8-2017

Concurrent and Predictive Validity of the Universal Nonverbal Intelligence Test – Group Ability Test

Brooke Lauren Browarnik
University of Tennessee, Knoxville, bbrowarn@vols.utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Brooke Lauren Browarnik entitled "Concurrent and Predictive Validity of the Universal Nonverbal Intelligence Test – Group Ability Test." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in School Psychology.

R. Steve McCallum, Major Professor

We have read this dissertation and recommend its acceptance:

Sherry M. Bell, Christopher Skinner, Merilee McCurdy

Accepted for the Council:
Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Concurrent and Predictive Validity of the Universal Nonverbal Intelligence Test – Group Ability Test

A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Brooke Lauren Browarnik
August 2017

Acknowledgements

Thank you to everyone who has helped me reach this point in my life. I particularly would like to thank the greatest niece and nephew in the world, Taylor and Tony. Your support means so much.

I would also like to thank Dr. R. Steve McCallum for all of your guidance and assistance over the past four years. I really appreciate all the time you put into helping me complete this study. To Dr. Sherry Bell, Dr. Christopher Skinner, and Dr. Merilee McCurdy: thank you so much for serving on my committee; your suggestions and input were incredibly helpful.

Abstract

Within the framework of a pretest/posttest design, relations among the Universal Nonverbal Intelligence Test – Group Ability Test (UNIT-GAT) and two measures of reading achievement: the Test of Silent Word Reading Fluency, Second Edition (TOSWRF-2) and the Test of Silent Contextual Word Reading Fluency, Second Edition (TOSCRF-2) were examined for 140 children between the ages of 6 and 15, enrolled in one of three Boys & Girls Clubs in the eastern United States. Based on a counterbalanced administration at pretest, UNIT-GAT Analogical Reasoning (AR) scores moderately correlated with TOSWRF-2 Form A scores (r = .45, p < 01), TOSWRF-2 Form B scores (r = .41, p < .01), TOSCRF-2 Form A scores (r = .45, p < .01), and TOSCRF-2 Form B scores (r = .49, p < .01). Further, among students who attended at least 15 tutoring sessions, only 1 of 4 correlation coefficients between UNIT-GAT AR scores and the posttest reading change scores was significant, UNIT-GAT AR and TOSWRF-2 Form B; (p < 01), but this value is essentially meaningless, as it was obtained from only 2 participants. Even though the concurrent validity between the UNIT-GAT AR and the two reading scores is moderately strong, it does not predict change scores after tutoring under the conditions of this study. Implications are discussed.

Table of Contents

List of Tables

Chapter I

Introduction and Literature Review

"I've spent about 5% of my time at school eating lunch and the other 95% taking tests."

-- Jared, a high school Senior reflecting on his years in the public school system

The aforementioned lamentation, though hyperbolic, provides some indication of the amount of testing that occurs in schools. According to some estimates, students take an average of one state or district required test per month (Center for American Progress, 2014). Given that students are frequently assessed and the results of these assessments are often used to make important decisions, the *Principles for Professional Ethics* [The National Association of School Psychologists (NASP), 2010] reminds school psychologists to "conduct valid and fair assessments" (p. 7). One assessment goal for many school systems includes screening for academic at-risk and/or giftedness status. One type of test often used for this purpose is a group-administered test of cognitive or intellectual ability. Group-administered nonverbal intelligence tests were developed as an efficient means of validly assessing the cognitive abilities of individuals who may be disadvantaged by traditional, verbal measures (e.g., children who are hearing impaired, from environments where non-standard English is spoken). One such nonverbal test is the Universal Nonverbal Intelligence Test – Group Ability Test (UNIT-GAT; McCallum & Bracken, in press), a nonverbal group-administered cognitive screener. Currently, the UNIT-GAT is being standardized, and validity studies are needed to ascertain the UNIT-GAT's technical properties; thus, the purpose of the present study is to determine the concurrent and predictive validity of the UNIT-GAT for use with public-school age children.

Intelligence Testing

Intelligence testing has long been of interest to psychologists; however, the assessment of intelligence did not become an important practice in psychology until the 20th century when Alfred Binet and Theodore Simon developed and began to use intelligence tests in France to determine academic at-risk status. These tests were eventually imported into the United States just after the turn of the last century and have been used for multiple purposes since (Gottfredson & Saklofske, 2009). For K-12 students, IQ tests have been used extensively to determine eligibility for special education and, more specifically, to ascertain whether a student has a learning disability, is intellectually disabled, or is gifted (Newton & McGrew, 2010). Even before IQ tests were routinely employed in schools, they were used by the U.S. military. For example, as early as World War I, potential recruits were administered either the Army Alpha (verbal) or Beta (nonverbal) test (Gottfredson & Saklofske, 2009). More recently, intelligence tests have been used for other purposes (e.g., to help adults identify appropriate vocational trajectories, provide students with necessary supports, and determine clients' resilience and mental wellbeing; Gottfredson & Saklofske, 2009).

Although intelligence tests have been the focus of considerable criticisms (e.g., Harris, Smith, & Harris, 2011), they continue to be used, primarily because they predict real world experiences well (and much of the criticism actually targeted misuse, rather than poor psychometrics). For example, Fuchs and Young (2006) found that intelligence predicted responsiveness to reading interventions in eight of thirteen studies. Moreover, because of their increasing sophistication, they currently are being used to identify cognitive underpinnings of specific academic deficits in reading and math (Kaufman & Kaufman, 2004; Mather & Jaffee, 2011; Naglier, 1996; Schrank, Mather, & McGrew, 2014). Although some early criticisms of tests were well founded (e.g., not all included representation of minority group members, nor students with disabilities in their standardization samples), too often intelligence tests have been misused (i.e., put to uses well beyond their intended goals), and much of the criticisms leveled against them is bogus (i.e., they are not biased in a technical sense against minority populations (Gutkin & Reynolds, 2009; Jensen, 1980). Consequently, they continue to be used in a variety of settings for multiple purposes (e.g., determine cognitive strengths/weaknesses in support of instructional planning, screen for intellectual disability or giftedness, predict important real-world academic and/or vocational outcomes).

Brief History of Nonverbal Intelligence Testing

During the First World War, two test developers, Yoakum and Yerkes, realized that traditional measures of intelligence may not be suitable for all individuals (Naglieri & Otero, 1997). They believed verbal intelligence tests may place individuals who are not fluent in English at a disadvantage; consequently, they created the first useful nonverbal intelligence test, the nonverbal (figural) Beta (Yoakum & Yerkes, 1920). The Army Beta, theoretically, could be used to assess the intelligence of Army recruits who were illiterate or, for some reason, not proficient in English (Yoakum & Yerkes, 1920).

Building on their rationale based on an increasingly diverse nation, present day researchers have developed better nonverbal measures. For instance, Abbott and McQuarrie (2015) write that individuals with limited English proficiency tend to obtain artificially low test scores if given verbal measures of intelligence because such measures are generally created with hearing, native English-speakers in mind. This finding is worrisome because, according to the United States Census Bureau (2013), 21% of people living in the United States who are over the age of five do not speak English at home. In addition, a recent report from the U.S. Department of Education (2013) showed 19 prominent languages spoken by students in English-Language Learner (ELL) programs. Spanish was the most prominent (71% of ELL students nationally). Other prominent languages included Chinese, Arabic, Vietnamese, Haitian, Russian, Navajo, and 12 other languages ranked second or third in frequency among the 21 U.S. States with 45,000 or more ELL students. Furthermore, cognitive abilities are often listed in the clinical criteria for disorders in the DSM-5 (American Psychiatric Association, 2013), frequently as “rule out” conditions. For example, for language, communication, and learning disorders (315.00, 315.1, 315.2, 315.39), the condition cannot be due to low intellectual ability to satisfy the criteria. Thus, using an estimate of IQ, found to be near average or above, would rule out intellectual deficiency as a reason for the disorder. Finally, from a Survey of Income and Program Participation (SIPP), Mitchell (2005) reports there are about 10,000,000 people who are deaf or hard-of-hearing in the United States.

Despite this need for nonverbal assessments, many current “nonverbal” tests are not truly nonverbal. They often require at least some verbal communication from the test administrator and examinee. For example, directions for the nonverbal portion of the Kaufman Assessment Battery for Children, Second Edition (KABC-II; Kaufman & Kaufman, 2004) and the Stanford-Binet, Fifth Edition (SB5; Roid, 2003) are verbal/oral. Additionally, some KABC tasks require examinees to orally respond to test items. Authors of nonverbal tests are increasingly aware and devoted to reducing as much as possible the language and cultural demands of nonverbal tests (see Bracken & McCallum, 1998, 2016).

Relation of Nonverbal Intelligence Tests to Achievement Testing.

Because there are significant numbers of U.S. citizens who may be disenfranchised by administration of verbally loaded tests, it is important to develop valid and reliable nonverbal intelligence measures for use with individuals who are not proficient in English and/or have hearing difficulties. There is ample evidence to suggest that scores on intelligence tests, even nonverbal ones, correlate with other important variables. In particular, scores on intelligence tests have been found to correlate significantly with academic achievement (Armstrong, 2012; McBee & Duke, 1960; McGrew & Wendling, 2010; Naglieri, 1996) and vocational success (Amdurer, Boyatzis, Saatcioglu, Smith, & Taylor, 2014).

Intuitively, it seems that verbal tests of intelligence should correlate more highly with academic success than nonverbal tests and, in the main, they do. However, as noted above, the relation between nonverbal intelligence tests and academic outcomes is strong. Several nonverbal intelligence measures have been found to correlate with academic achievement. According to the test's authors, the Wechsler Nonverbal Scale of Ability

(WNV; Wechsler & Naglieri, 2006), a measure of general cognitive ability for individuals under the age of 21, significantly correlates with the scores from the Wechsler Individual Achievement Test-Second Edition (WIAT-II; Wechsler, 2001) scores (*r* values are around .60). In addition, The Naglieri Nonverbal Ability Test (NNAT; Naglieri, 1996) has been found to correlate significantly with achievement as well. England (2003) found that scores on the NNAT correlate with scores on the Wechsler Individual Achievement Test-Second Edition (WIAT-II; Wechsler, 2001; around .55). The Test of Nonverbal Intelligence, Fourth Edition (TONI-4; Brown, Sherbenou, & Johnsen, 2010) also correlates with achievement as measured by several commonly used achievement tests (*r* = .55 to .78).

The Universal Nonverbal Intelligence Test (UNIT; Bracken & McCallum, 1998), a multidimensional test of cognition, also significantly correlates with achievement. According to Bracken and McCallum (1998), the UNIT Abbreviated, Standard, and Extended batteries Full Scale Intelligence Quotient (FSIQ) scores correlate with Wechsler Individual Achievement Test (WIAT; Wechsler, 1992) scores (*r* = .53, .62, and .59 respectively) for 31 students between the ages of 5 and 14. More recently, Bell, McConnell, Lassiter, and Matthews (2013) administered the UNIT and the Woodcock-Johnson III Tests of Achievement (WJ-III; Woodcock, McGrew & Mather, 2001) to 121 children and found that UNIT FSIQ scores moderately correlate with math achievement (*r* = .64). Similarly, the Universal Nonverbal Intelligence Test--Second Edition (UNIT2; McCallum & Bracken, in press) has been found to significantly correlate with the Diagnostic Achievement Battery–Fourth Edition (DAB-4; Newcomer, 2014), the

Kaufman Test of Educational Achievement–Second Edition (KTEA-II; Kaufman & Kaufman, 2005), the Woodcock-Johnson III Tests of Achievement (WJ III; Woodcock, McGrew, & Mather, 2001), the Test of Silent Word Reading Fluency–Second Edition (TOSWRF-2; Mather, Hammill, Allen, & Roberts, 2014), the Test of Silent Contextual Reading Fluency–Second Edition (TOSCRF-2; Hammill, Wiederholt, & Allen 2014), and the Test of Written Spelling–Fifth Edition (TWS-5; Larsen, Hammill, & Moats, 2013). Correlations typically range from .54 to .79, with UNIT2 Quantitative scores correlating more strongly with math achievement and UNIT2 Reasoning scores correlating more strongly with reading achievement.

Nonverbal, Group-Administered Tests. Most of the nonverbal assessments described above were developed for individual administration; however, some tests have the capacity to be administered to groups as well. However, it should be noted that those nonverbal tests that can be given in groups were often developed for one-on-one administration. For example, the Comprehensive Test of Nonverbal Intelligence, Second Edition (CTONI-2; Hammill, Pearson, & Wiederholt, 2009) can be administered in an individual or group format, but it was not standardized on groups. Whether they were developed for individual or group administration, nonverbal, group-administered tests have been found to moderately correlate with achievement. Naglieri and Ronning (2000a; 2000b) administered the Naglieri Nonverbal Ability Test (NNAT; Naglieri, 1996) in groups to African American, White, and Hispanic children and found that NNAT scores correlate significantly with SAT-9 math (r = .63) and reading scores (r = .52). Similarly, Wills (2012) examined the relation between NNAT scores in

Kindergarten and achievement scores in third grade. Wills' participants were divided into two groups – Cohort 1 and Cohort 2. Both cohorts were administered the NNAT and the Missouri Assessment Program (MAP), a measure of reading and writing achievement; however, Cohort 2 completed the assessments a year before Cohort 1. NNAT scores moderately correlate with MAP reading and writing scores for Cohort 1 (r = .50) and Cohort 2 (r = .44).

The Universal Nonverbal Intelligence Test- Group Ability Test (UNIT-GAT). Because group-administered, nonverbal tests are quick and easy to administer and can help practitioners balance large caseloads, more information is needed to determine if they relate significantly to real-world outcomes. Consequently, data currently are being collected on a new group-administered, nonverbal instrument: the Universal Nonverbal Intelligence Test – Group Ability Test (UNIT-GAT) to help determine validity and reliability. The UNIT-GAT is unique from most other nonverbal tests because it was standardized on groups, measures analogical and quantitative reasoning, and contains items that are sensitive to nonsymbolic (abstract) and symbolic thought. The UNIT-GAT is also efficiently, easily, and quickly administered. According to PRO-ED researchers (E. A. Allen, personal communication, February 15, 2016), the UNIT-GAT Full Scale IQ strongly correlates with the Universal Nonverbal Intelligence Test, Second Edition (UNIT2; McCallum & Bracken, in press) Memory Composite (r = 0.69), Reasoning Composite (r = 0.70), Quantitative Composite (r = 0.74), and Full Scale Intelligence Quotient (r = 0.81). UNIT-GAT FSIQ scores also correlate significantly with the TOSCRF-2 (r = 0.80) and TOSWRF-2 (r = 0.84). In addition, PRO-ED researchers found UNIT-GAT Quantitative Reasoning Scores and Analogical Reasoning Scores are strongly correlated (*r* = 0.88). Because the UNIT-GAT is currently in development, its psychometric properties need to be more firmly established. Assuming the UNIT-GAT's validity and reliability data are adequate, it has the potential to help practitioners screen individuals when group-administered cognitive scores are needed.

Statement of Purpose

With the multitude of achievement and intelligence measures available, choosing the best measure for a particular client can be challenging for a psychologist; thus, test developers and researchers work to determine the psychometric integrity of available instruments to inform test selection decisions. Psychometric integrity is particularly important when a new test is developed, and validity data are critical. Validity "concerns what an instrument measures and how well it does that task" (Whiston, 2013, p. 58). With this definition in mind, the aim of the present study is to provide data informing the validity of the Universal Nonverbal Intelligence Test – Group Ability Test (UNIT-GAT; McCallum & Bracken, in press), a nonverbal group-administered cognitive screener. Determining validity is an ongoing process that requires the results of many studies. In addition, there are various "types" of validity—a test may be valid for one purpose but not another. The specific goals of this study involve obtaining evidence regarding the concurrent and predictive validity of the UNIT-GAT when the criterion measures are brief, group-administered operationalizations of reading. In the service of this goal, on two occasions, the UNIT-GAT, the Test of Silent Word Reading Fluency, Second Edition (TOSWRF-2), and the Test of Silent Contextual Reading Fluency, Second Edition

(TOSCRF-2) were administered to 140 children between the ages of 6 and 15 in counterbalanced order. The specific research questions to be addressed are as follows:

1) What is the relation between UNIT-GAT Analogical Reasoning pretest scores and TOSWRF-2 pretest scores?
2) What is the relation between UNIT-GAT Analogical Reasoning pretest scores and TOSCRF-2 pretest scores?
3) To what extent do UNIT-GAT Analogical Reasoning pretest scores predict change scores on the TOSWRF-2? Follow up analysis investigated how UNIT-GAT scores predict change scores for students in one of two reading interventions: Reading fluency and comprehension tutoring and a books condition (These interventions will be described in more detail in a subsequent section).
4) To what extent do UNIT-GAT Analogical Reasoning pretest scores predict change scores on the TOSCRF-2? Follow up analysis investigated how UNIT-GAT scores predict change scores for students in one of two reading interventions: Reading fluency and comprehension tutoring and a books condition.

Hypotheses

Hypotheses address the relation between the UNIT-GAT Analogical Reasoning scores, TOSWRF-2 scores, and TOSCRF-2 scores and the predictive power of the UNIT-GAT. After culling the previously summarized research, several results are expected.

H_o1 It is hypothesized that UNIT-GAT Analogical Reasoning pretest scores will significantly correlate with TOSWRF-2 Form A pretest scores in a positive direction.

H_o2 It is expected that UNIT-GAT Analogical Reasoning pretest scores will significantly correlate with TOSCRF-2 Form A pretest scores in a positive direction.

H_o3 It is hypothesized that UNIT-GAT Analogical Reasoning pretest scores will significantly correlate with TOSWRF-2 Form B pretest scores in a positive direction.

H_o4 It is hypothesized that UNIT-GAT Analogical Reasoning pretest scores will significantly correlate with TOSCRF-2 Form B pretest scores in a positive direction.

The power of the statistical tests conducted to address these a priori hypotheses is increased, based on predictions from the literature. That is, the significance of the obtained values will be determined by examining *p* values under only one tail of the relevant distributions.

Chapter II

Methods

Participants

Participants were 140 children between the ages of 6 and 15 enrolled in one of three Boys & Girls Clubs in the southeastern United States. All of the children were entering the first, second, third, fourth, fifth, sixth, or seventh grade in the Fall of 2014, with the majority reading one grade level or more below their same-aged peers. Most of the students attended urban Title I schools in the area. Gender demographics were 56.4% (n = 79) males, 44.1% (n = 64) females. Race demographics were 61.4% African American, 24.3% White, and 14.3% multiracial. Six participants (4.3%) reported they were Hispanic. Participants were primarily from low-income families, with 85.7% of the children receiving free or reduced price lunch. Each participating student was assigned a number during the assessment stage and was tracked based on this number to ensure confidentiality of information. The University of Tennessee's Institutional Review Board guidelines were followed.

Setting

Boys & Girls Club A. The majority of the research was conducted on-site at a large Boys & Girls Club (BGC) facility in a mid-sized city in the southeastern U.S. The center is located in an older part of downtown in a building that is only a few years old. It is modern and bright, has open areas, and offers a variety of activities for children ages 5 through 17. Eighty-one participants were recruited from this site.

Boys & Girls Clubs B and C. Fifty-nine additional participants were obtained from two smaller BGC sites. Similar to the main BGC location, these sites are also located in the southeastern U.S. and served students from similar demographic backgrounds. Participants recruited from the smaller BGC sites were involved in a books-only treatment (described in later sections).

Instrumentation

Three instruments were used in this study: the UNIT-GAT, the TOSWRF-2, and the TOSCRF-2. The relations between these measures was examined in order to determine concurrent and predictive validity. A description of the measures follows:

The Universal Nonverbal Intelligence Test- Group Ability Test (UNIT-GAT). Riverside Publishing (n.d.) describes the Universal Nonverbal Intelligence Test- Group Ability Test (UNIT-GAT; McCallum & Bracken) as a measure of cognitive ability designed for use with groups of individuals between the ages of 4 and 30. Besides instructions which are provided verbally, the test is entirely nonverbal.

The UNIT-GAT is divided into two sections: Analogical Reasoning and Quantitative Reasoning. Two primary scales (Memory and Quantitative Reasoning) and items that assess symbolic and nonsymbolic thought are built into these sections (McCallum, personal communication, 2016). Each Analogical Reasoning question consists of a row (or rows) of items depicted in boxes, a question mark in a box, and four answer choices. Examinees identify relations between the boxed items as they move across the rows and choose the answer choice that belongs in the box with the question mark. Half of the boxed items are pictures of commonly seen objects (e.g., birds), and the rest of the boxed items are pictures of geometric shapes. Questions that contain pictures of commonly seen objects assess symbolic thought, while questions that contain pictures of geometric shapes assess nonsymbolic thought.

The Quantitative Reasoning section of the UNIT-GAT is similar to the Analogical Reasoning section; examinees identify relationships between items and choose the appropriate answer choice. Additionally, like the Analogical Reasoning nonsymbolic items, the Quantitative Reasoning nonsymbolic items contain geometric shapes. The Quantitative Reasoning symbolic items, however, contain numbers rather than pictures of common objects. Internal consistency coefficients have been obtained from the publisher, PRO-ED, and range from .91 to .99. The average coefficient alpha from the current sample (i.e., children between the ages of 6 and 15) for the Full Scale UNIT-GAT is .97. The average coefficient alpha from the current sample for the UNIT-GAT Analogical Reasoning subtest is .95 and .93 for UNIT-GAT Quantitative Reasoning subtest.

Test of Silent Word Reading Fluency, Second Edition (TOSWRF-2). The Test of Silent Word Reading Fluency, Second Edition (TOSWRF-2; Mather, Hammil, Allen, & Roberts, 2014) is presumed to be a measure of the ability of individuals between the ages of 6 and 24 to successfully identify printed words in 3 mins. Examiners are shown a series of words without spaces or punctuation and are asked to draw slashes where one word ends and another begins (e.g., d i m h o w f i g b l u e). Test-retest reliability for the TOSWRF-2 is robust (coefficients range from .84 to .91). Validity for the TOSWRF-2 is also high. TOSWRF-2 scores significantly correlate with scores on cognitive measures, such as the Stanford-Binet, Fifth Edition (SB5; Roid, 2003; *r* = .56), the Test of Nonverbal Intelligence, Fourth Edition (TONI-4; Brown, Sherbenou, & Johnsen, 2010; *r* = .55), and the Comprehensive Test of Nonverbal Intelligence, Second Edition (CTONI-2; Hammill, Pearson, & Wiederholt, 2009; *r* = .50).

Test of Silent Contextual Reading Fluency, Second Edition (TOSCRF-2). The Test of Silent Contextual Reading Fluency, Second Edition (TOSCRF-2; Hammill, Wiederholt, & Allen, 2014) is similar to the TOSWRF-2; however, test takers must separate strings of letters into words, and these words form sentences (e.g., A Y E L L O W B I R D W I T H B L U E W I N G S). Because test takers identify words in context, the TOSCRF-2 is presumed to be a measure of both fluency and comprehension. Test-retest reliability for this measure is robust (coefficients range from .82 to .90). Validity for the TOSCRF-2 is also high. TOSCRF-2 scores have been found to significantly correlate with Woodcock-Johnson III Tests of Achievement (WJ-III; Woodcock, McGrew & Mather, 2001) Broad Reading scores (*r* = .69). TOSCRF-2 scores also correlate with scores on cognitive measures, such as the Stanford-Binet, Fifth Edition (SB5; Roid, 2003; *r* = .48), the Test of Nonverbal Intelligence, Fourth Edition (TONI-4; Brown, Sherbenou, & Johnsen, 2010; *r* = .56), and the Comprehensive Test of Nonverbal Intelligence, Second Edition (CTONI-2; Hammill, Pearson, & Wiederholt, 2009; *r* = .57).

Procedures

Procedures were implemented and monitored by two University of Tennessee faculty members, three Special Education doctoral students, and two School Psychology doctoral students. Additional doctoral students and an undergraduate student were recruited and trained to assist with assessment administration and scoring.

Children completed assessments at the BGC in groups of approximately 20 students in early June 2014. Each participant was administered Form A of the TOSWRF-2, Form A of the TOSCRF-2, and Form B of either the TOSWRF-2 or TOSCRF-2 in counterbalanced order. Each student then completed the UNIT-GAT. Testing sessions lasted no more than 1.5 hours. Form C was administered after the interventions, as described below.

Tutoring. In the four weeks following pretest assessment, students were placed in one of two groups. Students in the first group received tutoring at one of the BGC sites. Tutoring was conducted by novice special educators under the supervision of Special Education faculty enrolled in the Modified Summer Institute (MSI), a program that provides 12 hours of intensive coursework related to high incidence disabilities. Coursework met the requirements from the Tennessee Department of Education to earn licensure in the area of modified special education. Seven novice special educators planned and implemented evidence-based instruction in the area of literacy daily. Specifically, instruction was focused on pre-reading, vocabulary, word work, and comprehension strategies. The teachers used a popular and culturally and curriculum relevant children's book each (e.g. *The Great Kapok Tree*) week to plan lessons. Four different books were used throughout the summer tutoring, one per week. Literacy strategies and lessons were developed with the help of one of the instructors of the MSI.

Tutoring occurred daily for four weeks throughout the summer. Each session was approximately two hours in length and included instruction related to the major areas of reading. Using pretest assessment data collected at the beginning of the study, students were grouped by ability levels in to one of three classrooms. Groups consisted of about 30 students with similar ability levels and used the same children's book each week; however, the lessons differed in order to meet the needs of all diverse learners. The MSI instructor, who was onsite daily during tutoring, reviewed lesson plans weekly. Each group also consisted of at least one doctoral student in Special Education. These students served as teacher mentors in order to help implement effective instruction.

Students in the tutoring condition also received intervention in the area of reading fluency. Every student participated in sustained silent reading (SSR). Known by numerous names (e.g., Sustained Silent Reading, Drop Everything and Read, Wide Independent Reading), SSR allows participants an extended, structured time to silently read (Faggella-Luby & Wardwell, 2011). SSR, which allows students to read without worrying about assignments or grades, promotes reading motivation and improves reading achievement scores (Faggella-Luby & Wardwell, 2011; Siah & Kwok, 2010).

During SSR, students self-selected books to read independently for approximately 15 to 20 minutes daily. Books were age-appropriate and culturally and curriculum relevant for students. The levels of books varied to provide access to all learners of various ages and abilities. The number of books read throughout the SSR period differed for each student, some children read longer, more-advanced chapter books, while others read picture books. The teachers and mentors in each classroom did not select the books for participants. If a student asked for help, the teachers simply prompted the students to select a book they want and are able to read.

Books-only. Participants placed in the second group had access to free, self-selected books throughout the summer because some research has shown that access to print materials can counteract summer reading loss and increase engagement and motivation to read for leisure (Allington & McGill-Franzen, 2013). Students in this condition participated in a book fair during the pre-assessment phase. After completion of the assessments, students were asked to select 12 books of their choice. Although encouraged to select a book that was appropriate for their specific reading level, students were allowed to select any book of their choice. Similarly to the procedure described by Allington and colleagues (2010), books varied by readability and topic. Books were chosen to represent the following categories: pop culture, series books, and culturally and curriculum relevant. Books were left at the participating BGC site for students to take home each week. Students were encouraged to maintain a simple book log throughout the summer, although it was not required for participation**.** The book log asked students to write about their favorite part of the story that they read each day. A collection box was stored at the BGC for students to drop off their logs at their convenience. Four weeks after pre-assessment, participants in groups one and two were again administered the TOSWRF-2 and TOSCRF-2.

Chapter III

Results

The concurrent and predictive validity of the UNIT-GAT was explored by the administration of the UNIT-GAT, TOSWRF-2, and TOSCRF-2 to a sample of 140 children between the ages of 6 and 15 who attended a summer tutoring program. Each of the hypotheses and research questions proposed in the present study was analyzed using Pearson product-moment correlation coefficients, and correlations were evaluated using Cohen's (1988) criteria for interpreting effect size and Coefficients of Determination (r^2) were reported, as appropriate. Coefficients of Determination reflect the proportion (or percentage) of shared variance between two measures.

Raw score means and standard deviations are reported in Tables 1 and 2, and standard score means and standard deviations are shown in Tables 3 and 4; all statistics are reported in tables located in the Appendix. For context, TOSWRF-2 (Forms A and B) and TOSCRF-2 (Forms A and B) standard scores are based on a mean of 100 and a Standard Deviation of 15. Most of the mean TOSWRF-2 and TOSCRF-2 standard scores ranged from approximately 84 to 95, and standard score standard deviations hover around 15. In general, these standard score means are lower than expected, relative to those from national peers; the standard deviations are roughly consistent with population variation, in general. The relatively low standard score means earned by these participants on the reading tests are not unexpected, given the demographics of this sample. That is, children from low socioeconomic backgrounds generally score lower than the general population when compared to peers nationwide on tests of achievement and cognition

(Lee & Burkam, 2002). TOSCRF-2 mean scores were even lower in this sample than the TOSWRF-2 scores, even though both are reading measures. A possible explanation for this finding is the discrepancy in the format between the two tests. That is, the TOSWRF-2 requires examinees to place slashes between unrelated words; on the other hand, the TOSCRF-2 requires examinees to place slashes between words that form sentences. Perhaps participants in this study found the comprehension component of the TOSCRF-2 more challenging than the word identification of the TOSWRF-2. Finally, skewness and kurtosis are reported in Table 1. In general, the distributions are slightly negatively skewed (range: -.480 to .005) and approximately normally distributed (range: -.429 to .583).

Standard scores are not yet available for the UNIT-GAT and, consequently, no comparison with peers nationwide is possible. Of note, the UNIT-GAT raw score means do show a consistent increase across age with one exception (see Table 3), as would be expected if the test is valid (i.e., cognitive ability is expected to increase as a function of chronological age). Raw score means and standard deviations are reported in Tables 1 and 3; skewness and kurtosis are reported in Table 1.

Relation Between UNIT-GAT and TOSWRF-2 and TOSCRF-2 Pretest Scores

Examination of correlation coefficients addressing the first and third hypotheses reveals that all are statistically significant. The relation between UNIT-GAT Analogical Reasoning pretest scores and TOSWRF-2 Form A pretest scores can be characterized as "moderately correlated" using Cohen's criteria (1988; r = .45, p < .01, n = 132), as can the relation between the UNIT-GAT pretest scores and the TOSWRF-2 Form B pretest scores (*r* = .41, *p* < .01, *n* = 46). Calculation of Coefficients of Determination indicates approximately 20% shared variance between UNIT-GAT and TOSWRF-2 Form A and 17% between UNIT-GAT and TOSWRF-2 Form B.

The coefficients between UNIT-GAT Analogical Reasoning pretest scores and TOSCRF-2 pretest scores are also moderately correlated, based on Cohen's criteria (1988). UNIT-GAT Analogical Reasoning pretest scores significantly correlated with TOSCRF-2 Form A pretest scores (*r* = .45, *p* < 01, *n* = 133) and TOSCRF-2 Form B scores (*r* = .49, *p* < 01, *n* = 81). Calculation of Coefficients of Determination indicates approximately 20% shared variance between UNIT-GAT and TOSCRF-2 Form A and 23% between UNIT-GAT and TOSCRF-2 Form B.

It is common practice to obtain corrections for restriction in range when appropriate. The above correlation coefficients were corrected based on consideration of population variance estimates (i.e., standard deviations) from the TOSCRF-2 and TOSWRF-2. It is impossible to correct for restriction in range for the UNIT-GAT because population standard deviations are not yet available for this measure. Population standard deviations for TOSWRF-2 and TOSCRF-2 raw scores were obtained from the respective test manuals. As is evident from Table 5, corrections did not significantly alter coefficients.

Relation Between UNIT-GAT Pretest Scores and TOSWRF-2 and TOSCRF-2 Change Scores

Analyses exploring the relations between UNIT-GAT scores and the change scores that occurred from the TOSWRF-2 Form A pretest administration to TOSWRF-2

Form C posttest administration were informed by subtracting Form A scores from Form C scores. Specifically, UNIT-GAT pretest scores were correlated with the change scores that occurred as a function of the reading interventions by subtracting pre-intervention TOSWRF-2 Form A and Form B scores from TOSWRF-2 Form C scores at posttest, which were obtained after the tutoring and books interventions.

In general, UNIT-GAT Analogical Reasoning pretest scores did not significantly correlate with TOSWRF-2 change scores (See Table 6). Among the total sample, UNIT-GAT Analogical Reasoning pretest scores did not significantly correlate with TOSWRF Form A – TOSWRF-2 Form C change scores (r = .038, p > .05) or TOSWRF-2 Form B – TOSWRF-2 Form C change scores (r = -.004, p > .05). Also, UNIT-GAT Analogical Reasoning pretest scores did not predict change scores for students in either of the two reading interventions (i.e., reading fluency and comprehension tutoring and a books condition). These coefficients are based on scores from the total sample; however, attendance was sporadic across the interventions.

In order to investigate the relations between the UNIT-GAT and TOSWRF-2 reading scores as a function of increasing participation, additional analyses were conducted (see Table 7). The pattern of coefficients is more consistent with the prediction that UNIT-GAT scores would predict change as a function of intervention (i.e., in general, the coefficients increased as participation increased). Nonetheless, with one exception, the coefficients did not approach statistical significance. Among students who attended at least 15 tutoring sessions, one correlation was significant: the one showing the relation between UNIT-GAT Analogical Reasoning pretest scores and

TOSWRF-2 Form B change scores (*r* = 1, *p* < 01). However, this value is meaningless as scores from only two participants were available for analysis.

Analyses mirroring those described to explore the predictive relation between the UNIT-GAT and TOSWRF-2 were conducted to explore the relation between UNIT-GAT scores and changes that occurred from administration of TOSCRF-2 Form A and Form B to post intervention administration of TOSCRF-2 Form C (i.e., TOSCRF-2 Form A and Form B scores were subtracted from Form C scores). UNIT-GAT Analogical Reasoning pretest scores did not significantly predict TOSCRF-2 Form A – TOSCRF2 –Form C change scores (*r* = -.041, *p* > .05) or TOSCRF-2 Form B – TOSCRF-2 Form C change scores (*r* = -.073, *p* > .05) among the total sample. See Table 7 for coefficients showing values for various levels of participation. Again, the pattern of coefficients as a function of participation was examined and, again, those coefficients were not consistent with expectations.

It is possible that these nonsignificant results occurred because the interventions were unsuccessful (i.e., there were no changes as a function of the interventions). To explore this possibility, paired-samples *t* tests were conducted to determine whether TOSWRF-2 and TOSCRF-2 pretest scores differed significantly from TOSWRF-2 and TOSCRF-2 posttest scores (see Tables 8-10). The difference between TOSWRF-2 Form A pretest scores (*M* = 71.073, *SD* = 26.9402) and TOSWRF-2 Form C posttest scores (*M* = 80.78, *SD* = 27.463); *t*(81) = 6.451, *p* < .001 is statistically significant. Similarly, there is also a significant difference between TOSWRF-2 Form B pretest scores (*M* = 64.125, *SD* = 30.6023) and TOSWRF-2 Form C posttest scores (*M* = 77.69, *SD* = 32.104); *t*(31) =

5.656, $p < .001$. The same pattern of scores was obtained for the differences between TOSCRF-2 pretest and posttest scores. Specifically a significant difference was found between TOSCRF-2 Form A pretest scores ($M = 50.060$, $SD = 25.3583$) and TOSCRF-2 Form C posttest scores ($M = 60.42$, $SD = 29.057$); $t(83) = 5.169$, $p < .001$. Similarly, there is a significant difference in TOSCRF-2 Form B pretest scores ($M = 48.149$, $SD = 25.3240$) and TOSCRF-2 Form C posttest scores ($M = 60.28$, $SD = 26.021$); $t(46) = 4.544$, $p < .001$.

To further explore these mean differences, analyses were conducted for the two interventions separately (i.e., paired-samples t-tests were conducted to determine the statistical significance of TOSWRF-2 and TOSCRF-2 pre- to posttest differences among students in the tutoring and books conditions). Among students in the tutoring condition, a significant difference in TOSWRF-2 Form A pretest scores ($M = 79.273$, $SD = 21.7762$) and TOSWRF-2 Form C posttest scores ($M = 86.61$, $SD = 20.459$) exists; $t(32) = 3.145$, $p < .004$. Also, there is a significant difference between TOSWRF-2 Form B pretest scores ($M = 79.667$, $SD = 24.8408$) and TOSWRF-2 Form C posttest scores ($M = 97.50$, $SD = 14.896$); $t(5) = 3.422$, $p < .019$. Similarly, there is a significant difference between TOSCRF-2 Form A pretest scores ($M = 53.971$, $SD = 23.8974$) and TOSWRF-2 Form C posttest scores ($M = 66.46$, $SD = 26.064$); $t(34) = 4.159$, $p < .01$). Finally, there is a significant difference in TOSCRF-2 Form B pretest scores ($M = 46.083$, $SD = 20.7300$) and TOSWRF-2 Form C posttest scores ($M = 59.38$, $SD = 21.017$; $t(23) = 3.848$, $p < .01$).

The same pattern of analyses was conducted among students in the books condition, with similar results. There is a significant difference between TOSWRF-2 Form A pretest scores (*M* = 65.551, *SD* = 28.8315) and TOSWRF-2 Form C posttest scores (*M* = 76.86, *SD* = 30.898); $t(48) = 5.784$, $p < .01$. Similarly, there is a significant difference between TOWCRF-2 Form B pretest scores (*M* = 60.538, *SD* = 31.0886) and TOSWRF-2 Form C posttest scores (*M* = 73.12, *SD* = 33.432); $t(25) = 4.650$, $p < .01$. Also, there is a significant difference between TOSCRF-2 Form A pretest scores (*M* = 47.265, *SD* = 26.2351) and TOSWRF-2 Form C posttest scores (*M* = 56.10, *SD* = 30.547); $t(48) = 3.287$, $p < .002$. Finally, there is a significant difference between TOSCRF-2 Form B pretest scores (*M* = 50.304, *SD* = 29.7007) and TOSWRF-2 Form C posttest scores (*M* = 61.22, *SD* = 30.857; $t(22) = 2.624$, $p < .015$). In summary, a significant difference between pre- and posttest means was found in every case, effectively ruling out lack of change as an explanation for the low predictive ability of the UNIT-GAT.

Chapter IV

Discussion

The purpose of the present study was to determine the technical properties of a group-administered, nonverbal intelligence test, the Universal Nonverbal Intelligence Test – Group Ability Test (UNIT-GAT). In order to determine the concurrent and predictive validity of the UNIT-GAT, 140 children between the ages of 6 and 15 completed the UNIT-GAT and two measures of reading achievement, before and after a reading intervention. In general, results suggest the UNIT-GAT has acceptable concurrent validity. On the other hand, the UNIT-GAT did not predict reading change scores after intervention under the conditions of this study. In the next section, these results are put into context, followed by a description of several limitations. Given that validity is only established over time and across multiple studies, further research is recommended.

Concurrent Validity

UNIT-GAT Analogical Reasoning pretest scores significantly correlated with Test of Silent Word Reading Fluency, Second Edition (TOSWRF-2) Form A and Form B pretest scores. UNIT-GAT Analogical Reasoning pretest scores also significantly correlated with Test of Silent Contextual Reading Fluency, Second Edition (TOSCRF-2) Form A and Form B pretest scores. Correlation coefficients were not significantly altered when corrections for restriction in range were made. Corrections were initially made because of the difference in magnitude of coefficients in this study and coefficients reported by PRO-ED, the company publishing UNIT-GAT. According to data from

PRO-ED from 55 participants, correlation coefficients between the UNIT-GAT and the TOSWRF-2 and TOSCRF-2 strengthened after correcting for range restrictions. Coefficients between UNIT-GAT and TOSWRF-2 scores changed from .51 to .84, and UNIT-GAT--TOSCRF-2 coefficients changed from .46 to .80. Typically, corrections for restriction in range make a difference when the variance estimates from one or both of the two sample distributions used to calculate the coefficients are different from population parameters. That is, if the variance estimates (e.g., standard deviations) are lower in one or both of the samples, corrected coefficients will increase; on the other hand, if the variance estimates are greater in the sample(s), the corrected coefficients will decrease. Apparently, the variance estimates in the sample data reported by PRO-ED are different from the pattern of variance estimates within this study.

Based on data from this study explicating the relation between the UNIT-GAT and the TOSWRF-2 and TOSCRF-2, it is possible to conclude that the UNIT-GAT concurrent validity is sufficient to recommend its use to practitioners. That is, coefficients between UNIT-GAT scores and TOSWRF-2 and TOSCRF-2 scores are similar to those found in the literature between verbal cognitive test scores and scores from achievement tests and between nonverbal cognitive measures and achievement test scores. For example, Armstrong (2012) and McBee and Duke (1960) examined the relation between verbal cognitive tests and achievement measures and report coefficients ranging from .31 to .81. Similarly, individually-administered, nonverbal intelligence tests have been found to correlate significantly with academic outcomes. For instance, researchers comparing the Universal Nonverbal Intelligence Test (Bracken & McCallum,

1998) or the Wechsler Nonverbal Scale of Ability (Wechsler & Naglieri, 2006) and various measures of achievement (e.g., Woodcock-Johnson III Tests of Achievement; Woodcock, McGrew, & Mather, 2001) have reported correlation coefficients that range from .53 to .79 (e.g., see Bell, McConnell, Lassiter, & Matthews, 2013; Bracken & McCallum, 1998, 2016; and Wechsler & Naglieri, 2006). Most relevant to the present study is the finding that group-administered, nonverbal tests tend to correlate with achievement also. For example, England (2003), Naglieri and Ronning (2000a; 2000b), and Wills (2012) found that Naglieri Nonverbal Ability Test (NNAT; Naglieri, 1996) scores significantly correlates with various measures of achievement, such as the Wechsler Individual Achievement Test-Second Edition (WIAT-II; Wechsler, 2001; coefficients range from .52 to .78).

In general, the correlations between UNIT-GAT scores and the reading achievement scores from the sample of students comprising this study are consistent with, but slightly weaker than, many of the correlations reported between other nonverbal, individual and/or group-administered intelligence tests and achievement tests. There are multiple explanations for this outcome. For example, current results reflect the relation between only one UNIT-GAT subtest, rather than scores from the entire test (i.e., UNIT-GAT AR plus QR scores); and, all other things being equal, correlations are sensitive to test length. That is, shorter tests typically yield coefficients that are lower than longer tests. In fact, when total UNIT-GAT FSIQ scores are compared to TOSWRF-2 and TOSCRF-2 scores, coefficients are higher than between the UNIT-GAT Analogical Reasoning subtest and the TOSWRF-2 and TOSCRF-2, and range from .44 to

.60, even though the second UNIT-GAT subtest is a measure of quantitative reasoning and is not theoretically linked to reading performance. In addition, several limitations of the present study (described in more detail in a subsequent section) may have impacted the relation between the UNIT-GAT and the reading achievement measures.

Predictive Validity

Logic suggests that children who score high on cognitive tests should also score high on academic measures following the implementation of an intervention. Several researchers have substantiated this assumption. For example, Stage, Abbott, Jenkins, and Berninger (2003) found that verbal IQ predicted growth on several measures of reading achievement for students who participated in a reading-based intervention. Although mean reading achievement change scores were significant among participants of the present study, a significant correlation was not found between cognitive scores and reading achievement pretest-posttest change scores. That is, unlike the results from the Abbot et al. study, UNIT-GAT Analogical Reasoning pretest scores generally did not significantly correlate with TOSWRF-2 and TOSCRF-2 change scores. There may be multiple explanations for this finding.

One explanation for the lack of predictive validity of the UNIT-GAT under the conditions of this study is that the psychometric properties of the test are suspect. That is not the most likely explanation, however. In fact, the UNIT-GAT demonstrates promising psychometric properties, even though standardization is not complete. For example, initial reliability estimates obtained from the publishing company are relatively high for the UNIT-GAT and for the AR subtest in particular (reliability estimates range from .91 to .99), as reported in the Methods section; and, reliability sets the upper bound estimate of validity. That is, a validity estimate cannot be higher than the reliability estimate squared. In addition, the pattern of increasing UNIT-GAT scores from this sample as a function of increasing chronological age reveals that it is sensitive to cognitive growth. These values are relatively impressive given that they are obtained from only one UNIT-GAT subtest. Therefore, the AR subtest, which is the focus of this study, appears relatively robust. Consequently, the lack of predictive validity may result from other influences (e.g., sample characteristics), as discussed below.

It is possible that the UNIT-GAT's lack of predictive capacity for this sample is attributable to an artificial increase in TOSWRF-2 and TOSCRF-2 scores from pretest to posttest. That is, scores may have increased from pretest to posttest as a result of variables unrelated to either the participants' cognitive abilities or the effectiveness of the two reading interventions. For example, increases in TOSWRF-2 and TOSCRF-2 scores from pretest to posttest may have occurred due to practice effects. Robust test-retest reliability coefficients obtained from the TOSWRF-2 (Mather, Hammil, Allen, & Roberts, 2014) and TOSCRF-2 (Mather, Hammil, Allen, & Roberts, 2014) manuals, however, suggest practice effects are unlikely (see Methods section). Increases in TOSWRF-2 and TOSCRF-2 scores may also have been observed due to regression to the mean, a statistical phenomenon that suggests that a variable that is extreme when first measured will be less extreme when measured again (Healy & Goldstein, 1978). Participants in the present study earned below average TOSWRF-2 and TOSCRF-2 pretest score; thus, regression to the mean suggests that their posttest scores should be closer to the mean.

Perhaps the lack of UNIT-GAT predictive capacity for this sample is related to sample size and length of intervention. Although the means increased as a function of intervention under both conditions, reading tutoring and books, attendance was somewhat sporadic, as shown in Table 7. Of the 16 coefficients reported in this table, four were obtained for 6 or fewer participants. So, limited sample size may have reduced the sensitivity of the scores. Related to this point is the observation that the limited number of intervention hours may have been insufficient to produce more robust findings. Even children who attended every tutoring session only received approximately 40 hours of instruction; and, research suggests more hours of tutoring may be necessary. For example, in his analysis of the literature, Torgesen (2004) found that the reading achievement of children between the ages of 9 and 12 only improved after 50-100 hours of "relatively intense" tutoring. Similarly, children in the books condition may not have read the books they were given. No participant returned all of their reading logs; therefore, there is no way to know with certainty whether participants read the books they received. These arguments seem less than compelling given that all pretest to posttest means increased significantly. However, these comparisons are global and predictive correlations are reported by group. In any case, given current results, it may be that the UNIT-GAT scores are too insensitive to predict effectively change scores.

Possible Outliers

Visual analysis of the means and standard deviations suggests that sixth and seventh grade participants may be "outliers" in comparison with the rest of the sample. Consequently, the data were reexamined without use of the five sixth and seventh grade participants' scores. The pattern of coefficients did not change appreciably when these students were removed from the dataset. All four of the coefficients between the UNIT-GAT AR and the TOSWRF-2 and the TOSCRF-2 decreased slightly. Similarly, two coefficients between UNIT-GAT AR scores and TOSWRF-2 and the TOSCRF-2 change scores decreased slightly. One coefficient between UNIT-GAT AR scores and TOSWRF-2 and the TOSCRF-2 change scores remained unchanged, and the final coefficient increased slightly. Based on these results, it can be concluded that the sixth and seventh grade participants may not be outliers, and it is appropriate to include their scores in analyses.

Limitations and Direction for Future Research

One possible strength of the present study is its population (and sample) of focus. Most participants were African American children from a low socioeconomic background, a historically underserved and under-researched group. However, in some important ways, the sample is also a limitation. For example, the sample is homogeneous, which limits generalizability. All participants were children from the Southeastern United States. In addition, most of the participants were minorities from a low socioeconomic background. Future researchers should obtain participants from across the country from a variety of backgrounds. Future researchers should also extend the results of the present study by administering the UNIT-GAT to other age ranges. In addition, future researchers should determine the concurrent and predictive validity of the UNIT-GAT by comparing participants' scores on the assessment with measures of reading achievement other than the TOSWRF-2 and the TOSCRF-2.

The relatively short period of time between pretest and posttest administration was a limitation of the present study as well. According to the TOSWRF-2 (Mather, Hammil, Allen, & Roberts, 2014) and TOSCRF-2 (Mather, Hammil, Allen, & Roberts, 2014) manuals, the assessments are meant to be given no more than four times a year; however, this sample completed posttest assessments just four weeks after completing pretest assessments. Future researchers should choose a reading measure that is more sensitive to change within a relatively short period of time.

Although not the focus of study, UNIT-GAT Quantitative Reasoning scores strongly correlate with TOSWRF-2 and TOSCRF-2 pretest scores, based on Cohen’s criteria (1988; for example, correlations ranged from .60 to .73). This finding is surprising because the rational link between UNIT-GAT AR and reading is stronger than the link between UNIT-GAT QR and reading. One explanation for this result is the order in which the subtests were administered. The UNIT-GAT AR subtest was completed before the UNIT-GAT QR subtest and, therefore, may have been more subject to error; because the response format for both subtests is similar, participants may have become more familiar with the assessment by the time they reached the QR portion of the assessment. It is also possible that UNIT-GAT QR scores and reading scores are related because success on the reading measures relies on swift word recognition and processing speed has been found to influence math skills and fluid reasoning, a cognitive skill measured by the UNIT-GAT QR subtest (Ferrer, O'Hare, & Bunge, 2009; McGrew & Wendling, 2010). Future researchers should further examine the relation between the UNIT-GAT Quantitative Reasoning and reading achievement.

Another limitation of the present study is its lack of a control group. Consequently, there is no way to determine whether the pre- to posttest increases in reading scores resulted from the two interventions or some other influence (e.g., maturation, history). Additionally, most children in the books condition did not return their reading logs; thus, their participation in the books condition is unknown.

Conclusions

In general, the Universal Nonverbal Intelligence Test – Group Ability Test appears to have acceptable concurrent validity. However, the UNIT-GAT AR subtest may not be sensitive enough to predict academic achievement change scores. Because of the limitations of the study, future research should be conducted to further explore the concurrent and predictive validity of the UNIT-GAT in general and the UNIT-GAT AR subtest specifically.

References

Allington, R. L., & McGill-Franzen, A. (2013). Eliminating summer reading setback: How we can close the rich/poor achievement gap. *Reading Today, 30*(5), 10.

Allington, R.L., McGill-Franzen, A., Camilli, G., Williams, L., Graff, J., Zeig, J., & Nowak, R. (2010). Addressing summer reading setback among economically disadvantaged elementary students. *Reading Psychology, 31*, 411-427.

Abbott, M. L., & McQuarrie, L. M. (2015). Language learners: An investigation of the impact of verbal load on PASS procedures. In T.C. Papadopoulos, R. K. Parrila, & J. R. Kirby (Eds.), *Cognition, Intelligence, and Achievement: A tribute to J. P. Das* (291-308). San Diego, CA: Elsevier.

Amdurer, E., Boyatzis, R. E., Saatcioglu, A., Smith, M. L., & Taylor, S. N. (2014). Long term impact of emotional, social and cognitive intelligence competencies and GMAT on career and life satisfaction and career success. *Frontiers in Psychology, 5*, 1447.

Armstrong, M. M. (2012). The relationship between intelligence, achievement, and psychological diagnosis: Applying the CHC theory to clinical diagnosis (Doctoral dissertation). Retrieved from http://search.proquest.com/psycinfo/docview/883991790/fulltextPDF/D6EEAD9A0D1D46DBPQ/1?accountid=14766

Bell, N. L., McConnell, J. E., Lassiter, K. S., & Matthews, T. D. (2013). The validity of the universal nonverbal intelligence test with the woodcock-johnson III tests of achievement. *North American Journal of Psychology, 15*(2), 243-256.

Bliss, S. L. (2008). Concurrent and Predictive Validity of the Universal Nonverbal

Intelligence Test-Group Ability Test (Doctoral dissertation). Retrieved from http://trace.tennessee.edu/utk_graddiss/412

Bracken, B. A. & McCallum, R. S. (1998). *The Universal Nonverbal Intelligence Test.* Chicago: Riverside.

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

Center for American Progress. (2014). *Testing overload in America's schools.* Washington, DC: Lazarin.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New Jersey: Lawrence Erlbaum.

Common Core State Standards Initiative. (2010). *Frequently ask questions*. Retrieved from http://www.corestandards.org/about-the-standards/frequently-asked-questions/

England, C. T. (2003). [Review of the Naglieri Nonverbal Ability Test-Individual Administration]. In R. A. Spies & B. S. Plake (Eds.), *The sixteenth mental measurements yearbook*. Retrieved from Mental Measurements Yearbook database.

Faggella-Luby, M., & Wardwell, M. (2011). RTI in a middle school: Findings and practical implications of a tier 2 reading comprehension study. *Learning Disability Quarterly*, *34*(1), 35-49.

Ferrer, E., O'Hare, E. D., & Bunge, S. A. (2009). Fluid reasoning and the developing brain. *Frontiers in Neuroscience, 3*(1), 46-51.

Fuchs, D. & Young, C. L. (2006). On the irrelevance of intelligence in predicting responsiveness to reading instruction. *Exceptional Children, 73(1),*8-30.

Goldstein, S. (2013). The science of intelligence testing: Commentary on the evolving nature of interpretations of the wechsler scales. *Journal of Psychoeducational Assessment, 31*(2), 132-137.

Gottfredson, L., & Saklofske, D. H. (2009). Intelligence: Foundations and issues in assessment. *Canadian Psychology/Psychologie Canadienne, 50*(3), 183-195.

Gutkin, T. B., & Reynolds, C. R. (2009). *The Handbook of School Psychology, Fourth Edition.* Hoboken, NJ: Wiley.

Hammill, D. D., Pearson, N., & Wiederholt, J. L. (2009). *Comprehensive Test of Nonverbal Intelligence, Second Edition.* Austin, TX: PRO-ED.

Hammill, D.D., Wiederholt, J.L., & Allen, E.A. (2014). *TOSCRF: Test of Silent Contextual Reading Fluency—Second Edition Examiner's Manual.* Austin, TX: PRO-ED.

Harris, P., Smith, B. M., & Harris, J. (2011). *The myths of standardized tests: Why they don't tell you what you think they do.* Lanham, MD: Rowman & Littlefield Publishers, Inc.

Healy, M. J. R., & Goldstein, H. (1978). Regression to the mean. *Annals of Human Biology, 5*(3), 277-280.

Jensen, A. R. (1980). *Bias in mental testing*. New York, NY: The Free Press.

Kaufman, A. S., & Kaufman, N. L. (2004). *Kaufman Assessment Battery for Children, Second Edition.* San Antonio, TX: Pearson Education.

Larsen, S., Hammill, D. D., & Moats, L. (2013). *Test of Written Spelling* (5th ed.). Austin, TX: PRO-ED.

Lee, V. E. (2002). *Inequality at the starting gate: Social background differences in achievement as children begin school.* Washington, DC: Economic Policy Inst.

Lohman, D. F. (2011). *Cognitive Abilities Test*, *Form 7*. Rolling Meadows, IL: Riverside Publishing.

Massa, I., & Rivera, V. (2009). Test review of Wechsler nonverbal scale of ability. *Journal of Psychoeducational Assessment, 27*(5), 426-432.

Mather, N. Hammil, D.D., Allen, E.A. & Roberts, R. (2014). *Test of Silent Word Reading Fluency—Second Edition Examiner's Manual.* Austin, TX: PRO-ED.

Mather, N., & Jaffe, L. E. (2011). Comprehensive Evaluations: Case Reports for Psychologists, Diagnosticians, and Special Educators. New York, NY: John Wiley.

McBee, G., & Duke, R. L. (1960). Relationship between intelligence, scholastic motivation, and academic achievement. *Psychological Reports, 6*, 3-8.

McCallum, R. S., & Bracken, B. (2016). *Universal Nonverbal Intelligence Test -- Second Edition Examiner's Manual.* Austin, TX: PRO-ED.

McCallum, R. S., & Bracken, B. (in press). *Universal Nonverbal Intelligence Test -- Group Ability Test.* Austin, TX: PRO-ED.

McGrew, K. S., & Wendling, B. J. (2010). Cattell-horn-carroll cognitive-achievement relations: What we have learned from the past 20 years of research. *Psychology in the Schools, 47*(7), 651-675.

Mitchell, R. E. (2005). How many deaf people are there in the United States: Estimates from the survey of income and program participation. *Journal of Deaf Studies and Deaf Education, 11(1)*, 112-119.

Naglieri, J. A. (1996). *Naglieri Nonverbal Ability Test.* San Antonio, TX: Harcourt Brace Educational Measurement.

Naglieri, J. A. & Bardos, A. N. (1997). *General Ability Measure for Adults*. Minneapolis, MN: National Computer Systems.

Naglieri, J. A., & Otero, T. M. (1997). The Wechsler nonverbal scale of ability: Assessment of diverse populations. In D. P. Flanagan, J. L. Genshaft & P. L. Harrison (Eds.). Contemporary intellectual assessment: Theories, tests, and issues (pp. 436-455). New York, NY: The Guilford Press.

Naglieri, J. A., & Ronning, M. E. (2000a). Comparison of white, African American, Hispanic, and Asian children on the Naglieri nonverbal ability test. *Psychological Assessment, 12*(3), 328-334.

Naglieri, J. A., & Ronning, M. E. (2000b). The relationship between general ability using the Naglieri nonverbal ability test (NNAT) and Stanford achievement test (SAT) reading achievement. *Journal of Psychoeducational Assessment, 18*(3), 230-239.

National Association of School Psychologists. (2010). *Principles for Professional Ethics.* Chicago, IL: Author.

Newcomer, P. (2014). *Diagnostic Achievement Battery* (4th ed.). Austin, TX: PRO-ED.

Newton, J. H., & McGrew, K. S. (2010). Introduction to the special issue: Current research in Cattell-Horn-Carroll-based assessment. *Psychology in the Schools, 47*(7), 621-634.

Reesman, J. H., Day, L. A., Szymanski, C. A., Hughes-Wheatland, R., Witkin, G. A., Kalback, S. R., & Brice, P. J. (2014). Review of intellectual assessment measures for children who are deaf or hard of hearing. *Rehabilitation Psychology, 59*(1), 99-106.

Riverside Publishing. (n.d.). UNIT Group Abilities Test National Standardization. Retrieved February 2, 2015, from http://www.riverpub.com/research/pdf/Website-UNITGAT6-29.pdf

Roid, G. H. (2003). *Stanford-Binet Intelligence Scales, Fifth Edition.* Rolling Meadows, IL: Riverside.

Roid, G. H., & Miller, L. J. (2013). *Leiter International Performance Scale, Third Edition.* Torrance, CA: WPS

Schrank, F. A., Mather, N., & McGrew, K.S. (2014). *Woodcock-Johnson IV Tests of Achievement.* Rolling Meadows, IL: Riverside.

Siah, P. C., & Kwok, W. L. (2010). The value of reading and the effectiveness of sustained silent reading. *The Clearing House*, *83*(5), 168-174.

Stage, S. A., Abbott, R. D., Jenkins, J. R., & Berninger, V. W. (2003). Predicting response to early reading intervention from verbal IQ, reading-related language abilities, attention ratings, and verbal IQ-reading discrepancy: Failure to validate discrepancy method. *Journal of Learning Disabilities, 36(1)*, 24-33.

Torgesen, J. K. (2004). Avoiding the devastating downward spiral: The evidence that early intervention prevents reading failure. *American Educator.* Retrieved from http://www.aft.org/periodical/american-educator/fall-2004/avoiding-devastating-downward-spiral

United States Census Bureau (2013). Language use in the United States: 2011. Washington, DC: Ryan.

Warne, R. T. (2015). Review of cognitive abilities test, form 7 (CogAT7). *Journal of Psychoeducational Assessment, 33*(2), 188-192.

Wechsler, D. (1992). *Wechsler Individual Achievement Test.* San Antonio: The Psychological Corporation.

Wechsler, D. (2001). *Wechsler Individual Achievement Test–Second Edition*. San Antonio, TX: Psychological Corporation.

Wechsler, D., & Naglieri, J.A. (2006). *Wechsler Nonverbal Scale of Ability.* San Antonio, TX: Harcourt Assessment

Whiston, S. C. (2013). Principles and applications of assessment in counseling (4th ed.). Belmont, CA: Brooks/Cole.

Wills, A. J. (2012). A study of the relationship between kindergarten nonverbal ability and third-grade reading achievement (Doctoral dissertation). Retrieved from http://gradworks.umi.com/35/05/3505945.html

Woodcock, R.W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III Tests of Achievement.* Itasca: Riverside Publishing.

Yoakum, C. S., & Yerkes, R. M. (1920). *Army mental tests.* New York, NY: Holt.

Appendix

Table 1

Descriptive Statistics of UNIT-GAT, TOSWRF-2, and TOSCRF-2 for the Total Sample

Measure	Mean Raw Score	SD (Raw Score)	Mean Standard Score/ Standard Deviation	Skewness/ Kurtosis
UNIT-GAT Analogical Reasoning (*n* = 133)	27.76	8.12	--	-.10/ .11
TOSWRF-2, Form A (*n* = 139)	71.08	28.16	91.35/ 13.37	-.48/ -.03
TOSCRF-2, Form A (*n* = 139)	50.60	28.01	85.40/ 15.56	.01/ -.20
TOSWRF-2, Form B (*n* = 52)	66.98	32.02	90.31/ 13..88	-.39/ .07
TOSCRF-2, Form B (*n* = 82)	48.50	26.01	84.33/ 16.22	-.161/ -.43
TOSWRF-2, Form C (*n* = 83)	80.95	27.34	95.23/ 13.30	-.24/ .18
TOSCRF-2, Form C (*n* = 84)	60.42	29.06	90.55/ 14.93	-.09/ .58

Note. Forms A and B of the TOSWRF-2 and TOSCRF-2 were given at pretest; Form C of the TOSWRF-2 and TOSCRF-2 was given at posttest.

Table 2

Descriptive Statistics of UNIT-GAT, TOSWRF-2, and TOSCRF-2 for the Tutoring and Books Conditions

	Tutoring Condition		Books Condition	
Measure	Mean Raw Score	SD	Mean Raw Score	SD
UNIT-GAT Analogical Reasoning	28.85 (*n* = 78)	8.12	26.22 (*n* = 55)	7.94
TOSWRF-2, Form A	78.36 (*n* = 80)	24.69	61.20 (*n* = 59)	29.73
TOSCRF-2, Form A	55.77 (*n* = 81)	28.11	43.36 (*n* = 58)	26.44
TOSWRF-2, Form B	85.95 (*n* = 19)	22.40	56.06 (*n* = 33)	31.87
TOSCRF-2, Form B	48.71 (*n* = 56)	24.74	48.04 (*n* = 26)	29.08
TOSWRF-2, Form C	86.85 (*n* = 34)	20.20	76.86 (*n* = 49)	30.90
TOSCRF-2, Form C	66.46 (*n* = 35)	26.06	56.10 (*n* = 49)	30.55

Note. Forms A and B of the TOSWRF-2 and TOSCRF-2 were given at pretest; Form C of the TOSWRF-2 and TOSCRF-2 was given at posttest.

Table 3

Means and Standard Deviations of UNIT-GAT Analogical Reasoning Raw Scores for Each Grade Level

	Grade						
	First (n = 9)	Second (n = 38)	Third (n = 31)	Fourth (n = 36)	Fifth (n = 11)	Sixth (n = 3)	Seventh (n = 2)
Mean	21.78	25.55	26.97	31.25	29.09	33.33	34.00
Standard Deviation	7.89	8.08	9.12	6.29	6.89	6.81	11.31

Table 4

Means/Standard Deviations of TOSWRF-2 and TOSCRF-2 Standard Scores for Each Grade Level.

	Grade						
	First	Second	Third	Fourth	Fifth	Sixth	Seventh
TOSWRF-2, Form A	93.00 /11.26 (*n* = 9)	91.95 /15.46 (*n* = 40)	91.74 /12.50 (*n* = 31)	93.05 /12.91 (*n* = 38)	85.08 /13.57 (*n* = 12)	93.33 /1.53 (*n* = 3)	77.50 /12.02 (*n* = 2)
TOSCRF-2, Form A	86.11 /8.43 (*n* = 9)	80.93 /18.22 (*n* = 40)	87.19 /14.35 (*n* = 31)	88.18 /14.52 (*n* = 38)	84.67 /16.59 (*n* = 12)	101.00 /16.52 (*n* = 3)	78.00 /8.49 (*n* = 2)
TOSWRF-2, Form B	94.14 /7.79 (*n* = 7)	81.33 /15.55 (*n* = 6)	90.79 /14.93 (*n* = 14)	92.25 /15.02 (*n* = 12)	91.71 /17.39 (*n* = 7)	97.00 /2.83 (*n* = 2)	74.00 /-- (*n* = 1)
TOSCRF-2, Form B	85.50 /12.02 (*n* = 2)	81.88 /18.59 (*n* = 32)	84.83 /14.26 (*n* = 18)	90.22 /15.50 (*n* = 22)	74.40 /10.31 (*n* = 5)	80.00 /-- (*n* = 1)	82.00 /-- (*n* = 0)
TOSWRF-2, Form C	92.25 /16.97 (*n* = 8)	96.59 /12.21 (*n* = 22)	94.38 /14.64 (*n* = 21)	96.08 /13.76 (*n* = 24)	97.00 /7.83 (*n* = 4)	98.00 /8.49 (*n* = 2)	-- /-- (*n* = 0)
TOSCRF-2, Form C	87.38 /12.92 (*n* = 8)	94.73 /14.08 (*n* = 22)	88.95 /14.92 (*n* = 21)	92.92 /15.01 (*n* = 24)	79.00 /19.72 (*n* = 4)	87.50 /4.95 (*n* = 2)	-- /-- (*n* = 0)

Note. Forms A and B of the TOSWRF-2 and TOSCRF-2 were given at pretest; Form C of the TOSWRF-2 and TOSCRF-2 was given at posttest.

Table 5

Corrected and Uncorrected Correlation Coefficients

	UNIT-GAT	
	Uncorrected	Corrected
TOSWRF-2, Form A	.45	.48
TOSWRF-2, Form B	.41	.36
TOSCRF-2, Form A	.45	.47
TOSCRF-2, Form B	.48	.50

Note. Both Form A and Form B of the TOSWRF-2 and TOSCRF-2 were given at pretest.

Table 6

Correlations Between UNIT-GAT Analogical Reasoning Pretest Scores and TOSWRF-2 and TOSCRF-2 Change Scores

	Total Sample	Tutoring Condition	Books Condition
TOSWRF-2 Change Scores from Pretest Form A to Posttest Form C	.04 (81)	-.17 (32)	.20 (49)
TOSWRF-2 Change Scores from Pretest Form B to Posttest Form C	-.01 (32)	-.32 (6)	.03 (26)
TOSCRF-2 Change Scores from Pretest Form A to Posttest Form C	-.04 (83)	-.17 (34)	.01 (49)
TOSCRF-2 Change Scores from Pretest Form B to Posttest Form C	-.07 (46)	-.06 (23)	-.10 (23)

Note. No Correlation Coefficient is statistically significant; values in parentheses are sample sizes

Table 7

Correlations Between UNIT-GAT Analogical Reasoning Pretest Scores and TOSWRF-2 and TOSCRF-2 Change Scores Among Students in the Tutoring Condition

	Total Sample	Attended 10+ Sessions	Attended 12+ Sessions	Attended 15+ Sessions
TOSWRF-2 Change Scores from Pretest Form A to Posttest Form C	-.17 (32)	.09 (19)	-.07 (16)	-.03 (12)
TOSWRF-2 Change Scores from Pretest Form B to Posttest Form C	-.32 (6)	-.15 (5)	.14 (3)	1.00* (2)
TOSCRF-2 Change Scores from Pretest Form A to Posttest Form C	-.17 (34)	-.02 (20)	-.13 (16)	-.02 (12)
TOSCRF-2 Change Scores from Pretest Form B to Posttest Form C	-.06 (23)	.25 (13)	-.27 (12)	.22 (9)

Note. * indicates $p < .001$; values within parentheses indicate sample size.

Table 8

Paired-Samples t*-test Results comparing TOSWRF-2 and TOSCRF-2 Pretest Scores and TOSWRF-2 and TOSCRF-2 Posttest Scores for the Total Sample*

		Mean	SD	T	df	P
Pair 1	TOSWRF-2 Form C -TOSWRF-2 Form A	9.71	13.63	6.451	81	.001
Pair 2	TOSWRF-2 Form C -TOSWRF-2 Form B	13.56	13.56	5.656	31	.001
Pair 3	TOSCRF-2 Form C- TOSCRF-2 Form A	10.36	18.36	5.169	83	.001
Pair 4	TOSCRF-2 Form C- TOSCRF-2 Form A	12.13	18.30	4.544	46	.001

Note. Forms A and B of the TOSWRF-2 and TOSCRF-2 were given at pretest; Form C of the TOSWRF-2 and TOSCRF-2 was given at posttest.

Table 9

Paired-Samples t*-test Results comparing TOSWRF-2 and TOSCRF-2 Pretest Scores and TOSWRF-2 and TOSCRF-2 Posttest Scores for the Tutoring Condition*

		Mean	SD	T	Df	P
Pair 1	TOSWRF-2 Form C -TOSWRF-2 Form A	7.33	13.39	3.145	32	.004
Pair 2	TOSWRF-2 Form C -TOSWRF-2 Form B	17.83	12.77	3.422	5	.019
Pair 3	TOSCRF-2 Form C- TOSCRF-2 Form A	12.49	17.76	4.159	34	.001
Pair 4	TOSCRF-2 Form C- TOSCRF-2 Form A	13.29	16.92	3.848	23	.001

Note. Forms A and B of the TOSWRF-2 and TOSCRF-2 were given at pretest; Form C of the TOSWRF-2 and TOSCRF-2 was given at posttest.

Table 10

Paired-Samples t-*test Results comparing TOSWRF-2 and TOSCRF-2 Pretest Scores and TOSWRF-2 and TOSCRF-2 Posttest Scores for the Books Condition*

		Mean	SD	T	df	P
Pair 1	TOSWRF-2 Form C -TOSWRF-2 Form A	11.31	13.68	5.784	48	.001
Pair 2	TOSWRF-2 Form C -TOSWRF-2 Form B	12.58	13.79	4.650	25	.001
Pair 3	TOSCRF-2 Form C- TOSCRF-2 Form A	8.84	18.82	3.287	48	.002
Pair 4	TOSCRF-2 Form C- TOSCRF-2 Form A	10.91	19.94	2.624	22	.015

Note. Forms A and B of the TOSWRF-2 and TOSCRF-2 were given at pretest; Form C of the TOSWRF-2 and TOSCRF-2 was given at posttest.

Vita

Brooke Browarnik is originally from Miami, Florida. She graduated from Connecticut College with a B.A. in Psychology in 2012. Brooke began the doctoral School Psychology program at the University of Tennessee-Knoxville in the fall of 2012. In December of 2014, she received an M.S. in Applied Educational Psychology. Brooke will receive her Ph.D. in August 2017 after completion of a year-long, pre-doctoral internship.