening tools?
CHAPTER THREE: METHODS

Participants

The Classroom Teachers. Eight female teachers participated in the study from four different schools in the same district, which was located in a rural parish of Southeastern, LA. The United States Census from 2008 – 2012 for this parish indicates a median household income of $46,410 and 18.8% of individuals living below the poverty line (U.S. Bureau of the Census, 2014). One teacher was African-American and the others were White. Each served as the lead teacher of a regular kindergarten classroom at her school. The teachers were assigned questionnaires for students in their classrooms who were participating in the larger study. The number of questionnaires completed by each teacher ranged from 4 to 16 (total = 77). See Table 1 for information on the classroom teachers who participated in the study. Four of the teachers had been teaching for five or less years, and four had been teaching for 15 years or more. Two independent-sample t-tests were conducted to compare the CCC-2 and TROLL scores based on the teachers’ years of experience. There was no significant difference in scores based on the teachers’ years of experience for either the CCC-2 ($p = .66$) or TROLL ($p = .93$).

Table 1. Demographics of the Classroom Teachers

<table>
<thead>
<tr>
<th>Teacher</th>
<th>Gender</th>
<th>Race</th>
<th>School</th>
<th># of Students/Questionnaires Completed</th>
<th>Years of Experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teacher 1</td>
<td>F</td>
<td>W</td>
<td>School A</td>
<td>16</td>
<td>20+</td>
</tr>
<tr>
<td>Teacher 2</td>
<td>F</td>
<td>W</td>
<td>School B</td>
<td>14</td>
<td>0 - 5</td>
</tr>
<tr>
<td>Teacher 3</td>
<td>F</td>
<td>W</td>
<td>School B</td>
<td>13</td>
<td>20+</td>
</tr>
<tr>
<td>Teacher 4</td>
<td>F</td>
<td>W</td>
<td>School B</td>
<td>7</td>
<td>0 - 5</td>
</tr>
<tr>
<td>Teacher 5</td>
<td>F</td>
<td>W</td>
<td>School C</td>
<td>4</td>
<td>20+</td>
</tr>
<tr>
<td>Teacher 6</td>
<td>F</td>
<td>W</td>
<td>School C</td>
<td>6</td>
<td>15 - 20</td>
</tr>
<tr>
<td>Teacher 7</td>
<td>F</td>
<td>AA</td>
<td>School D</td>
<td>11</td>
<td>0 - 5</td>
</tr>
<tr>
<td>Teacher 8</td>
<td>F</td>
<td>W</td>
<td>School D</td>
<td>6</td>
<td>0 - 5</td>
</tr>
</tbody>
</table>

Note. F = Female. AA = African American. W = White.
The Child Participants. The 77 child participants from whom the questionnaires were completed reflected a subset of 669 children who participated in the larger study. The 77 children averaged 65.35 months (SD = 4.00) which was similar to the average age of 65.81 (SD = 3.95) of the children in the larger study. In addition, like all of the children in the larger study, those who contributed data to the current study presented with normal hearing as confirmed by a pass on a pure-tone hearing screening at 30 dB for 1, 2, and 4 Hertz in each ear. Thirty-nine (51%) of the 77 children were males, which was similar to the 51% of the children in the larger sample. The race and ethnicity of the 77 children per caregiver report was: 29 (38%) African American, 41 (53%) White children, and 7 (9%) Other (American Indian, Mixed, or Not Reported by caregiver but classified as African American based on school records). Like gender, the race and ethnic distribution of the current sample was similar to the larger sample, which was 32% African American, 63% White, and 5% Other.

Of the 77 participants, 35 children were classified as African American English speakers (29 classified as African-American, 3 classified as Mixed; 3 classified as not reported by caregiver but school records indicated African American) and 42 children were classified as Southern White English (SWE) speakers based on their race (41 classified as White, 1 classified as American Indian). In previous studies in this area of Louisiana, race has been shown to be highly correlated with blind listener judgments of the children’s nonmainstream dialect type (Oetting & McDonald, 2002; Oetting & Richardson, 2011). To confirm the relationship between the children’s race and their dialect, the nonmainstream dialect statuses of the children were further examined through the DELV-ST Dialect Variation subtest. The DELV-ST allows for the classification of each child as a speaker of either MAE, some variation from MAE, or a strong variation from MAE. Each child’s responses on the DELV-ST can also be converted to a
percentage of nonmainstream responses, which is often referred to in the literature as a Dialect Density Measure (DDM). See Table 2 for classification of the children’s dialects based on the DELV-ST scoring system. As is evident, more children in the AAE group earned ratings of strong dialect variation and more children in the SWE group earned mainstream dialect ratings. The AAE group’s percentage of nonmainstream responses \((M = .84; SD = .18)\) was also statistically higher than the percentage obtained by the SWE group \((M = .46; SD = .29)\); an independent sample t-test indicated a significant difference between the two groups \((p < .001)\). These findings are consistent with other studies of AAE and SWE child speakers in Louisiana (Oetting & McDonald, 2002).

Table 2. DELV-ST Degree of Dialect Variation.

<table>
<thead>
<tr>
<th>Degree of Variation</th>
<th>SWE = 42</th>
<th>AAE = 35</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mainstream</td>
<td>25 (60%)</td>
<td>3 (9%)</td>
<td>28 (37%)</td>
</tr>
<tr>
<td>Some Variation from Mainstream</td>
<td>7 (17%)</td>
<td>5 (14%)</td>
<td>12 (16%)</td>
</tr>
<tr>
<td>Strong Variation from Mainstream</td>
<td>10 (24%)</td>
<td>27 (77%)</td>
<td>38 (48%)</td>
</tr>
<tr>
<td>DELV-ST DDM</td>
<td>.46</td>
<td>.84</td>
<td>.63</td>
</tr>
</tbody>
</table>

Finally, listener judgments of the children’s dialects had been collected on 40 of the 77 children as part of the larger study; these data were analyzed as part of the current study to test the validity of race to determine each child’s type of American English dialect. Results were that two out of three blind listener judges agreed on 85% of the children’s nonmainstream dialect type (AAE or SWE), and these dialects aligned with the children’s race (African-American vs. non-African-American). This rate of agreement is also consistent with other studies of AAE and
SWE child speakers in Louisiana (Oetting & McDonald, 2002). See Table 3 for the percentage of listener judgment agreements based on the children’s type of American English dialect. These findings help confirm the validity of the children’s dialect classifications.

Table 3. Listener Judgment Agreements

<table>
<thead>
<tr>
<th>Dialect Type</th>
<th>% Agreement 2/3 Listener Judgments</th>
</tr>
</thead>
<tbody>
<tr>
<td>SWE</td>
<td>12/15 (80%)</td>
</tr>
<tr>
<td>AAE</td>
<td>22/25 (88%)</td>
</tr>
<tr>
<td>Total</td>
<td>34/40 (85%)</td>
</tr>
</tbody>
</table>

For the purpose of the current study, 26 of the 77 children were classified as SLI, and 51 were classified as TD. The clinical status of the participants (SLI vs. TD) was determined through a review of the children’s family and academic profiles and test scores that had been collected as part of the larger study. Regarding the children’s family profiles, caregivers were asked to report if other members of their family (siblings, parents, grandparents, aunts and uncles) presented with speech, language, or reading impairments as part of the parental consent for their child to participate in the study. Caregivers who reported other family members with impairments were classified as presenting a positive family history, and those who reported no impairments within their family were classified as presenting a negative family history. Although the families of 31 of the children did not complete the family history question on the consent form, of those that did, 7 (27%) of the SLI group and 9 (18%) of the TD group reported a positive family history of speech, language, or reading impairment. Higher rates of a positive family history for children with SLI is consistent with other studies that have shown a positive
family history to be ~two times higher in children with SLI as compared to children without SLI (e.g., Rice, Haney, & Wexler, 1998; Pruitt, Garrity, & Oetting, 2010).

Regarding the children’s academic profiles, 5 (19%) children in the SLI group and none of the children in the TD group were receiving services by a speech-language pathologist. Although only 19% of those with SLI were in receipt of services, low rates of children receiving services is consistent with a large epidemiology study by Tomblin et al. (1997) who found that less than 30% of kindergartners who qualify for the diagnosis of SLI receive SLP services. In addition, 16 (62%) of the children classified as SLI within the current study have also been included in two other language studies that have found statistically significant differences between those with and without SLI (McDonald, Seidel, Hammarlund, & Oetting, under review; Oetting, McDonald, Seidel, & Hegarty, in press). Finally, in support of the clinical status of the participants, the percentage of children classified as SLI within the larger study (53/669 = 8%) is consistent with the 7% prevalence rate of SLI in kindergarten as documented by the epidemiology study by Tomblin et al. (1997).

Regarding test scores of the children, all of the children were administered the Primary Test of Nonverbal Intelligence (PTONI; Erhler & McGhee, 2008) and the Goldman Fristoe Test of Articulation-2 (GFTA-2; Goldman & Fristoe, 2000). The PTONI assesses reasoning abilities in young children and is often used as an index of children’s nonverbal IQ. It was normed on a sample of 1,010 children, ages 3 to 9 years of age, residing in 38 states. Testing takes approximately 5 to 15 minutes. It includes minimal oral directions and a pointing response format. The items are arranged in order of difficulty. Early items measure lower order reasoning and later items focus on higher order reasoning abilities. A standard score of 100 (SD = 15) is considered within normal limits.
The GFTA-2 assesses an individual’s articulation of English consonants. This tool is appropriate for children ages 3 through 21 years and requires 5 to 15 minutes to be administered. It samples both spontaneous and imitative sound production. For the current study, only the single word subtest was administered. Children provide a single word response to a picture. A standard score of 100 (SD = 15) is considered within normal limits.

The children’s test scores by their clinical status are reported in Table 4. As is evident by the ranges reported, all of the participants scored above -1 SD on the GFTA-2, which indicated that their articulation abilities were in the average range. However, 16 children (14 SLI and 2 TD) scored below – 1 SD on the PTONI. Traditionally, children with SLI and children classified as TD within SLI studies are required to present with a nonverbal IQ (i.e., PTONI) score that is within or above -1 or -2 SD of the normative mean. Of those 16 who earned PTONI scores below – 1 SD, all but two fell within or above -2 SD. The two children who did not were both in the SLI group, and they earned PTONI standard scores of 53 and 68. Both of these children may be inappropriately classified as SLI in the current study; however, they were kept in the analysis because their vocabulary test scores were higher than their PTONI scores (making their PTONI scores suspect), and neither of these children nor the others with low PTONI score were identified as presenting with a cognitive disability based on school reports.

**Teacher Ratings**

CCC-2. Teachers returned a CCC-2 questionnaire for 75 of the 77 children. The CCC-2 is a 70-item measure designed to assess children’s communication skills in the areas of pragmatics, syntax, morphology, semantics, and speech for children, aged 4 to 16 years. The 70 items are divided into 10 scales that are on a Likert scale (speech, syntax, semantics, coherence, initiation, scripted language, context, nonverbal communication, social relations, interests) and
each scale contains seven items each. Out of the seven items in each scale, five of the items address difficulties and two focus on the child’s strengths. The tool has three main sections: (1) the first four scales address articulation, phonology, language structure, vocabulary, and discourse, (2) scales five through eight address pragmatic aspects of communication, and (3) the last two scales assess behaviors commonly impaired in children with Autism Spectrum disorder.

The tool takes approximately 10 to 15 minutes to complete.

Table 4. Means, (SD), Range for PTONI and GFTA -2 based on Clinical Status

<table>
<thead>
<tr>
<th>Clinical Status Groups</th>
<th>PTONI</th>
<th>GFTA-2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean (SD)</td>
<td>Range</td>
</tr>
<tr>
<td>SLI</td>
<td>85.38 (11.28)</td>
<td>53-105 (5.21)</td>
</tr>
<tr>
<td></td>
<td>53-105</td>
<td>91-114</td>
</tr>
<tr>
<td>TD</td>
<td>106.08 (12.91)</td>
<td>84-140 (4.25)</td>
</tr>
<tr>
<td></td>
<td>84-140</td>
<td>94-116</td>
</tr>
<tr>
<td>Total</td>
<td>99.09 (15.76)</td>
<td>108.45 (4.99)</td>
</tr>
<tr>
<td></td>
<td>53-140</td>
<td>91-116</td>
</tr>
</tbody>
</table>

The CCC-2 yields one composite score - the General Communication Composite (GCC). The GCC is a norm referenced standard score (M = 100, SD = 15) that reflects the overall communication skills. The cut off score for the GCC is a standard score of 85. In this study, participants presented with a GCC mean standard score of 102.05 (SD = 17.08). The scores ranged from 52 to 139. This score is calculated by summing up 8 of the scales.

TROLL. The TROLL was developed to track the language skills of children, aged 3 to 5 years. Although the children in the current study were slightly older than 5 years, previous studies of 5 year olds indicate that children of this age have not scored at ceiling on this tool. The tool has 25 items and has a 4-point Likert scale. Total scores on the TROLL vary from a
minimum of 24 to a maximum of 98. The total score provides the researcher and clinician with an indication of a child’s development relative to other children as assessed by the child’s teacher. The cut-off score for the TROLL is 65 out of 98. In this study, participants presented with a TROLL mean score of 87.04 (SD = 12.34). The scores ranged from 41 to 98.

**Tools Used to Evaluate Convergent Validity of Teacher Ratings**

*DELV-ST.* Part II of the DELV-ST screens children’s risk for language disorder and includes 17 items. In this section, 7 questions target morphology (e.g., *hers vs. theirs; it was raining vs. it’s raining*), 4 items target a wh-question (e.g., “A girl and her brother were eating breakfast. The girl ate some cereal, and the boy ate an apple. Who ate what?”) and 6 target nonword repetitions (e.g., The child is asked to repeat nonsense words that are two to four syllables (e.g., goyfowm)). A diagnostic error score is determined based on responses that are more indicative of a language disorder, errors on non-word repetition items, and the amount of times a participant presented with no response to a question. Responses scored as such are added together to yield a diagnostic risk score. Based on the total diagnostic risk score and the age of the participant, the child’s diagnostic risk status is estimated. The risk score can range from 0 to 14+. The diagnostic risk status for a language disorder results in the following categories: lowest, low to medium, medium to high, and highest risk. It is important to note that the diagnostic error score for the DELV-ST presents cut-offs that vary by age. For age 5, the cut-off is 9 and for age 6 the cut-off is 7. However, only 1 child out of 75 was age 6, and interpretation of his diagnostic error score did not vary based on his age. As a result, the diagnostic error score was set at age 5 and this score was 9. This score was utilized to determine who failed the DELV-ST. In this study, participants presented with a mean DELV-ST diagnostic risk score of 7.35 (SD = 4.03). The scores ranged from 0 to 16.
DIBELS NEXT. Only the Fall DIBELS scores were included within this study. Although, four subtests (First Sound Fluency, Letter Naming Fluency, Phoneme Segmentation Fluency, and Nonsense World Fluency) of the DIBELS Next (Good & Kaminski, 2011) are administered throughout the kindergarten year, only the First Sound Fluency and Letter Naming Fluency subtest are administered in the Fall.

The First Sound Fluency subtest examines children’s phonemic awareness abilities by having them identify the initial sounds in words. During the administration, the examiner verbally presents 30 words to the student, and the student is instructed to say the initial sound of each word. The child is given one minute to complete the thirty items. Two points are given for the correct initial phoneme and one point is provided for correct initial blends and correct initial syllables. Good and Kaminski (2011) have a discontinue rule for this subtest, a total score of 0 is to be recorded for all students who respond incorrectly for the first 5 words in this subtest. This subtest is given in the beginning and middle of the kindergarten year.

The Letter Naming Fluency subtest measures a child’s recognition of letters of the alphabet, it requires students to name as many upper- and lower-case letters arranged in random order as they can. During DIBELS administration, each student is provided randomly arranged upper and lower case letters and given one minute to name as many letters as they can. If a child responds incorrectly to the first row of Letter Naming Fluency items (i.e., the first ten letters), examiners are instructed to stop and give a total score of 0 for this subtest. The child is provided with one point for each correctly named letter. This subtest is administered throughout the three-benchmark periods in kindergarten. The cut-off score for the DIBELS Next was at the 25th percentile rank. In this study, participants presented with a DIBELS Fall mean score of 39.87 (SD = 22.05). The scores ranged from 0 to 81.
PPVT-4. The PPVT-4 is a standardized test that measures receptive vocabulary in both children and adults from age 2 years 6 months through 90 years and older. For each item, children are shown four pictures and asked to identify the picture that depicted the word presented by the examiner. The items on the test increase in difficulty from “cat” to “bouquet.” The test contains 192 items, which are broken into 16 subtests that each contains 12 items. Testing is terminated when a child misses 8 or more items in a subtest. The test has a normative mean of 100 and a standard deviation of 15. The PPVT-4 has a cut-off score that is a standard score of 85. In this study, participants presented with a PPVT-4 mean score of 99.35 (SD = 14.88). The scores ranged from 66 to 128.

Tools Used to Evaluate Predictive Validity of Teacher Ratings

The Syntax subtest of the Diagnostic Evaluation of Language Variation-Norm Referenced (DELV-NR) was used to evaluate predictive validity. The DELV-NR Syntax subtest examines three domains: Wh-questions, comprehension of passive sentences, and article usage. Ten items address the child’s comprehension of wh-questions. The child is shown a set of pictures and asked to listen to a short story based on pictures. Immediately after hearing the story, the child is asked a wh-question about the content. For example, “How does the boy eat what?” The examiner is required to read the question exactly as it is written and to include emphasis on the italicized wh-words. To assess the comprehension of passive sentences, a child is shown three pictures and asked to identify the picture that corresponds to the passive sentence that the examiner reads. This subtest includes 10 passive sentences such as “The ball was rolling by the boy.” The article usage subtest measures a child’s ability to produce appropriate articles given a specific context. The examiner reads eight prompts that require a child to respond using either “the” or “a/an” such as “a car.” Scores from all three domains are combined to form a
standard syntax score. A standard score of 10 (SD = 3) is considered to be typically developing for a child. The DELV-NR Syntax has a standard score cut-off that is 7. In this study, participants presented with a DELV-NR Syntax mean score of 8.22 (SD = 2.87). The scores ranged from 2 to 15.

**Procedures**

A consent form was distributed to the parents through the children’s school. All participants whose caregivers provided consent were assigned an identification number and alpha code in order to maintain confidentiality. Testing was completed in three to four 30-minute sessions. The order of testing for most children was: the DELV-ST, PPVT-4, GFTA-2, PTONI, and the DELV-NR. Data were collected in a quiet room in the child’s school. The examiners were master’s level graduate students, Ph.D. graduate students in psychology, linguistics, or communication sciences and disorders, or a professor in communication sciences and disorders. At the end of the school year, information regarding the participants’ DIBELS NEXT scores were collected from the schools. Teachers were asked to complete the CCC-2 and TROLL rating scales in the spring semester or in the summer immediately after the school year ended.

**Reliability**

Reliability for the DELV-ST, PPVT-4, GFTA-4, PTONI, and DELV-NR was deemed reliable given that they are commercially available standardized tests that were deemed reliable as part of the standardization process. For the teacher ratings, twenty percent of the children’s data was randomly selected and scored independently by a second researcher to assess reliability of the scoring and data entry. There were 840 (21 children x 40 scores) opportunities for agreement. The rate of agreement was 97% (814/840).
CHAPTER FOUR: RESULTS

Convergent Validity of Teacher Ratings

Recall that convergent validity is established if two similar constructs correspond with one another. As such, analysis of convergent validity involved the evaluation of the teacher rating tools against each other and the DELV-ST, DIBELS, and the PPVT-4. Convergent validity was examined in three ways: (1) ANOVAs were done to examine if each screening tool varied by the children’s dialect, school, gender, and clinical status in the same way, (2) correlations were completed to examine how well the teacher ratings related to each other and the other screening tools, and (3) pass and fail rates were examined to see if there was consistency in clinical outcomes across the two teacher ratings and the other screeners. For these analyses, scores from each tool were examined to see if they met the assumptions for parametric statistics. Both the TROLL and PPVT-4 violated the homogeneity of variance assumption. As such, the data for these measures were transformed using a square root transformation (McDonald, 2014). Also, for all of the analyses of the children who passed or failed the tools, results were based on data from 75 children because two did not have a CCC-2 rating form.

The CCC-2 and TROLL. First, I examined whether the teacher ratings from the CCC-2 and TROLL varied by the children’s dialect, school setting, gender, and clinical profile. The first teacher rating was the CCC-2, and the dependent measure for this tool was the general communication composite score. The second teacher rating was the TROLL, and the dependent measure from this tool was the TROLL total score. Total scores for both teacher ratings are reported in Table 6.
Table 6. Test Scores for Teacher Ratings: Means, (SD), and Range

<table>
<thead>
<tr>
<th></th>
<th>CCC-2</th>
<th>TROLL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>102.74</td>
<td>87.43</td>
</tr>
<tr>
<td></td>
<td>(17.17)</td>
<td>(12.26)</td>
</tr>
<tr>
<td>52-139</td>
<td>41-98</td>
<td></td>
</tr>
</tbody>
</table>

The first analysis for the CCC-2 involved a 2 x 4 x 2 x 2 ANOVA, and all four independent variables (i.e. dialect, school, gender, clinical status) were between subject variables. The analysis revealed no main effect for dialect, $F(1, 55) = .346, p = .56, \eta^2 = .006$, school, $F(3, 55) = 1.28, p = .29, \eta^2 = .065$, or gender, $F(1, 55) = 3.84, p = .06, \eta^2 = .065$. A main effect was revealed for clinical status, $F(1, 55) = 23.99, p < .001, \eta^2 = .304$. The group classified as SLI ($M = 89.69; SD = 15.00$) scored lower than the TD group ($M = 108.61; SD = 14.36$).

The effect of the children’s dialect, school, gender, and clinical status on the children’s TROLL scores was also examined utilizing a 2 x 4 x 2 x 2 ANOVA. Results revealed a main effect for school, $F(3, 57) = 2.78, p < .05, \eta^2 = .128$, and clinical status, $F(1, 57) = 14.21, p < .001, \eta^2 = .200$. No main effect or interactions were revealed for dialect, $F(1, 57) = 3.19, p = .08, \eta^2 = .053$, or gender, $F(1, 57) = 1.13, p = .29, \eta^2 = .019$. The SLI group ($M = 77.96; SD = 12.84$) scored lower than the TD group ($M = 91.67; SD = 9.18$), and the main effect for school was related to the children who attended School C ($M = 78.00; SD = 20.34$) earning lower teacher ratings than the children who attended School A ($M = 92.82; SD = 6.56$). School C presented with a higher percentage of questionnaires from children with SLI (40%) than School A; however so did School D, which had 47% of the questionnaires from children classified as having SLI. Given this, the school effects were not tied to the number of children with SLI.
enrolled at School C. See Table 7 for results by school on the TROLL and the percentage of children per school classified as SLI.

Table 7. TROLL Scores by School: Means (SD) and Percentage diagnosed with SLI

<table>
<thead>
<tr>
<th>School</th>
<th>TROLL Score</th>
<th>% Questionnaires from Children with SLI</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>92.56 (6.68)</td>
<td>12% 2/17 children</td>
</tr>
<tr>
<td>B</td>
<td>87.06 (11.72)</td>
<td>33% 12/36 children</td>
</tr>
<tr>
<td>C</td>
<td>78.00 (20.34)</td>
<td>40% 4/10 children</td>
</tr>
<tr>
<td>D</td>
<td>87.12 (9.14)</td>
<td>47% 8/17 children</td>
</tr>
</tbody>
</table>

To compare the teacher rating tools, Table 8 shows the results of the 2 x 4 x 2 x 2 ANOVAs for both the CCC-2 and TROLL. As is evident, both the CCC-2 and TROLL showed a significant difference for clinical status. In addition, the scores for the TROLL showed a significant difference for the children’s school.

Table 8. Results of 2 x 4 x 2 x 2 ANOVA for the Teacher Ratings

<table>
<thead>
<tr>
<th>Test</th>
<th>Dialect</th>
<th>School</th>
<th>Gender</th>
<th>Clinical Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCC-2</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>p &lt; .001</td>
</tr>
<tr>
<td>TROLL</td>
<td>--</td>
<td>p = .049</td>
<td>--</td>
<td>p &lt; .001</td>
</tr>
</tbody>
</table>

Next, relationships between the children’s scores on the CCC-2 and TROLL were examined. See Table 9 for correlations between the teacher rating tools. The overall composite score for the CCC-2 was highly correlated to the overall composite score for the TROLL, r = .70, p < .001. Very high correlations ranging from .90 to .96 were also seen between the overall CCC-2 composite score and the CCC-2 Semantic and Coherence subtests, and between the TROLL overall score and the TROLL subtests for Language, Reading, and Writing. Because the
items used to calculate the subtest scores for the CCC-2 and TROLL were also used to calculate the overall scores for both teacher ratings, it was expected that high correlations would exist between the tools’ composite scores and their subtests. In fact, both the CCC-2 and TROLL presented with higher correlations to their own subtests in comparison to correlations with the subtests from the other teacher rating tool. Nevertheless, moderate correlations were found between the overall composite score for the CCC-2 and both the Reading subtest, $r = .64$, and Writing subtest $r = .61$ of the TROLL. The highest correlations for the TROLL composite score and the CCC-2 subtests were seen with Coherence, $r = .70$, and Semantics, $r = .69$.

Table 9. Correlations between Teacher Rating Scales

<table>
<thead>
<tr>
<th></th>
<th>CCC-2 Composite</th>
<th>TROLL Composite</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCC-2 Speech</td>
<td>.84**</td>
<td>.58**</td>
</tr>
<tr>
<td>CCC-2 Syntax</td>
<td>.81**</td>
<td>.61**</td>
</tr>
<tr>
<td>CCC-2 Semantic</td>
<td>.92**</td>
<td>.69**</td>
</tr>
<tr>
<td>CCC-2 Coherence</td>
<td>.90**</td>
<td>.70**</td>
</tr>
<tr>
<td>CCC-2 Initiation</td>
<td>.68**</td>
<td>.37*</td>
</tr>
<tr>
<td>CCC-2 Scripted Language</td>
<td>.83**</td>
<td>.56**</td>
</tr>
<tr>
<td>CCC-2 Nonverbal Comm.</td>
<td>.85**</td>
<td>.65**</td>
</tr>
<tr>
<td>TROLL Language</td>
<td>.70**</td>
<td>.90**</td>
</tr>
<tr>
<td>TROLL Reading</td>
<td>.64**</td>
<td>.96**</td>
</tr>
<tr>
<td>TROLL Writing</td>
<td>.61**</td>
<td>.92**</td>
</tr>
</tbody>
</table>

Note. *p < .05 **p < .001

Lastly, the overall pass rates were very high for both the CCC-2 (87%) and TROLL (97%), and as a consequence the fail rates for the CCC-2 and TROLL were very low. The CCC-2 overall fail rate was 13%, with 10 of 75 children failing. The TROLL presented with a 5% fail
rate, with 4 of 75 children failing. In addition, 64 (85%) children passed both the CCC-2 and TROLL, and 3 (4%) children failed both the CCC-2 and TROLL. Of the 75 children with complete questionnaire data, only 8 (11%) were not classified by both tools as either passing or failing. The eight included 7 participants who failed the CCC-2 and not the TROLL and 1 who failed the TROLL but not the CCC-2.

In summary, both the CCC-2 and TROLL were able to differentiate the children based on clinical status, but the CCC-2 was less sensitive than the TROLL to other factors such as the children’s dialect spoken and school. Both the TROLL and CCC-2 were highly correlated ($r = .70$) with each other. As expected, the correlations between each test’s subtests were highest with their own composite but all subtests were correlated at moderate to high levels to each other. Lastly, the pass and fail outcomes for both teacher ratings were consistent with each other, with only 11% of children failing one tool but not the other.

The DELV-ST, DIBELS, and PPVT-4. Initially, I looked at the total scores for the DELV-ST, DIBELS, and the PPVT-4. See Table 10 for the total scores and means for each of the convergent validity tools.

<p>| Table 10. Total Group Scores for Convergent Validity Tools: Means, (SD), and Range |
|---------------------------------|----------------|----------------|</p>
<table>
<thead>
<tr>
<th>DELV-ST Risk Score</th>
<th>DIBELS</th>
<th>PPVT-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.35 (4.03)</td>
<td>39.81 (22.24)</td>
<td>99.9 (14.92)</td>
</tr>
<tr>
<td>0-16</td>
<td>0-81</td>
<td>66-128</td>
</tr>
</tbody>
</table>

A 2 x 4 x 2 x 2 ANOVA was conducted to examine the effects of dialect type, school, gender, and clinical status on the children’s DELV-ST risk error score. A significant main effect was revealed for clinical status, $F (1, 57) = 13.20, p < .05, \eta^2 = .032$. The SLI group ($M = 9.69; SD = 3.50$) earned higher risk scores than the TD group ($M = 6.16; SD = 3.78$). No main effect
or interaction was revealed for dialect, \( F (1, 57) = 2.17, p = .15, \eta^2 = .037 \), gender, \( F (1, 57) = 1.86, p = .18, \eta^2 = .032 \), or school, \( F (3, 57) = .91, p = .44, \eta^2 = .046 \).

A 2 x 4 x 2 x 2 ANOVA was also conducted to examine the effects of the children’s dialect type, school, gender, and clinical status on their DIBELS scores. The analysis revealed no significant effect for gender, \( F (1, 57) = .211, p = .65, \eta^2 = .004 \), but significant effects for dialect type, \( F (1, 57) = 4.49, p < .05, \eta^2 = .073 \), school, \( F (3, 57) = 4.32, p < .05, \eta^2 = .185 \), and clinical status, \( F (1, 57) = 8.17, p < .05, \eta^2 = .125 \). The AAE group (\( M = 38.00; SD = 22.87 \)) scored lower than the SWE group (\( M = 41.43; SD = 21.50 \)). School A (\( M = 35.25 \)) earned lower scores than both Schools B (\( M = 42.91 \)) and D (\( M = 42.35 \)). Recall that School A included 12% of children with SLI, School B reported 33%, and School D reported 47% of children classified with SLI. Given this, results were not related to greater numbers of children with SLI enrolled in School A. See Table 11 for DIBELS scores by school. Finally, the SLI group (\( M = 28.19; SD = 21.36 \)) scored lower than the TD group (\( M = 45.82; SD = 20.10 \)).

<table>
<thead>
<tr>
<th>School</th>
<th>DIBELS</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>35.25</td>
</tr>
<tr>
<td></td>
<td>(22.29)</td>
</tr>
<tr>
<td>B</td>
<td>42.91</td>
</tr>
<tr>
<td></td>
<td>(20.04)</td>
</tr>
<tr>
<td>C</td>
<td>32.70</td>
</tr>
<tr>
<td></td>
<td>(24.64)</td>
</tr>
<tr>
<td>D</td>
<td>42.35</td>
</tr>
<tr>
<td></td>
<td>(24.38)</td>
</tr>
</tbody>
</table>

The last convergent validity tool was the PPVT-4. A 2 x 4 x 2 x 2 ANOVA was conducted to examine the effects of the children’s dialect, school, and gender on the children’s PPVT-4 scores. No significant difference was revealed for dialect, \( F (1, 57) = 2.98, p = .09, \eta^2 = .050 \), school, \( F (3, 57) = .257, p = .86, \eta^2 = .013 \), or gender, \( F (1, 57) = .822, p = .37, \eta^2 = .014 \).
The analysis revealed a main effect for clinical status, $F(1, 57) = 106.07, p < .001, \eta^2 = .650$. The SLI group ($M = 81.88; SD = 8.92$) earned lower scores than the TD group ($M = 108.25; SD = 7.57$).

For comparison purposes, Table 12 shows the results of the ANOVAs for the two teacher rating scales and the DELV-ST Risk, DIBELS, and PPVT-4. As is evident, all five screeners showed a significant difference for clinical status. Ideally, a clinician or researcher would choose a tool that shows effects for clinical status without effects for other types of child socio-demographic variables. Both the TROLL and DIBELS scores varied based on the children’s school. The DIBELS scores varied by the most variables, and these included the children’s clinical status, dialect, and school. The CCC-2, DELV-ST, and PPVT-4 showed differences for clinical status only. These results suggest that although both teacher rating tools showed differences for clinical status like the other screeners, the CCC-2 showed fewer differences in the child’s scores as a function of other variables as did the DELV-ST and PPVT-4. As such, I conclude from the ANOVA results that evidence for convergent validity of the teacher ratings was stronger for the CCC-2 than for the TROLL.

Table 12. Results of $2 \times 4 \times 2 \times 2$ ANOVA for Convergent Validity Tools

<table>
<thead>
<tr>
<th>Test</th>
<th>Dialect</th>
<th>School</th>
<th>Gender</th>
<th>Clinical Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCC-2</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>$p &lt; .001$</td>
</tr>
<tr>
<td>TROLL</td>
<td>--</td>
<td>$p = .049$</td>
<td>--</td>
<td>$p &lt; .001$</td>
</tr>
<tr>
<td>DELV-ST Risk</td>
<td>--</td>
<td></td>
<td>--</td>
<td>$p = .001$</td>
</tr>
<tr>
<td>DIBELS</td>
<td>$p = .038$</td>
<td>$p = .008$</td>
<td>--</td>
<td>$p = .006$</td>
</tr>
<tr>
<td>PPVT-4</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>$p &lt; .001$</td>
</tr>
</tbody>
</table>
Next, I looked at how well the teacher ratings correlated with the DELV-ST, DIBELS, and the PPVT-4. See Table 13 for results of the correlational analysis. The relationship between the overall composite score for the CCC-2 and the TROLL and the other screening tools showed both negative and positive correlations. Note that negative correlations were found between the DELV-ST risk score and the other subtests because the higher the risk score the lower the child’s language abilities. A low correlation was found between the CCC-2 composite and both the DELV-ST Risk score, $r = -.25, p < .05$, and DIBELS score, $r = .39, p < .05$, and a moderate correlation was found between the CCC-2 and the PPVT-4, $r = .55, p < .001$. A moderate correlation was also found between the TROLL overall score and the DELV-ST risk score, $r = -.40, p < .001$, DIBELS score $r = .60, p < .001$, and PPVT-4, $r = .51, p < .001$.

Of the seven subtests of the CCC-2, the Speech ($r = -.34$) and Syntax ($r = -.32$) subtests yielded the highest correlations with the DELV-ST. Of the three subtests of the TROLL, the Reading subtest ($r = .58$) had the highest correlation with the DIBELS. The CCC-2 Syntax subtest and the Language subtest of the TROLL yielded the highest overall correlations with the PPVT-4 ($r = .64$ and $60$), respectively.

In summary, although both teacher questionnaires correlated with the other screeners, the correlations for the TROLL were higher than those of the CCC-2. The TROLL presented with more moderate correlations with the other screeners in comparison to the CCC-2. These results suggest that the TROLL shows stronger evidence for convergent validity than the CCC-2 when convergent validity is examined with correlational analysis.
Table 13. Correlations: Convergent Validity

<table>
<thead>
<tr>
<th>Teacher Rating Scales</th>
<th>DELV-ST</th>
<th>DIBELS</th>
<th>PPVT-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCC-2 Composite</td>
<td>-.25*</td>
<td>.39*</td>
<td>.55**</td>
</tr>
<tr>
<td>CCC-2 Subtest Speech</td>
<td>-.34*</td>
<td>.28*</td>
<td>.56**</td>
</tr>
<tr>
<td>CCC-2 Subtest Syntax</td>
<td>-.32*</td>
<td>.32*</td>
<td>.64**</td>
</tr>
<tr>
<td>CCC-2 Subtest Semantics</td>
<td>-.23*</td>
<td>.39*</td>
<td>.52**</td>
</tr>
<tr>
<td>CCC-2 Subtest Coherence</td>
<td>-.28*</td>
<td>.43**</td>
<td>.48**</td>
</tr>
<tr>
<td>CCC-2 Subtest Scripted Language</td>
<td>-.18</td>
<td>.26*</td>
<td>.34*</td>
</tr>
<tr>
<td>CCC-2 Subtest Context</td>
<td>-.21</td>
<td>.42**</td>
<td>.50**</td>
</tr>
<tr>
<td>CCC-2 Subtest Nonverbal Communication</td>
<td>-.23*</td>
<td>.32*</td>
<td>.48**</td>
</tr>
<tr>
<td>TROLL Composite</td>
<td>-.40**</td>
<td>.60**</td>
<td>.51**</td>
</tr>
<tr>
<td>TROLL Language Subtest</td>
<td>-.34*</td>
<td>.54**</td>
<td>.60**</td>
</tr>
<tr>
<td>TROLL Reading Subtest</td>
<td>-.35*</td>
<td>.58**</td>
<td>.42**</td>
</tr>
<tr>
<td>TROLL Writing Subtest</td>
<td>-.45**</td>
<td>.54**</td>
<td>.40**</td>
</tr>
</tbody>
</table>

Note. * p < .05. **p < .001
Next, I examined the percentage of children who passed and failed the CCC-2 and TROLL relative to those who passed and failed the DELV-ST, DIBELS, and the PPVT-4. Table 14 illustrates the breakdown of children who passed one or more of the screening tools. As reported earlier, sixty-four (85%) children passed both the CCC-2 and TROLL, but this number ranged from 57% to 77% when the children were required to pass at least one other screener.

When the CCC-2 and TROLL were examined separately, a slightly higher passing rate was obtained for tools combined with the TROLL in comparison to tools combined with the CCC-2. The two tools with the lowest passing rate (57%) were the CCC-2 and DELV-ST, and the highest passing rate (84%) was the TROLL combined with the DIBELS. Finally, 36 (48%) of the 75 children passed all five screening tools.

Table 14. Pass Rates Across Convergent Validity Tools

<table>
<thead>
<tr>
<th>Tool(s)</th>
<th>Number who Passed</th>
<th>Pass Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCC-2</td>
<td>65</td>
<td>87%</td>
</tr>
<tr>
<td>TROLL</td>
<td>71</td>
<td>95%</td>
</tr>
<tr>
<td>CCC-2 &amp; TROLL</td>
<td>64</td>
<td>85%</td>
</tr>
<tr>
<td>CCC-2, TROLL, &amp; DELV-ST</td>
<td>43</td>
<td>57%</td>
</tr>
<tr>
<td>CCC-2, TROLL, &amp; DIBELS</td>
<td>58</td>
<td>77%</td>
</tr>
<tr>
<td>CCC-2, TROLL, &amp; PPVT-4</td>
<td>54</td>
<td>72%</td>
</tr>
<tr>
<td>CCC-2 &amp; DELV-ST</td>
<td>43</td>
<td>57%</td>
</tr>
<tr>
<td>CCC-2 &amp; DIBELS</td>
<td>61</td>
<td>81%</td>
</tr>
<tr>
<td>CCC-2 &amp; PPVT-4</td>
<td>55</td>
<td>73%</td>
</tr>
<tr>
<td>TROLL &amp; DELV-ST</td>
<td>46</td>
<td>61%</td>
</tr>
<tr>
<td>TROLL &amp; DIBELS</td>
<td>63</td>
<td>84%</td>
</tr>
<tr>
<td>TROLL &amp; PPVT-4</td>
<td>59</td>
<td>79%</td>
</tr>
<tr>
<td>Passed all 5 Screeners</td>
<td>36</td>
<td>48%</td>
</tr>
</tbody>
</table>

Before interpreting the pass rate data, it is important to consider the fail rates. Table 15 illustrates the breakdown of children who failed one or more of the screening tools. Recall that the fail rates for the teacher rating tools were relatively low for the CCC-2 (13%) and the TROLL (5%). These fail rates were relatively low in comparison to the fail rates of the DELV-ST and PPVT-4. Also note that the Table shows the cumulative number of participants who
failed the tools. Note that the DELV-ST had 39 (52%) of 75 children failing. A dramatic drop in fail rate occurred when a child was required to fail one or more tools, from 52% to a range of 3% to 20%. The CCC-2 when combined with the DELV-ST led to the highest fail rate with 6 (8%) children failing both of these tools.

Table 15. Fail Rates Across Convergent Validity Tools

<table>
<thead>
<tr>
<th>Tool(s)</th>
<th>Number who Failed</th>
<th>Fail Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCC-2</td>
<td>10</td>
<td>13%</td>
</tr>
<tr>
<td>TROLL</td>
<td>4</td>
<td>5%</td>
</tr>
<tr>
<td>DELV-ST</td>
<td>28</td>
<td>37%</td>
</tr>
<tr>
<td>DIBELS</td>
<td>10</td>
<td>13%</td>
</tr>
<tr>
<td>PPVT-4</td>
<td>15</td>
<td>20%</td>
</tr>
</tbody>
</table>

  | Failed 1 screener | 39 | 52% |
  | Failed 2 screeners | 15 | 20% |
  | Failed 3 Screeners | 8  | 11% |
  | Failed 4 screeners | 4  | 5%  |
  | Failed all 5 Screeners | 2 | 3%  |

<table>
<thead>
<tr>
<th>Tool(s)</th>
<th>Number who Failed</th>
<th>Fail Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCC-2 &amp; DELV-ST</td>
<td>6</td>
<td>8%</td>
</tr>
<tr>
<td>CCC-2 &amp; DIBELS</td>
<td>4</td>
<td>5%</td>
</tr>
<tr>
<td>CCC-2 &amp; PPVT-4</td>
<td>5</td>
<td>7%</td>
</tr>
<tr>
<td>TROLL &amp; DELV-ST</td>
<td>3</td>
<td>4%</td>
</tr>
<tr>
<td>TROLL &amp; DIBELS</td>
<td>3</td>
<td>4%</td>
</tr>
<tr>
<td>TROLL &amp; PPVT-4</td>
<td>3</td>
<td>4%</td>
</tr>
</tbody>
</table>

In summary, 85% of the children passed both the CCC-2 and TROLL, but this number decreased significantly when the children were required to pass at least one additional screener, and it dropped to 48% when the children were required to pass all five screening tools. The teacher rating tools presented with low overall fail rates (i.e., between 5 -13%), especially when compared to the DELV-ST and the PPVT-4. Overall, the CCC-2 appeared to have better convergent validity in comparison to the TROLL when examining the children’s pass/fail rates. This conclusion is based on the CCC-2 pass/fail rates being slightly more consistent with the other screeners than the TROLL.
Predictive Validity of Teacher Ratings: The DELV-NR

The predictive validity of the teacher ratings were examined with the DELV-NR. Recall that the overall mean score for the DELV-NR was 8.22 (SD = 2.87). The range of scores was from 2 to 15. As before, an ANOVA was first completed to examine the effect of dialect type, school, gender, and clinical status on the DELV-NR. The 2 x 4 x 2 x 2 ANOVA revealed no significant difference for dialect, $F(1, 57) = .191$, $p = .66$, $\eta^2 = .003$, school, $F(3, 57) = .180$, $p = .91$, $\eta^2 = .009$, or gender, $F(1, 57) = .134$, $p = .72$, $\eta^2 = .002$. The analysis revealed a main effect for clinical status, $F(1, 57) = 99.67$, $p < .001$, $\eta^2 = .636$. The SLI group (M = 4.92; SD = 1.55) earned lower scores than the TD group (M = 9.90; SD = 1.68).

See Table 16 for results of the ANOVA in comparison to the two teacher ratings. As is evident, the CCC-2 was similar to the DELV-NR because it showed effects for clinical status without differences for the children’s dialect type, school, or gender. Given this, the CCC-2 showed stronger predictive validity than the TROLL as measured by the ANOVA results.

<table>
<thead>
<tr>
<th>Test</th>
<th>Dialect</th>
<th>School</th>
<th>Gender</th>
<th>Clinical Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCC-2</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>$p &lt; .001$</td>
</tr>
<tr>
<td>TROLL</td>
<td>--</td>
<td>$p = .044$</td>
<td>--</td>
<td>$p &lt; .001$</td>
</tr>
<tr>
<td>DELV-NR</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>$p &lt; .001$</td>
</tr>
</tbody>
</table>

Next, I looked at how well the two teacher rating scales correlated with the children’s scores on the DELV-NR. See Table 17 for correlations between the teacher ratings and the DELV-NR. The CCC-2 composite score ($r = .53$) and TROLL overall score ($r = .52$) were moderately correlated with the DELV-NR. The three teacher rating subtests that presented with
the highest correlations with the DELV-NR were the CCC-2 Speech subtest, CCC-2 Syntax subtest, and the TROLL Language subtest (all $r = .58$).

Table 17. Correlations with Teacher Rating Subtests: Predictive Validity

<table>
<thead>
<tr>
<th></th>
<th>DELV-NR</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCC-2 Composite</td>
<td>.53**</td>
</tr>
<tr>
<td>CCC-2 Speech</td>
<td>.58**</td>
</tr>
<tr>
<td>CCC-2 Syntax</td>
<td>.58**</td>
</tr>
<tr>
<td>CCC-2 Semantic</td>
<td>.52**</td>
</tr>
<tr>
<td>CCC-2 Coherence</td>
<td>.47**</td>
</tr>
<tr>
<td>CCC-2 Scripted Language</td>
<td>.33*</td>
</tr>
<tr>
<td>CCC-2 Context</td>
<td>.49**</td>
</tr>
<tr>
<td>CCC-2 Nonverbal Comm.</td>
<td>.39*</td>
</tr>
<tr>
<td>TROLL Composite</td>
<td>.52**</td>
</tr>
<tr>
<td>TROLL Language</td>
<td>.58**</td>
</tr>
<tr>
<td>TROLL Reading</td>
<td>.44**</td>
</tr>
<tr>
<td>TROLL Writing</td>
<td>.43**</td>
</tr>
</tbody>
</table>

Note. *p < .05. **p < .001.

Recall that the DELV-NR was utilized to help classify the clinical status of participants. As such, 100% of the children with SLI failed the DELV-NR and 100% of the TD children passed it. Since 34% of the children in the study were classified as SLI, 34% of the sample failed the DELV-NR. Table 18 shows the pass rates for the two teacher rating scales in comparison to the DELV-NR. As can be seen, all the teacher rating tools yielded a higher pass rate in comparison to the pass rate of the DELV-NR.
Table 18. Pass Rates: Teacher Ratings in Comparison to the DELV-NR

<table>
<thead>
<tr>
<th>Tool</th>
<th>Number who Passed</th>
<th>Pass Rate Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCC-2</td>
<td>65</td>
<td>87%</td>
</tr>
<tr>
<td>TROLL</td>
<td>71</td>
<td>95%</td>
</tr>
<tr>
<td>DELV-NR</td>
<td>49</td>
<td>65%</td>
</tr>
</tbody>
</table>

As shown in Table 19, very few of the children who failed the DELV-NR also failed the CCC-2 and the TROLL. Eight (11%) failed both the CCC-2 and DELV-NR whereas only three (4%) failed the TROLL and DELV-NR. Interestingly, the same three children who failed the TROLL and DELV-NR also failed the CCC-2. Given this, only three children failed both teacher ratings and the DELV-NR.

Table 19. Fail Rates Across Predictive Validity Tools

<table>
<thead>
<tr>
<th>Tools</th>
<th>Number who Failed</th>
<th>Fail Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCC-2</td>
<td>10</td>
<td>13%</td>
</tr>
<tr>
<td>TROLL</td>
<td>4</td>
<td>5%</td>
</tr>
<tr>
<td>DELV-NR</td>
<td>26</td>
<td>35%</td>
</tr>
<tr>
<td>CCC-2 &amp; DELV-NR</td>
<td>8</td>
<td>11%</td>
</tr>
<tr>
<td>TROLL &amp; DELV-NR</td>
<td>3</td>
<td>4%</td>
</tr>
<tr>
<td>CCC-2, TROLL, &amp; DELV-NR</td>
<td>3</td>
<td>4%</td>
</tr>
</tbody>
</table>

Using the pass/fail data just presented, diagnostic sensitivity and specificity values were calculated for each teacher rating scale. Recall that sensitivity is the percentage of children with SLI who were identified as SLI, and specificity is the percentage of children with typical development who were identified as typical. Both tools show less than ideal accuracy rates, with extremely low levels of sensitivity. The overall diagnostic accuracy of the CCC-2 to identify
individuals correctly was 74%. Sensitivity and specificity for this tool was .31 (8/26) and .96 (47/49), respectively. The overall accuracy of the TROLL to identify individuals correctly was 70%. Sensitivity and specificity for the TROLL was (3/26) .12 and (48/49) .98, respectively.

In summary, I found mixed results depending on the analyses conducted. The CCC-2 and the DELV-NR were able to differentiate the clinical status of the participants with no effects for the children’s dialect status, gender, and school. Both teacher ratings also presented with moderate correlations with the DELV-NR. However, overall fail rates were lower for the teacher ratings in comparison to the DELV-NR, with only 8 (11%) children failing the CCC-2 and DELV-NR and only 3 (4%) failing the TROLL and the DELV-NR. In other words, both teacher rating tools showed strong predictive validity when analyses focused on ANOVAs and correlations, but neither showed even a fair degree of diagnostic accuracy when pass/fail accuracy rates (i.e., sensitivity and specificity values) were considered.

**Combining Screeners to Improve Diagnostic Sensitivity and Specificity**

The clinical cut-offs for the teacher questionnaires as well as the means and standard deviations for the SLI and TD clinical groups are re-reported in Table 20 along with similar data for the other three screeners that were used to examine convergent validity. The second column shows the clinical cut-off scores based on each of the screener’s manuals. These are the same cut-off scores that were used in the previous analyses. The next two columns in the Table report the sensitivity and specificity for each of the tools based on those clinical cut-offs. The final two columns show the mean scores of the SLI and TD groups.
Table 20. Sensitivity and Specificity for Screeners Relative to DELV-NR: Means and (SD)

<table>
<thead>
<tr>
<th>Clinical Cut-offs</th>
<th>Se.</th>
<th>Sp.</th>
<th>TD Group</th>
<th>SLI Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCC-2 85 Composite Score</td>
<td>.31</td>
<td>.96</td>
<td>108.61 (14.36)</td>
<td>89.69 (15.00)</td>
</tr>
<tr>
<td>TROLL 65 total score</td>
<td>.12</td>
<td>.98</td>
<td>91.67 (9.18)</td>
<td>77.96 (12.84)</td>
</tr>
<tr>
<td>DELV-ST 9 Error Score</td>
<td>.62</td>
<td>.76</td>
<td>6.16 (3.78)</td>
<td>9.69 (3.50)</td>
</tr>
<tr>
<td>DIBELS 25th percentile rank</td>
<td>.27</td>
<td>.94</td>
<td>45.82 (20.10)</td>
<td>28.19 (21.36)</td>
</tr>
<tr>
<td>PPVT-4 85 Standard Score</td>
<td>.58</td>
<td>1.0</td>
<td>108.25 (7.57)</td>
<td>81.88 (8.92)</td>
</tr>
</tbody>
</table>

Note. Se. = sensitivity. Sp. = specificity.

As can be seen, the sensitivity was low for both the CCC-2 and TROLL; however it was also low for the other screeners. At the same time, all five screeners led to statistically significant differences between the SLI and TD groups. This finding suggests that it may not be the screening tools that lack predictive validity but the screeners’ cut-off scores that are used to define who passes and who fails. To explore this hypothesis, I examined the data to determine the optimal cut score for each screener. I did this manually by finding the score that corresponded to the fewest number of children who were misclassified on each screener relative to the DELV-NR. See Table 21 for new sensitivity and specificity values utilizing the empirically-derived cut-off scores.

As can be seen, adjusting the cut-off scores improved the sensitivity for all five screeners. Using the empirically-derived cut scores, both the teacher rating tools yielded the highest levels of sensitivity in comparison to the other three screening tools, however, higher rates of sensitivity also led to lower rates of specificity. By comparison, the PPVT-4 yielded the highest specificity (48/49 = .98), and it also yielded a fair level of sensitivity (23/26 = .88).
Table 21. Sensitivity and Specificity for Screeners Using Empirically-Derived Cut-Off Scores

<table>
<thead>
<tr>
<th>Clinical Cut-offs</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Overall Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCC-2 106 Composite Score</td>
<td>.92</td>
<td>.61</td>
<td>72%</td>
</tr>
<tr>
<td>TROLL 90 Total Score</td>
<td>.96</td>
<td>.73</td>
<td>81%</td>
</tr>
<tr>
<td>DELV-ST 8 Error Score</td>
<td>.69</td>
<td>.76</td>
<td>73%</td>
</tr>
<tr>
<td>DIBELS 68th percentile rank</td>
<td>.73</td>
<td>.63</td>
<td>67%</td>
</tr>
<tr>
<td>PPVT-4 89 Standard Score</td>
<td>.88</td>
<td>.98</td>
<td>95%</td>
</tr>
</tbody>
</table>

Next, I ran a discriminant function analysis on all of the screening tools to see if two or more of the tools could be combined to improve the accuracy of the screening process. Before doing this, I had to determine if each tool met the assumptions of a discriminant function analysis. A discriminant function analysis must meet five assumptions: (1) normal distribution of data, (2) homogeneity of variance, (3) absence of multicollinearity, (4) independence of participants, and (5) no outliers. Recall that both the TROLL and PPVT-4 did not meet the assumption of homogeneity of variance, which was important for the earlier ANOVAs. For the discriminant function analysis, the assumption of normality is also important, and the DIBELS along with the TROLL and PPVT-4 did not meet this assumption.

To address the normality of the data, scores for the TROLL, PPVT-4, and DIBELS were examined for outliers. For each tool, each child’s case was tested against the group mean utilizing case-wise statistics. Outliers were identified if the difference between the child’s score and group mean was greater than what one would expect from chance (with the alpha set at .05). Only two participants presented as outliers on the TROLL, and with scores of 41 and 45, respectively. These scores were statistically different from the TROLL’s group mean of 87.43. The score of 41 came from a child classified as SLI (Group mean = 77.96) and the score of 45
was from a child classified as TD (Group mean = 91.67). As a result, these two outliers were removed, and the discriminant function analysis was based on 73 participants. Finally, as before, square root transformations were completed in an attempt to normalize the distribution (McDonald, 2014) of the TROLL, DIBELS, and PPVT-4 data.

The stepwise discriminant function analysis selected the TROLL and PPVT-4 as the best predictor variables, with a diagnostic accuracy level of 95%. Sensitivity and specificity was .88 and .98, respectively. No other variables (i.e., screeners) were selected for the model. The overall Chi-square test of the discriminant function was significant (Wilks $\lambda = .274$, Chi-square = 87.981, df = 2, canonical correlation = .852, $p < .001$); and the two screeners together accounted for 73% of the variance in the children’s DELV-NR scores. Visual inspection of the cut scores for this model indicated that children were classified as SLI if their PPVT-4 score was 89 or below and their TROLL score was below 92. If a child’s scores were above these cut scores, he or she was classified as TD.

It is important to note that the two children excluded from the discriminant function analysis had no impact on the sensitivity and specificity values of the empirically-derived cut scores for PPVT-4. Recall that the PPVT-4 was 95% accurate (Sensitivity = 23/26 = .88; Specificity= 48/49 = .98) with an empirically-derived cut-off of 89 when considering data from 75 participants. When data from 73 participants were included in the discriminant function analysis, the cut score remained 89, and the diagnostic accuracy of the PPVT-4 remained at 95%, yielding sensitivity and specificity values of .88 (22/25) and .98 (47/48), respectively. The only difference between the two models (one with 73 participants and the other with 75) was the amount of variance accounted for by the PPVT-4 (73% vs. 70%). In other words, the same percent of participants were accurately classified by the PPVT-4 by itself with data from 73 or 75
children. The child with SLI who was excluded because of the TROLL earned a PPVT-4 score of 75 and the child classified as TD who was excluded earned a PPVT-4 score of 113. In contrast, excluding the two children who were outliers of the TROLL resulted in an 82% accuracy level for the TROLL, yielding sensitivity and specificity values of .96 (24/25) and .75 (36/48), respectively. However, the TROLL by itself only accounted for 38% of the variance in the children’s DELV-NR scores. Combining both the TROLL and PPVT-4 accounted for 73% of the variance in the model, which led to 3% more of the variance being accounted for than when the PPVT-4 was considered by itself.

In summary, utilizing empirically-derived cut-offs led to increased diagnostic accuracy for all of the screeners. The empirically-derived cut-offs led to the two teacher rating tools having the highest sensitivity of the tools, although it came at a price, with lower levels of specificity. Overall, the PPVT-4 yielded the highest level of diagnostic accuracy (95%) as compared to the other tools, and it by itself accounted for 70% of the variance in the children’s DELV-NR scores. The PPVT-4 and TROLL together also yielded a diagnostic accuracy level of 95%, but was able to account for 73% (as compared to 70%) of the variance in the children’s DELV-NR scores.
CHAPTER FIVE: DISCUSSION

The purpose of this study was to examine the validity of teacher ratings for measuring children’s language skills. Teacher ratings were measured with two tools, the *Children’s Communication Checklist-2* and the *Teacher Rating of Oral Language and Literacy*, and these two tools were examined for their convergent and predictive validity. Convergent validity was examined by comparing the two teacher rating scales to each other and to three other language and literacy screeners. Predictive validity was examined by comparing the two teacher rating scales to the *Diagnostic Evaluation of Language Variation: Norm Referenced*. Finally, empirically-derived cut-offs and empirically-derived combinations of tools were examined to determine if these methods would improve the predictive validity of the screeners.

This chapter is divided into four sections. The first section focuses on the findings as they relate to the research questions. The second section focuses on the findings as they relate to previous studies. The third section considers limitations of the current study and directions for future studies. The fourth section suggests clinical implications of the findings.

**Findings as Related to the Research Questions**

The first three questions focused on convergent validity, and each question required a different type of statistical analysis: ANOVAs, correlations, and an examination of pass/fail rates. The first question was, *Are the children’s scores on the CCC-2 and TROLL affected by their clinical status and by other sociodemographic variables (nonmainstream dialect use, school, gender) in a way that is consistent with each other and with the DELV-ST, DIBELS, and the PPVT-4?* Both teacher rating tools showed effects for clinical status like the other screeners; however, the CCC-2 led to fewer statistical differences in the children’s scores as a result of other child socio-demographic variables as did the DELV-ST and PPVT-4. Given this, the
ANOVA results indicated that the CCC-2 showed stronger evidence of convergent validity than the TROLL.

The second question was, *Are the children’s scores on the CCC-2 and TROLL correlated with each other and to the children’s scores on the DELV-ST, DIBELS, and the PPVT-4?* The CCC-2 and TROLL composites showed a high correlation with each other. The CCC-2 and TROLL were also correlated at low to moderate levels with the other three screeners, and the correlations for the TROLL and other screeners were slightly higher than those of the CCC-2. These results suggest that although both teacher rating tools were correlated with each other and the other screening tools, the TROLL showed stronger evidence for convergent validity than the CCC-2 when convergent validity was examined with correlational analysis.

The third question was, *What percentage of children pass and fail the CCC-2 and TROLL relative to each other and to those who pass and fail the DELV-ST, the DIBELS, and the PPVT-4?* Eighty-five percent of the children passed both for the CCC-2 and TROLL, but this number decreased to 57% - 77% when the children were required to pass at least one additional screener. Moreover, only 48% of the children passed all five screening tools. The fail rates from the current study were 13% for the CCC-2 and 5% for the TROLL. Both of these fail rates were lower than the fail rates of the DELV-ST (37%) and PPVT-4 (20%). Given that the fail rates of the CCC-2 were closer to those of two of the other screeners, the CCC-2 showed stronger evidence for convergent validity than the TROLL when pass and fail rates were considered.

As shown by a review of the findings, for two out of three of the analyses, the CCC-2 presented with better evidence of convergent validity in comparison to the TROLL. However, one could argue that neither the CCC-2 nor TROLL led to pass/fail rates that were highly consistent with the pass/rail rates of the other screeners, and the pass/fail rates of the other
screeners were also not consistent with each other. This is unfortunate and suggests that the number and type of children who fail speech and language screenings varies as a function of the screener utilized.

Questions four through six focused on predictive validity, and these questions also required three different types of statistical analysis: ANOVAs, correlational analysis, and pass/fail rates. The fourth question was, *Are the children’s scores on the CCC-2 and TROLL affected by their clinical status and by other sociodemographic variables (nonmainstream dialect use, school, gender) in a way that is consistent with the effects of these variables on the DELV-NR?* The CCC-2, but not the TROLL, showed effects for clinical status without effects of the children’s dialect type, school, or gender, and these findings were similar to those obtained for the DELV-NR. Based on the result from the ANOVAs, I conclude that the CCC-2 shows stronger predictive validity than the TROLL.

The fifth research question was, *Are the children’s scores on the CCC-2 and TROLL correlated with the children’s scores on the DELV-NR?* Both teacher ratings were moderately correlated with the DELV-NR. Based on these results, I conclude that both the CCC-2 and TROLL showed similar levels of predictive validity.

The sixth research question was, *What percentage of children pass and fail the CCC-2 and TROLL relative to those who pass and fail the DELV-NR?* The pass and fail rates for the current study were 87% and 13% for the CCC-2, 95% and 5% for the TROLL, and 66% and 34% for the DELV-NR. As is evident, the fail rates for both teacher ratings were much lower than the DELV-NR. This finding was surprising because screeners by design should over-identify children rather than under-identify children. In addition, only three (4%) children who failed the DELV-NR also failed both teacher ratings.
In summary, I again have mixed findings for the teacher ratings based on the analyses conducted. Both the CCC-2 and TROLL led to group differences between those with and without SLI but only the CCC-2 was like the DELV-NR in showing no effects for any other child socio-demographic variable. In addition, although both teacher ratings presented with moderate correlations with the DELV-NR, overall fail rates were lower for the teacher ratings in comparison to the DELV-NR. It was both surprising and disappointing that neither teacher rating scale led to pass/fail rates that were similar to those of the DELV-NR. This finding indicates that teacher ratings cannot be used as the only screening tool for determining who should be referred to a SLP for a comprehensive evaluation. In fact, if only the teacher ratings were used for screening, most of the children with speech and language weaknesses would never be evaluated for services. Given this, I must conclude that neither teacher rating tool, when it is administered as intended by the test authors, yielded a level of predictive validity that is clinically useful.

The final research question was, If low predictive validity of the teacher ratings are revealed, does predictive validity increase if empirically derived cut-offs are established and the teacher ratings are combined with the other screening tools? Sensitivity was improved by changing the cut-off scores for all five screeners based on an empirically-derived method. This method involved identifying the optimal cuts that would lead to the fewest number of children who were misclassified on each screener. Based on empirically-derived cut scores, the sensitivity values improved for all five screeners, and both teacher rating tools yielded the highest sensitivity values in comparison to the other screeners. Of all the screeners, however, the PPVT-4 yielded the highest levels of diagnostic accuracy (95%; sensitivity = .88; specificity = .98). The PPVT-4 and TROLL in combination also led to the same overall accuracy level of
95% (sensitivity = .88; specificity = .98), but it accounted for 73% of the variance in the model. The only difference between using the PPVT-4 by itself and combining it with the TROLL related to the variance accounted for within the models. The PPVT-4 by itself accounted for 70% of the variance, whereas the PPVT-4 and TROLL together accounted for 73%.

Given this, I conclude that the PPVT-4 by itself or both the PPVT-4 and TROLL together generate the strongest evidence of predictive validity of the five screeners examined in this dissertation. Both of these options lead to a diagnostic accuracy level of 95% and sensitivity and specificity values of .88 and .98, respectively. According to Plante and Vance (1994), levels of sensitivity and specificity at or above .80 are considered fair and levels at or above .90 are considered good. As such, findings for the discriminant function analysis indicate that when empirically-derived cut points are employed, the PPVT-4 by itself or the PPVT-4 and the TROLL together lead to fair to good levels of predictive validity.

Findings Related to Previous Studies

The literature review focused on eleven studies that evaluated the usefulness of teacher ratings of children’s speech and language abilities. None of the studies reviewed examined two different teacher rating scales or examined both convergent and predictive validity. The current study did this with three different types of analyses (i.e., ANOVAs, correlations, pass/fail rates, and/or sensitivity/specificity). Of the previous studies, three focused on correlational analysis, two focused on pass and fail rates, and two focused on diagnostic accuracy levels using sensitivity and specificity values. Below, I compare my findings to these nine studies when a comparison is possible.

Similar to the current study, Elliot, Lee, and Tollefson (2001) examined teacher ratings in comparison to the DIBELS. The participants were enrolled in kindergarten (63% White and
37% non-White) from four classrooms in three elementary schools. The teachers were administered a pre-reading questionnaire and the scores were compared to the children’s DIBELS scores. The teacher’s pre-reading rating questionnaire and DIBELS scores were correlated at \( r = .67 \). In the current study, the TROLL also correlated to the DIBELS at a similar level \( (r = .60) \), while the correlation between the CCC-2 and DIBELS was lower \( (r = .39) \).

Recall that the TROLL questionnaire included a reading and writing subtest. Correlations for the TROLL subtests with DIBELS in the current study were: the Language Subtest \( (r = .54) \), Reading Subtest \( (r = .58) \), and Writing Subtest \( (r = .54) \). The Reading subtest for TROLL had the highest correlation with the DIBELS as expected since the DIBELS examine the pre-cursors to reading. As such, the findings in the current study are very similar to those reported by Elliot, Lee, and Tollefson (2001).

The TROLL developers reported that total scores were moderately correlated to children’s scores on three measures of language, including the PPVT-3 raw score \( (r = .47) \), the Emergent Literacy Profile total score \( (r = .43) \), and the Early Phonemic Awareness Profile total score \( (r = .45) \). In the current study, the teacher ratings were correlated at similar levels with the PPVT-4 and the DIBELS which is a tool that examines phonemic awareness. The correlations between the TROLL and the PPVT-4 \( (r = .51) \) and the CCC-2 and the PPVT-4 were similar \( (r = .55) \). The correlation between the TROLL and DIBELS were higher \( (r = .60) \) than what was reported by the TROLL’s test developers, while the correlation between the CCC-2 and DIBELS were lower \( (r = .39) \).

Rodriguez and Guiberson (2011) examined children, aged 4 years, who were English-speaking, Spanish-speaking, and English/Spanish bilingual children. All of the children in the study were typically developing. For the English-speaking children, results showed that the
TROLL was significantly correlated with the PLS expressive ($r = .22$) and the receptive ($r = .20$) subtest scores. For the Spanish-speaking children, the TROLL was correlated with only the PLS receptive ($r = .22$) subtest scores. For the bilingual children, the TROLL was not correlated with either the expressive ($r = .07$) or receptive ($r = .05$) language subscale of the PLS-4. For the current study, correlations between the TROLL and the other language measures were higher ($r = -.40$ and .51) than those reported for all three groups by Rodriguez and Guiberson (2011). The TROLL total mean score for TD children in the current study was also higher than those reported for all three groups by Rodriguez and Guiberson. In the current work, the TD group mean was 91.67. Rodriguez and Guiberson’s means were: English-speaking children (54.93), Spanish-speaking children (47.88), and bilingual children (50.19). Finally, Rodriguez and Guiberson (2011) reported a 20% fail rate on the TROLL for the English-speaking group, a 35% fail rate for the Spanish-speaking group, and a 32% fail rate for the bilingual children. By comparison, in the current study the fail rate of the TROLL was only 5%. This 5% fail rate was much lower than the fail rates reported by Rodriguez and Guiberson (2011).

Some of the reasons for differences across studies may be related to the type of children included in the studies. Rodriguez and Guiberson (2011) included children living in urban areas of the Southwestern United States, whereas the current study examined students who lived in rural regions of Louisiana. However, the children’s ages also differed across studies. Recall that the children in the current study were older than the children in the Rodriguez and Guiberson (2011) study, and scores on the TROLL are expected to increase with children’s age. As such, it might be expected that four-year olds would fail the TROLL at higher rates compared to the 5- and 6-year olds studied here.
Interestingly, Washington and Craig (2004) also documented higher fail rates for children in pre-kindergarten in comparison to children in kindergarten. Their study included 196 African-American children. One-hundred and forty children were in preschool, and fifty-six children were in kindergarten. Sixty-four children (33%) failed at least one screener, however most of the children that failed a screener were preschoolers (89%). By comparison, the fail rate for the kindergarten children was only 11%.

Moland (2011) also examined the consistency of pass/fail rates across as screeners as was done here. Recall that her participants spoke AAE and attended Pre-K and Head Start programs. The children were administered the DELV-ST, the Fluharty Preschool Speech and Language Screen Test-Second Edition, and the Washington-Craig Language Screener. These three screeners led to identical clinical outcomes for 54% of the children. For the current study, the screeners led to identical clinical outcomes for 51% of the children. As is evident, the degree of consistency in pass/fail rates across the screeners studied here was very similar to results from Moland’s (2011) study.

Finally, Cabell et al. (2009) and Jessup et al. (2008) each examined diagnostic sensitivity and specificity values of teacher ratings. Cabell et al.’s (2009) teacher scale included an abbreviated, 12-item version of the Pre-Literacy Rating Scale from the Clinical Evaluation of Language Fundamentals Preschool – Second Edition, while Jessup et al. (2009) utilized the Kindergarten Development Check. Both Cabell et al. (2009) and Jessup et al. (2008) included children who were the same age as was studied in the current study. Both studies collected teacher ratings without training the teachers, which was also similar to what was done in the current study. See Table 22 for a comparison of diagnostic sensitivity and specificity values of the current study to those of past studies.
Table 22. Sensitivity and Specificity of Teacher Ratings across Studies

<table>
<thead>
<tr>
<th>Study</th>
<th>Age</th>
<th>Findings for Teacher Ratings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cabell et al. (2009)</td>
<td>3–5 years</td>
<td>Sen. = .52 Spec. = .88</td>
</tr>
<tr>
<td></td>
<td>N = 209</td>
<td></td>
</tr>
<tr>
<td>Jessup et al. (2008)</td>
<td>4–5 years</td>
<td>Sen. = .15 Spec. = .97</td>
</tr>
<tr>
<td></td>
<td>N = 286</td>
<td></td>
</tr>
<tr>
<td>Current Study</td>
<td>5–6 years</td>
<td>CCC-2 Sen. = .31 Spec. = .96</td>
</tr>
<tr>
<td></td>
<td>N = 77</td>
<td>TROLL Sen. = .12 Spec. = .98</td>
</tr>
</tbody>
</table>


As is evident in the table, the sensitivity and specificity values of the current study are similar to those obtained by Cabell et al. (2009) and Jessup et al. (2008). The CCC-2 and TROLL presented with similar diagnostic specificity values as the other studies; however, the sensitivity values were lower.

**Limitations and Directions for Future Studies**

Several limitations of the current study limit the generalizations of the findings. These limitations included the small size of the sample, the narrow age range of the children studied, the low number of schools solicited for participation, and the use of a loose PTONI score for determining who could be included in the study. A sample size greater than 77 children would increase the generalizability of the results to a larger population of AAE and SWE speakers in the South. Despite the small sample size, the current study provided a more detailed set of data analysis than has been done in previous studies.

The narrow age range of five- to six-year old children also limits the generalizability of the findings. Future studies should focus on the usefulness of teacher ratings with different age
groups of children to see if teachers are able to better identify childhood language impairment in young or older children. However, as mentioned in the literature review, kindergarten is an important year for SLPs to conduct screenings because it is often the year a child first attends school.

The children were recruited from only four schools in one rural region in Louisiana. Different results may have been obtained had I examined data from other schools in and outside Louisiana. The TROLL and DIBELS scores were the only two tools out of the five screeners that varied based on the children’s school, and the school effects were not the same for these two tools. For the TROLL, the main effect for school was related to School C earning lower teacher ratings than School A. For the DIBELS, the main effect for school related to School C earning lower ratings than both School B and D. Also, neither set of school effects were tied to the number of children with SLI enrolled at the schools or to a particular teacher or school. Nevertheless, data from teachers who work at different types of schools and in different types of communities and regions of the country are needed to examine the generalization of the current findings.

Another limitation of the study related to the loose criteria that was used for the PTONI. Recall that two children classified as SLI in the study earned PTONI scores that were -2 SD below the normative mean and an additional 14 children earned a PTONI score between – 1 SD and – 2 SD of the normative mean. Ideally, a study focusing on children with SLI would include only those children who earned PTONI scores that were within or above -1 SD above the normative mean. In the current work, a loose criteria was used to increase the number of children who could contribute data to the study. Recall that none of these children had been classified in the schools as presenting a cognitive disability and for the two children who earned the lowest
PTONI scores, their vocabulary scores (as measured by the PPVT-4) were higher than their PTONI scores. Nevertheless, for completeness, I have listed the test scores and teacher ratings for the 16 children who earned low PTONI scores in Appendix B. As can be seen, the teacher ratings of these children varied. Moreover, using the cut-offs recommended by the test developers, only two of these children failed the CCC-2 and the TROLL, and only one of these children failed both of these tools.

Although not mentioned previously, 16 (21%) children reached the ceiling on the TROLL. Ideally, one would want to choose a tool that does not have any children reaching the ceiling on the test. In spite of 21% of the children reaching ceiling, the TROLL presented with an overall accuracy of 81% when an empirically-derived cut-off was employed. Although the CCC-2 was not selected as a predictor tool within the current study, none of the children who were tested reached ceiling. Given this, future studies are needed to examine other teacher rating tools along with other types of screeners and other types of speech and language assessment tools. Regardless of the tools selected, future studies should include various statistical analyses (ANOVA, correlations, discriminant function analysis) to examine convergent and predictive validity. As was shown in the current study, different types of analyses lead to different types of findings, and findings from multiple types of analyses are needed to fully evaluate the clinical usefulness of a tool.

Clinical Implications

Results from the current study indicate that neither the CCC-2 nor TROLL should be used in isolation to determine who should be referred to the SLP for a language evaluation in kindergarten. The results also suggest that the PPVT-4 in isolation or the PPVT-4 and TROLL together can be used for screening purposes as long as empirically-derived cutoffs are used.
When thinking about these two options for the school-based SLP, one might argue that administration of the PPVT-4 is the best option because it involves one tool and does not require input from the teachers. On the other hand, administering the PPVT-4 to all kindergarteners would be very expensive and also extremely time-consuming, especially since any unnecessary time an SLP spends on screening is time taken away from serving children who have already been identified as needing services. One option, then, might be to ask kindergarten teachers to complete the TROLL on all of their children. Using an empirically-derived cut score of 90 on the TROLL, a clinician could expect 96% of the children with language impairments to score lower than 90 and 73% of children without language impairments to score higher. Recall that these percentages reflect the TROLL’s diagnostic accuracy of 81% when the empirically-derived cut score of 90 is used. With this information, the SLP could then administer the PPVT-4 to all of the children who failed the TROLL. This would mean that the SLP would screen 96% of the children with impairments (and miss 4%) and 27% (100 – 73%) of the children without an impairment. At this time, any child who scored below 89 on the PPVT-4 would be classified as failing the screening, and a full speech and language evaluation would be recommended.

Assuming a school enrolls 200 kindergartners, and 7% (n = 14) of the children presents with SLI, this screening option would require teachers to complete 200 TROLLs, and the SLP to administer the PPVT-4 to 63 children (13 with SLI and 50 false positives – children without SLI). With the PPVT-4 yielding an accuracy rate of 95% (sensitivity = .88 and specificity = .98), two of the 13 children with SLI and 49 of the 50 children without SLI who failed the TROLL would likely pass the PPVT-4. This outcome would result in 12 children for whom the SLP would conduct a full speech and language evaluation. Although two children with SLI would be missed by this screening option (one would be missed by the TROLL and another missed by the
PPVT-4) and one child without SLI would receive an unnecessary speech and language
evaluation, 86% of the kindergarteners with SLI enrolled at the school would be identified. This
percentage is notable given the less than 30% of kindergarteners with SLI who were receiving
services by an SLP in Tomblin et al.’s (1997) large epidemiology study.

Conclusions

Identifying children with SLI during the early school years is imperative because
targeting these children for early intervention in the schools may improve long term reading
outcomes and academic achievement. The current study was conducted to examine the validity
of teacher ratings within the screening process. The current study was the first to examine
teacher ratings of children who speak a non-mainstream dialect of American English. This
current study was also the first to examine both the convergent and predictive validity of two
teacher rating tools, the CCC-2 and TROLL. It was also the first to utilize three different types
of statistical analyses (i.e. ANOVAs, correlations, pass/fail rates, discriminant function analysis)
and an empirically-derived cut score approach to examine the validity the teacher rating scales.

Results for the current study provide some support for convergent and predictive validity
of teacher ratings as measured by the CCC-2 and TROLL. Two out of three of the analyses (i.e.
ANOVA, pass/fail rates) indicated that the CCC-2 presented with better convergent validity than
the TROLL. Analysis of predictive validity also indicated that only the CCC-2 was similar to
the DELV-NR in showing no effects for any other child socio-demographic variable other than
clinical status. However, the overall fail rates for the teacher ratings as measured by both the
CCC-2 and TROLL were much lower than those of the DELV-NR. This same finding was also
observed for the three other screeners. Given this, neither teacher rating tool nor the other three
screeners led to pass/fail rates that were highly consistent with the DELV-NR. This finding
indicates that neither teacher rating tool nor the other three screeners, when they are administered as intended by the test authors, yield a level of predictive validity that is clinically useful.

Nevertheless, the teacher ratings were sensitive to the children’s language abilities as both tools and the other three screeners led to statistical differences between the children with SLI and those classified as TD. Given this, empirically-derived cut-offs were explored, and when these were used, the diagnostic accuracy of both teacher ratings and the other screeners improved significantly. Moreover, a discriminate function analysis with the empirically-derived cut-offs resulted in the selection of the PPVT-4 and TROLL as the best tools to utilize in combination for a screening. Together, these two tools yielded a diagnostic accuracy level of 95% and accounted for 73% of the variance in the children’s DELV-NR scores. The PPVT-4 by itself also led to a diagnostic accuracy level of 95%, with 70% of the variance in the children’s DELV-NR scores explained. For screening purposes, these findings support the use of teacher ratings as measured by the TROLL when empirically-derived cut scores are used and when the TROLL is combined with the PPVT-4.
NOTES

1 The significant difference noted in this table was based on an omnibus 2 x 4 x 2 x 2 ANOVA that utilized a modified population mean.
REFERENCES


Eickhoff, J., Betz, S., & Ristow, J. (2010, June). *Clinical procedures used by speech-language pathologists to diagnose SLI.* Poster session presented at the symposium on Research in Child Language Disorders, Madison, WI.


Cambrium Learning Group.


Hernandez, M. (2013). *Comparison of Performance on Static and Dynamic Language Tasks among Typically Developing Preschoolers with Bilingual Experience* (Master’s Thesis. The University of New Mexico). Retrieved from https://repository.unm.edu/bitstream/handle/1928/23396/M.H.Thesis.2.pdf?sequence=1

Literacy Screening: Preschool, Charlottesville, VA: University of Virginia.


### APPENDIX A: SUMMARY OF TEACHER RATING STUDIES

<table>
<thead>
<tr>
<th>Authors</th>
<th>Participants</th>
<th>Tool Utilized</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Botting et al.</td>
<td>242 children</td>
<td>Teacher Judgments</td>
<td>66% agreement level between SLP/teacher judgments in the domains of articulation, phonology, syntax/morphology</td>
</tr>
<tr>
<td>(1997)</td>
<td>Aged 6 – 8 years</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bishop et al.</td>
<td>196 children</td>
<td>CCC</td>
<td>Teacher ratings statistically higher for TD children ($d = .82$)</td>
</tr>
<tr>
<td>(2006)</td>
<td>SLI = 64</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>TD = 132</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Age = 6 years</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| Cabell et al.    | 209 children | 12-item version of the Pre-Literacy Rating Scale from the *Clinical Evaluation of Language Fundamentals Preschool- 2nd Edition* | Overall Accuracy = 79%  
Sen. = .52  
Spec. = .88 |
| (2009)           | Aged 3 -5 years |                                         |                                                                                             |
| Jessup et al.    | 286 children | The *Kindergarten Development Check*    | Overall Accuracy = 71%  
Sen. = .15  
Spec. = .97 |
| (2008)           | Aged 4 -5 years |                                         |                                                                                             |
| Williams         | Amount of    | Teacher Judgments                      | Sen. = .86  
Spec. = .68                                                                                       |
| (2006)           | children = 29 Teachers |                                         |                                                                                             |
|                  | Ages = Unknown |                                         |                                                                                             |
| Gilmore & Vance  | 52 children  | 8 item questionnaire on teacher’s perception of the children’s attentive listening & verbal comprehension skills | Ratings and children’s verbal comprehension skills ($r = .31$)  
Ratings and speech discrimination in noise ($r = -.32$) |
<p>| (2007)           | Aged 4 -5 years |                                         |                                                                                             |</p>
<table>
<thead>
<tr>
<th>Authors</th>
<th>Participants</th>
<th>Tool Utilized</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boynton-Hauerwas &amp; Stone (2000)</td>
<td>86 children Aged 5 – 7 years</td>
<td>Teacher Ratings</td>
<td>Children with SLI scores correlated with Expressive language subtest ($r = .55$) Auditory Comprehension Test ($r = .35$) TD children correlated with Expressive Language Subtest ($r = .12$) Auditory Comprehension Test ($r = .06$)</td>
</tr>
<tr>
<td>Bedore et al. (2001)</td>
<td>549 children Aged 4 – 11 years</td>
<td>Teacher Ratings</td>
<td>Correlations w/ a morphosyntax subtest ($r = .23$)</td>
</tr>
<tr>
<td>Bishop &amp; McDonald (2006)</td>
<td>245 children Aged 9 – 10 years</td>
<td>CCC-2 (the general communication composite &amp; social interaction index)</td>
<td>Sen. = .91 Spec. = .46</td>
</tr>
<tr>
<td>Timler (2014)</td>
<td>44 children Aged 5 – 8 years</td>
<td>CCC-2 (parents administered)</td>
<td>Sen. = 1.0 Spec. = .85</td>
</tr>
<tr>
<td>Rodriguez &amp; Guiberson (2011)</td>
<td>353 children Aged 4 years</td>
<td>TROLL</td>
<td>Correlated w/ language measures: Expressive ($r = .22$) Receptive ($r = .20$) Fail Rates: English- Speaking: 20% Spanish-Speaking: 35% Bilingual (English/ Spanish): 32%</td>
</tr>
</tbody>
</table>
APPENDIX B:
STANDARD SCORES OF CHILDREN WITH LOW PTONI SCORES

<table>
<thead>
<tr>
<th>Subject Number</th>
<th>Clinical Status</th>
<th>PTONI</th>
<th>GFTA-2</th>
<th>PPVT-4</th>
<th>DELV-NR</th>
<th>CCC-2</th>
<th>TROLL&lt;sup&gt;†&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>3215</td>
<td>SLI</td>
<td>53</td>
<td>91</td>
<td>73</td>
<td>2</td>
<td>73</td>
<td>66</td>
</tr>
<tr>
<td>978</td>
<td>SLI</td>
<td>68</td>
<td>106</td>
<td>75</td>
<td>3</td>
<td>52</td>
<td>41</td>
</tr>
<tr>
<td>994</td>
<td>SLI</td>
<td>73</td>
<td>110</td>
<td>96</td>
<td>4</td>
<td>107</td>
<td>89</td>
</tr>
<tr>
<td>965</td>
<td>SLI</td>
<td>74</td>
<td>107</td>
<td>77</td>
<td>3</td>
<td>88</td>
<td>83</td>
</tr>
<tr>
<td>984</td>
<td>SLI</td>
<td>76</td>
<td>114</td>
<td>97</td>
<td>5</td>
<td>106</td>
<td>86</td>
</tr>
<tr>
<td>962</td>
<td>SLI</td>
<td>78</td>
<td>103</td>
<td>74</td>
<td>4</td>
<td>95</td>
<td>84</td>
</tr>
<tr>
<td>974</td>
<td>SLI</td>
<td>81</td>
<td>107</td>
<td>103</td>
<td>6</td>
<td>103</td>
<td>90</td>
</tr>
<tr>
<td>948</td>
<td>SLI</td>
<td>82</td>
<td>102</td>
<td>76</td>
<td>5</td>
<td>85</td>
<td>79</td>
</tr>
<tr>
<td>956</td>
<td>SLI</td>
<td>82</td>
<td>94</td>
<td>82</td>
<td>6</td>
<td>104</td>
<td>87</td>
</tr>
<tr>
<td>989</td>
<td>SLI</td>
<td>82</td>
<td>102</td>
<td>85</td>
<td>7</td>
<td>98</td>
<td>68</td>
</tr>
<tr>
<td>947</td>
<td>SLI</td>
<td>84</td>
<td>108</td>
<td>86</td>
<td>7</td>
<td>94</td>
<td>83</td>
</tr>
<tr>
<td>3157</td>
<td>SLI</td>
<td>84</td>
<td>104</td>
<td>89</td>
<td>4</td>
<td>71</td>
<td>71</td>
</tr>
<tr>
<td>964</td>
<td>SLI</td>
<td>84</td>
<td>107</td>
<td>86</td>
<td>5</td>
<td>96</td>
<td>82</td>
</tr>
<tr>
<td>967</td>
<td>SLI</td>
<td>84</td>
<td>105</td>
<td>66</td>
<td>3</td>
<td>61</td>
<td>59</td>
</tr>
<tr>
<td>972</td>
<td>TD</td>
<td>84</td>
<td>107</td>
<td>108</td>
<td>10</td>
<td>114</td>
<td>95</td>
</tr>
<tr>
<td>988</td>
<td>TD</td>
<td>84</td>
<td>110</td>
<td>104</td>
<td>10</td>
<td>88</td>
<td>76</td>
</tr>
</tbody>
</table>

Notes.  <sup>†</sup>TROLL scores are total scores.
APPENDIX C: IRB CONSENT FORM

Institutional Review Board Document

Project Report and Continuation Application
(Complete and return to IRB, 131 David Boyd Hall. Direct questions to IRB Chairman Robert Mathews 578-8692.)

IRB#: 2792 Your Current Approval Expires On: 7/15/2013
Review type: Expedited Risk Factor: Minimal
PI: Janna Oettl Dept: COMD Phone: (225) 578-3932
Student/Co-Investigator: see below

Project Title: Tense and Subject-Verb Agreement in SAAE and SWE by Dialect Density and SLI Status
Number of Subjects Authorized: 500

Please read the entire application. Missing information will delay approval.

I. PROJECT FUNDED BY: NICDC LSU proposal #: 33813

II. PROJECT STATUS: Check the appropriate blank(s); and complete the following:
   1. Active, subject enrollment continuing; # subjects enrolled: 449
   2. Active, subject enrollment complete; # subjects enrolled: 
   3. Active, subject enrollment complete; work with subjects continues.
   4. Active, work with subjects complete; data analysis in progress.
   6. Project start postponed
   7. Project cancelled; no human subjects used.

III. PROTOCOL: (Check one).
   Protocol continues as previously approved
   Changes are requested*
      • List (on separate sheet) any changes to approved protocol.

IV. UNEXPECTED PROBLEMS: (did anything occur that increased risks to participants):
   > State number of events since study inception: 0 since last report 0
   > If such events occurred, describe them and how they affect risks in your study, in an attached report.
   > Have there been any previously unreported events? Y/IN?
      (If YES, attach report describing event and any corrective action).

V. CONSENT FORM AND RISK/BENEFIT RATIO:
   Does new knowledge or adverse events change the risk/benefit ratio? Y/IN;
   Is a corresponding change in the consent form needed? Y/IN

VI. ATTACH A BRIEF, FACTUAL SUMMARY of project progress/results to show continued participation of subjects
   is justified; or to provide a final report on project findings.

VII. ATTACH CURRENT CONSENT FORM (only if subject enrollment is continuing); and check the appropriate blank:
   1. Form is unchanged since last approved
   2. Approval of revision requested herewith: (Identify changes)

Signature of Principal Investigator: [Signature] Date: 5/15/13

IRB Action: ✓ Continuation approved: Approval Expires: 5/10/14
   Disapproved
   File closed

Signed: [Signature] Date: 5/12/13

Form date: April 16, 2008

87
APPENDIX D: IRB CONTINUATION FORM

ACTION ON PROTOCOL CONTINUATION REQUEST

TO: Janna Oetting  
COMD

FROM: Dennis Landin  
Chair, Institutional Review Board

DATE: January 12, 2015

RE: IRB# 2792

TITLE: Tense and Subject-verb Agreement in SAAE and SWE by Dialect Density and SLI Status

New Protocol/Modification/Continuation: Continuation

Review type: Full ___ Expedited X ___ Review date: 1/8/2015

Risk Factor: Minimal ___ X ___ Uncertain ________ Greater Than Minimal ________

Approved ___ X ___ Disapproved ________

Approval Date: 1/8/2015  Approval Expiration Date: 1/7/2016

Re-review frequency: (annual unless otherwise stated)

Number of subjects approved: 500

LSU Proposal Number (if applicable): 33813

Protocol Matches Scope of Work in Grant proposal: (if applicable) ___

By: Dennis Landin, Chairman

PRINCIPAL INVESTIGATOR: PLEASE READ THE FOLLOWING –
Continuing approval is CONDITIONAL on:
1. Adherence to the approved protocol, familiarity with, and adherence to the ethical standards of the Belmont Report, and LSU’s Assurance of Compliance with DHHS regulations for the protection of human subjects*
2. Prior approval of a change in protocol, including revision of the consent documents or an increase in the number of subjects over that approved
3. Obtaining renewed approval (or submittal of a termination report), prior to the approval expiration date, upon request by the IRB office (irrespective of when the project actually begins); notification of project termination.
4. Retention of documentation of informed consent and study records for at least 3 years after the study ends.
5. Continuing attention to the physical and psychological well-being and informed consent of the individual participants, including notification of new information that might affect consent.
6. A prompt report to the IRB of any adverse event affecting a participant potentially arising from the study,
8. SPECIAL NOTE:
   *All investigators and support staff have access to copies of the Belmont Report, LSU’s Assurance with DHHS, DHHS (45 CFR 46) and FDA regulations governing use of human subjects, and other relevant documents in print in this office or on our World Wide Web site at http://www.lsu.edu/irb
VITA

Kyomi Dana Gregory was born in May 1983, in Brooklyn, New York. She is a graduate of Queens Gateway to Health Sciences Secondary School. After high school, Kyomi attended the State University of New York at New Paltz and majored in communication sciences and disorders. After graduating from New Paltz College in 2005, Kyomi attended the State University of New York at Buffalo where she received the Arthur A. Schomburg Fellowship. Kyomi graduated from the State University of New York at Buffalo with her Master of Arts degree in 2007. Upon completing her master’s degree, Kyomi completed her clinical fellowship in both adult and pediatric settings. After the completion of her clinical fellowship, Kyomi worked for the New York City Board of Education for three years at an elementary and junior high school. In the fall of 2011, Kyomi enrolled in the doctoral program in the Department of Communication Sciences and Disorders at Louisiana State University. While at Louisiana State University, Kyomi was a member of the Child Language Development and Disorders Laboratory and was funded by a National Institute of Health Grant. Kyomi will join the faculty of Salus University in Elkins Park, Pennsylvania as an Assistant Professor in January 2016.