12-15-2014

An Analysis of the Cultural Invariance of a Visual-Motor Integration Measure

Emma Kate C. Wright
University of South Carolina - Columbia

AN ANALYSIS OF THE CULTURAL INVARIANCE OF A VISUAL-MOTOR INTEGRATION MEASURE

by

Emma Kate C. Wright

Bachelor of Science
College of Charleston, 2012

Submitted in Partial Fulfillment of the Requirements

For the Degree of Master of Arts in

School Psychology

College of Arts and Sciences

University of South Carolina

2014

Accepted by:

Scott L. Decker, Director of Thesis

E. Scott Huebner, Reader

Lacy Ford, Vice Provost and Dean of Graduate Studies

ACKNOWLEDGEMENTS

The author wishes to thank the University of South Carolina's school psychology faculty for the opportunity to participate in the scholarly community, Drs. Scott Decker and Scott Huebner for their guidance and support, and the graduate members of the Applied Cognitive Neuropsychology Laboratory at the University of South Carolina for their help in the completion of this study.

ABSTRACT

As diverse populations within schools increase, the need for culturally-sensitive assessment is essential; however, test of ability vary in their degree of influence from culture. No test is “culture free,” but the low-linguistic demands on test of visual-motor integration (VMI) make them appropriate for use with diverse populations. Variation in VMI test performance due to cultural factors has negative implications for test interpretation and use with diverse populations because of VMI’s significant association with school readiness, academic achievement, social-emotional functioning, and neuropsychological assessment. The current study explored the cultural invariance of the Bender Motor Gestalt Test, Second Edition (BG-II), a test of VMI, using Differential Item Functioning (DIF). Analyses were conducted using a subset of data from the normative sample of the BG-II, which included the BG-II’s copy phase items for 935 African-American, Hispanic, and Caucasian children ages 4 to 7 years. Overall, results indicated that the BG-II can be considered a culturally invariant measure, but caution should be used when interpreting item 3 of the copy phase, only for African-American 4-year-olds due to significant DIF. It is currently unclear why item 3 has significant DIF for African-American 4-year-olds, and continued research on the cultural invariance of the BG-II is needed to facilitate the development and use of culturally appropriate measures.

Table of Contents

List of Tables

CHAPTER 1

INTRODUCTION

School psychologists have been reported to spend approximately two-thirds of their time on assessment (Brown, Holcombe, Bolen, & Thomson, 2006), where they are responsible for identifying why a child is not meeting age- or grade-expectations for learning. When deciding children's eligibility for special education services or educational needs, one method that school psychologists frequently utilize is norm-referenced measures (Decker, 2008). Use of these tests for such high-stakes decision-making highlights the critical need for tests to accurately measure and represent an individual's ability, without undue influence from construct-irrelevant factors, such as cultural differences. In essence, tests of ability should be unbiased, illustrating cross-cultural invariance, where scores should not significantly differ across cultural groups (Flanagan, Ortiz, & Alfonso, 2013). When performance on a measure varies due to the influence of cultural factors, potential negative implications can occur with interpretation of test performance for diagnostic criteria and special education eligibility (Flanagan et al., 2013). Due to the substantial increase in diverse populations within the United States, the need for accurate assessment is critical to avoid measurement issues associated with cultural influence.

Increasing Diverse Populations

Interest in the potential effects of cultural differences on learning has increased in recent years due to the continued growth of diverse populations in America. The total population of the United States grew by 9.7% from 2000 to 2010, moving from 281.4 million individuals to 308.7 million individuals (U.S. Census Bureau, 2011a). Both African-American and Hispanic populations were reported to increase at faster rates than the overall population (U.S. Census Bureau, 2011a, 2011b, 2011c), while the Caucasian population was reported to grow more slowly (U.S. Census Bureau, 2011d). The 2010 U.S. Census reported that 40.2 million individuals identified themselves as African-American, which was a 12% increase from 2000 (U.S. Census Bureau, 2011b). The Hispanic population saw a dramatic 43% increase, with a total of 50.5 million individuals identified as Hispanic (U.S. Census Bureau, 2011c). This population growth was equal to roughly 15.2 million people (U.S. Census Bureau, 2011c). In comparison to African-American and Hispanic populations, the Caucasian population only had a 6% increase from 2000 to 2010, where a total of 223.6 million individuals were identified as Caucasian (U.S. Census Bureau, 2011d). The changes in ethnic distribution from the past decade are dramatic, and can be reflected within the United States' public school system.

Public school enrollment rates as a whole have increased from 44.7 million students to 49.5 million students from the fall of 2001 to 2011, but representations of African-American and Caucasian populations actually decreased. Specifically, African-American student enrollment decreased from 17% to 16% representation in the schools from 2001 to 2011. Caucasian student enrollment decreased at a higher rate, moving from 60% to 52% representation from 2001 to 2011. In contrast, the number of Hispanic students enrolled in public schools grew from 17 % to 24% from 2001 to 2011 (U.S.D.E., N.C.E.S., C.C.D, 2014), now representing almost a fourth of the total school population. The change in population distribution in public schools illustrates a significant need for educators to be sensitive to differences in cultural groups. The need for cultural sensitivity is especially true for school psychologists, where they should be cognizant of the potential influence of cultural factors on a measure prior to beginning an evaluation with a child who has a culturally or linguistically diverse background.

Assessing Diverse Children

With the increase in diverse populations within the United States being reflected in the schools, culturally sensitive assessment practices and appropriate interpretation of assessment with diverse populations is needed. When assessing children from diverse backgrounds, it is important to note several potential concerns. One potential concern involves lower performance on assessment measures for children with limited resources from lower SES groups. Research indicates higher cognitive outcomes in children that have increased access to resources when raised in high SES homes verses individuals in low SES homes (Bradley & Corwyn, 2002). Individuals with a low SES may be living with their families below the national poverty level, which is a measure of daily household income based on the number of individuals living in a home. Between the years of 2007 to 2011, a total of 25.8% of African-American individuals and 23.2% of Hispanic individuals were reported by the U.S. Census to be living below the national poverty level. The rates for African-American and Hispanic populations are significantly higher than the reported 11.6% of Caucasian individuals living below the national poverty level (U.S. Census Bureau, 2013b). Further, children from diverse backgrounds, specifically, African-American and Hispanic children, are the most overrepresented populations within special education services in the schools (National Center for

Learning Disabilities, 2014). This overrepresentation highlights the need for appropriate assessment of children from diverse groups.

Another major concern that arises when assessing children from culturally diverse backgrounds involves language. Language is an especially salient concern when assessing individuals who have non-native-English-speaking backgrounds or limited English proficiency (McCallum & Bracken, 1997). High linguistic demands can lead to systematic test biases that may lead to underestimation of a child's true abilities. A reported 35 million individuals 5 years and older living in the United States speak Spanish (U.S. Census, 2013a). Some school psychologists try to combat cultural factors influence on assessment by conducting bilingual assessments, but the use of another language creates additional concerns regarding the validity and reliability of an assessment measure to accurately portray an individual's ability. Even if the assessment measure itself was deemed reliable and valid for use with a bilingual evaluation, many school psychologists do not believe they have adequate training to administer bilingual assessments (Ochoa, Rivera, & Ford, 1997).

One alternative to bilingual evaluation is nonverbal assessment or assessment with low-linguistic demands. Tests of ability vary in their level of bias due to cultural influence, but those with a major verbal component (i.e., language) tend to be more culturally and linguistically demanding than those with fewer verbal components (Flanagan et al., 2013). This is because cultural diversity is often linked with linguistic diversity—especially in the case of the Hispanic school-aged population. A recent survey reported that 88% of all practitioners administered a nonverbal intelligence test when evaluating individuals who were considered culturally or linguistically diverse (Sotelo-

Dynega et al., 2011). Nonverbal tests are believed to increase validity because language has been removed from the test content (Brigham, 1923)—i.e., stimuli and responses do not require language processing. However, it is important to clarify that nonverbal tests do require some communication between the examinee and the examiner (i.e., for instructions), but when compared with traditional ability tests, the language demands are significantly reduced (Flanagan et al., 2013). Even with reduced language demands, it is possible that a test may contain culturally-loaded content (Flanagan et al., 2013). However, at this point in time, it is assumed that test with low-linguistic demands are less culturally biased than those with a heavy verbal component.

To address cultural influence with assessment, Ortiz (2004) highlights several pre-assessment recommendations for nondiscriminatory assessment. Ortiz's recommendations include considering: cultural and linguistic background, behavior or performance within the context of the learning environment, measuring performance and academic achievement through informal and direct methods, considering potential bias in the use of standardized assessment, and joining forces across disciplines when making decisions for a child. By considering Ortiz's recommendations before starting the evaluation process with a child, it is possible to reduce bias, particularly cultural bias (e.g., variance across test performance) in assessment (Ortiz, 2004). Take note that there are no suggestions for "culture-free" assessment from Ortiz, but rather recommendations for reducing cultural bias, once again illustrating the predicament that school psychologists face when assessing children from assorted cultural backgrounds.

Establishing how much culture influences performance on standardized measures is complex (Ostrosky-Solis, Ramirez, & Ardila, 2004). At this point in time, psychometric research has not been able to eliminate bias from testing. As Flanagan, Ortiz, and Alfonso (2013) have stated, "tests will always reflect specific values, utilize culture-specific content to one extent or another, and expect possessions of age- or grade- appropriate development in their content, design, and structure" (p.293). With no "culture-free" tests, school psychologists are left with the dilemma of determining how to appropriately assess children from diverse backgrounds. Knowing that no measure is without limitations, and that low-linguistically demanding test are more frequently used for assessing diverse populations, it is essential that researchers and practitioners are aware of potential biases within a measure from a psychometric standpoint.

Evaluating Bias in Measurement

When developing a measure, the goal is to create a valid and reliable test that appropriately examines a construct of interest. Test development requires careful planning and attention to the construct of interest so that comparisons of performance can be made across individuals, regardless of differences amongst individuals. However, no measure is without limitations, and awareness regarding a measure's specific weaknesses should be noted. Unfortunately, some limitations of a measure are not as clearly identifiable as others, including the issue of bias. One major form of bias within measurement is that of construct bias. When the construct a particular measure aims to evaluate results in dissimilar meanings for two groups being tested, the measure has construct bias. Construct bias results in difficulties with the interpretation and meaning of test scores for a population. When present, a measure would not provide a meaningful representation of ability on the intended construct of interest because comparison could not be made between different groups of individuals (Furr & Bacharach, 2014). One such difference between groups could be culture-linguistic factors.

Determining whether or not bias occurs within a measure based on cultural differences is a complex process. Consider some of the following culturally-loaded components of an assessment: administration format (e.g., paper and pencil vs. performance test), instruction format (e.g., printed vs. oral), examinee response format (e.g., written or oral), cultural loading of test items (e.g, written words vs. pictorial representations), specific knowledge for problem solving (e.g., pulling from crystallized knowledge vs. novel problem-solving), and the amount of language required to complete the test (e.g., high verbal loading vs. non-verbal or pantomime; Jensen, 1980), School psychologists should be aware of these potential sources of bias with assessment.

As previous discussed, one example of bias based on cultural factors is that of language. Ortiz and Lella (2005) indicated that bias based on cultural-linguistic factors has typically been described based on psychometric properties of an assessment. These properties include the reliability and validity of a measure, and are related to the processes involved in test development. Arthur Jensen (1980) highlighted the basic steps required for test construction. First, when creating a measure, items are constructed to assess ability of individuals on a specific construct of interest. Once items are created, items are administered to a pool of relevant individuals that test developers are seeking to measure performance on for the construct of interest. After initial data is collected from the items, the items are analyzed for item difficulty, item discrimination, and error. Item analysis will indicate biased items, which will allow test developers to remove those items from the final pool of items to be used on the measure. Once the items have been selected for the measures, the last step is to standardize the measure, by administering the created measure to a large, representative sample. This normative sample allows comparisons to be made between an individual's performance on the construct of interest and the general population (Jensen, 1980). Within this brief synopsis of test developed, they key steps where bias is presented an opportunity to harm psychometric properties (e.g. validity of a measure) occurs within the item selection and analysis stage.

Specifically, appropriate item selection helps ensure construct validity, where our items are accurately assessing the construct of interest for a test. If test performance varies on the construct of interest across cultural groups the measure may have cultural bias within the test items, where items favor one cultural group over another. One method to assess for bias involves the use of factor analysis. Factor analysis, looks at the correlation between items, where items with similar correlations are grouped together into a cluster or a "factor," indicating that these items are likely measuring the same construct. Factor Analyses can result in multiple clusters of items, indicating that the measure is multidimensional (i.e., measuring multiple constructs), or items can all have similar correlations where there is a single cluster of items, indicating that the measure is unidimensional (i.e., measuring a single construct). To compare performance across cultural groups using factor analysis, item correlations need to be conducted separately for the different groups of interests (Furr & Bacharach, 2014). If, for example, we were assessing cultural differences between Caucasian and African-American individual's performance on a measure, we would need to examine the factor structure for both of these groups separately. If we found all items were similarly correlated for Caucasian individuals (i.e., one factor), but African-American individual's performance on items results in two separate clusters of items (i.e., two factors), we could conclude that the measure likely has construct bias, where items are measuring different constructs based on an individual's ethnicity (i.e., cultural bias).

An alternate method to address cultural bias from a psychometric standpoint is to investigate item difficulty for different cultural-linguistic groups (Ortiz & Lella, 2005). Differential Item Functioning (DIF) using an Item Response Theory (IRT) model is one way to assess item difficulty (Stark, Chernyshenko, Drasgow, 2006). IRT models take into account item responses from examinees while factor analysis examines the covariance between items (Reise, Widaman, & Pugh, 1993), where IRT models are superior to factor analysis methods because items can be analyzed at the individual level as opposed to clusters of items. More specifically, IRT models investigate the probability of responding to an item while taking into account the individual's ability level on a construct of interest (Lord, 1980). A specific assumption of IRT models is that from test data it is possible to estimate an individual's true score, or true ability level for the construct of interest based on their responses to test items. Using this assumption, if individuals' responses from a test are known along with estimates of true scores for two groups of people, item bias can be demonstrate when true scores do not match item responses (Furr & Bacharach, 2014). Examining DIF follows this logic, where DIF involves calculating difficulty levels for a set of assessment items and comparing the average difficulty level of a particular item for one group versus another (Tennant et al., 2004). Understanding performance at the item level allows insight into performance differences between groups. DIF occurs when there is an interaction between persons and items, such that a particular item has a significantly higher (or lower) item difficulty value for one group of participants' scores than for the other group, on average, after taking into account participants' overall scores (Linacre, 2005).

DIF can be used to examine invariance across cultural-linguistic groups, where for example, if difficulty for a particular item was higher for Hispanic individuals than for Caucasian individuals, this would represent cultural variance and systematic bias in the test item, rather than a true difference in ability. If several items on the same test followed the same pattern, the test could be said to be culturally variant across Hispanic and Caucasian groups, implying a systematic bias that may lead to underestimation of Hispanic individuals' performance. If, however, item difficulty was similar for Hispanic and Caucasian groups for most items, the test's cultural invariance would be supported for those cultural-linguistic groups. The implications of finding that a measure has cultural variance through significant DIF depend on the specific construct being examined by the test. Specifically, it is important to consider how performance on a construct is being interrupted, applied, and used from a measure. One significant construct of interest, with widespread implications if measurement is culturally biased, is that of visual-motor integration.

Significance of Visual-Motor Integration

According to the American Psychological Association's dictionary of psychology (2007), visual-motor coordination is "the ability to synchronize visual information with the movements of different parts of the body" (p.986). Integrating both visual and motor skills allows the completion of variety of tasks. For example, in children within the schools, visual-motor integration plays a role in successful academic achievement where individuals must be able to coordinate their body's motor movements and visual information to complete tasks like writing with pencil and paper or typing on a computer. Visual-motor integration has been extensively studied in the literature, and has specifically been associated with school readiness (Bart, Hajami, & Bar-Heim, 2007; Carlton & Winsler, 1999), academic functioning (especially in the areas of reading and writing; Bart et al., 2007; Kulp, 1999), socio-emotional and/or behavioral difficulties (Cummins, Piek, & Dyck, 2005; Kurdek & Sinclair, 2000), and neuropsychological functioning (Sutton, et al., 2011; Williams, Griebel, & Dykman, 1998; Dawson & Watling, 2000).

In the area of school readiness, consider specific tasks that are required for success in school like cutting and coloring in the early years of education, followed by writing. To be successful with these types of tasks, individuals must be able to coordinate fine-motor movements with visual input. According to McHale and Cermak (1992) the majority of the school day is devoted to fine motor activities, which highlights the critical need for visual-motor integration skills when considering school readiness. In a study conducted by Bart, Hajami, and Bar-Haim (2007), seventy-one kindergarten students were assessed at two time periods, (1) when first transitioning to formal education and (2) one year later during first grade, on basic motor skills, as well as academic, social, and emotional functioning. Results indicated that visual-motor integration was significantly related to academic achievement, adaptation to education in the school, as well as social and emotional adjustment to school. Specifically, children with lower visual-motor integration skills were found to have higher rates of teacher reported negative behavior and anxious-withdrawn behavior, while children with higher visual-motor integration skills were found to engage in more pro-social behaviors in first grade (Bart et al., 2007). If visual-motor integration in children is associated with easier adaption to the educational environment, then we can expect to see associations with specific aspects of academic achievement.

Visual motor integration skills have been significantly correlated with children's reading, writing, and math achievement (Bart et al., 2007; Kulp, 1999; Chu, 1997). When reading, children learn to visually identify different, distinguishable letters (i.e., visual discrimination), which then become grouped into words. With the increase in knowledge of sight words, children are then able to begin to read clusters of words, which with semantic and syntactic knowledge form sentences. The task of reading requires visual discrimination between letters and sight words, which is similar to discriminating between numbers for math. Children with lower visual discrimination perform more poorly on reading and math tasks (Kulp, 1999). Additionally, more significant academic concerns are associated with writing and poor visual-motor integration skills. Writing is a complex skill that involves many components, including fine-motor coordination and knowledge of language. In order to write, children need to be able to grip a writing utensil, coordinate their mental thoughts to create a motor movement on paper that is legible and meaningful. Children with poor visual-motor integration skills struggle with writing legibly (Daly, Kelley, & Krauss, 2003), which may increase tasks avoidance due to anxiety (Bart et al., 2007). In addition to anxious/avoidant behavior, children with low visual-motor integration deficits have poor peer relationships and lower self-worth (Skinner & Piek, 2001). Children who have gross and fine motor deficits have difficulty playing games with peers, which can lower social acceptance from peers (Rose, Larkin, Berger, 1997), where some children have negative interactions with other children. Further, these negative interactions may cause children with visual-motor skill deficits to avoid social situations and develop negative self-wroth due to a lack of peer socialization (Skinner & Piek, 2001; Harter, 1987). Test of Visual-motor integration skills are also associated with neuropsychological assessment, which seeks to identify if there is damage to specific regions of the brain that are influencing behavior (Hebben, & Milberg, 2002). Identification of brain damage involves a comprehensive battery of multiple test instruments that are able to isolate specific brain functions to draw correlations between behavioral deficits (Lacks, 1999).

Test of visual-motor integration are appropriate for use with diverse populations due to their low-linguistic demands, where test content frequently involves geometric figures and limited verbal However, if performance on test of visual-motor integration is biased due to cultural factors, interpretation could have significant negative implications for individuals. One example of a potential negative effect of culturally biased visual-motor integration measures could be higher cost for assessment and treatment. Specifically, if a measure is biased and an individual is inaccurately diagnosed as having visual-motor impairment, more extensive assessment and treatment may occur, like expensive brain imaging studies or occupational therapy. Another example of a negative effect of culturally biased measures is that, in the schools, inaccurate assessment may perpetual the over-representation of cultural-linguistic groups receiving special education services. Visual-motor integration skills are clearly associated with many areas of functioning, which make appropriate, un-biased assessment measures essential. The Bender Visual-Motor Gestalt Test, Second Edition (BG-II) is a test of visual motor integration that could be considered one of the most widely used psychological measures (Archer & Newsom, 2000; Brannigan & Decker, 2003; Decker, Allen, & Choca, 2006; Piotrowski, 1995; Sullivan & Bowden, 1997), and it has not yet been assessed for cultural-bias.

Current Study

The current study seeks to examine the cultural invariance of the BG-II, as performance across cultural-linguistic groups has not yet been empirically tested. The BG-II has been found to be significantly correlated with both cognition (Woodcock, McGrew, & Mather, 2001; Psychological Corporation, 2001) and academic achievement (Wechsler, 1991; Psychological Corporation, 1997; Roid, 2003), and has been used within neuropsychological test batteries for the diagnosis of brain damage (Lacks, 1999; Lacks & Newport, 1980; Goldberg, 1959). The BG-II is an example of a nonverbal test that removes language demands from test content—stimuli and responses. Again, although no test is “culture-free,” the BG-II has low linguistic demands and may be considered ideal for use with those with diverse cultural backgrounds.

The BG-II is a norm-referenced measure that assesses visual-motor integration by asking children to draw (and later recall) a series of geometric designs that become progressively more complex (Brannigan & Decker, 2003). Items on the BG-II are presented on stimulus cards to examinees. Although the stimulus cards contain non-linguistic information (geometric designs) and require non-verbal responses (drawing geometric designs), it is possible that cultural-linguistic factors still influence performance. For example, communication is required when explaining the directions and describing the required tasks. The invariance of BG-II performance across cultural-linguistic groups has not yet been empirically tested.

In a study conducted by Decker and colleagues, performance on the BG-II was compared to performance on the Stanford-Binet Intelligence Scales, 5th edition (SB-V; Decker, Englund, Carboni, & Brooks, 2011; Roid, 2003), where the SB-V measures cognitive ability based on the Cattell-Horn-Carroll (CHC) model of intelligence (Carroll, 1993; Cattell, 1963). Results indicated that both the quantitative reasoning and fluid reasoning factors on the SB-V were significantly associated with motor performance on the Copy phase of the BG-II (Decker et al., 2011). The quantitative reasoning factor score on the SB-V represents an individual's ability to apply logical thinking and mathematical knowledge to arrive at a solution to a problem, while the fluid reasoning factor score represents an individual's ability to solve novel verbal and nonverbal problems (Roid, 2003). Additionally, Decker and colleagues found that nonverbal composites on the SB-V were more highly associated with performance on the BG-II than verbal factors (Decker et al., 2011), supporting the notion that the BG-II could be considered a non-verbal measure.

Other uses of the BG-II include identification of learning disorders, diagnosis of brain injury, and verifying anxious mannerisms (Tolor & Schulberg, 1963). Overall, when using the BG-II for assessments, information can be gained about an individual due to visual-motor integration being associated with different areas of functioning. As previously discussed, the BG-II's non-verbal nature may support the use of this assessment as a good measure for culturally diverse populations.

Historically, performance on visual-motor integration tasks was believed to vary across ethnicity and SES. Koppitz (1975) reported that when comparing African-American to Caucasian individuals performance on the Bender Gestalt Test (Bender, 1938), African-American children were delayed in their visual-motor skills, with more scoring errors than Caucasian children. Koppitz indicated that this delay in developing visual-motor skills was likely due to distinct cultural differences between African-American and Caucasian children. In a study conducted by Sattler and Gwynne, (1982), visual-motor performance differences existed between African-American and Caucasian children across the age range studied (ages 5- to 11-years old; Sattler & Gwynne, 1982). Additionally, individuals with lower socioeconomic status were found to make more errors on the Bender Gestalt Test than individuals from higher SES backgrounds (Hoffman, 1966), which is likely a reflection of increased access to resources in higher SES homes (Bradley & Corwyn, 2002). Due to historical biases on performance of visual-motor skills, including potential biases based on ethnicity and SES, performance across different cultural-linguistic groups should be assessed to enhance the known psychometric properties of the measure. The BG-II has been found to have strong reliability and validity (Brannigan & Decker, 2003); however, an analysis of the cultural invariance of the BG-II has not been previously conducted within the research literature.

The objective of this study is to evaluate the cultural invariance of BG-II performance across cultural-linguistic groups to rule out potential biases within this measure using DIF, which is a sophisticated methodology that is superior for assessing cultural invariance over other methods like factor analysis. It is hypothesized that the BG-II will be considered a culturally invariant measure due it’s reduced language demands and sound reliability and validity, with no significant DIF found across items. African-American and Hispanic populations were selected for this study because they are among the fastest growing populations in the U.S. (U.S. Census Bureau, 2012), and school psychologists are increasingly asked to assess individuals from diverse cultural backgrounds. Being aware of how widely used tests of ability like the BG-II are influenced by cultural-linguistic differences is critical for providing nondiscriminatory assessment in the schools.

CHAPTER 2

METHODS

Participants

Analyses were conducted from the Bender Visual-Motor Gestalt Test, Second Edition (BG-II) using a subset from the normative data sample, which includes 4,000 individuals, ages 4 to 85+ years. All individuals were administered the standardized full battery assessment of the BG-II, which provided scores for the 16 items on the Copy phase, as well as the 16 items on the Recall phase. A total of 747 variables were collected with the BG-II by original examiners, one of which provided demographic information regarding cultural group or ethnicity based on the 2000 U.S. Census ethnic group labels (i.e., Caucasian, Black/African American, Hispanic, Asian, and Other). This ethnic variable was used for the grouping variable in the current study.

Cases were selected for analyses when age was equal to or less than 7 years 11 months, and when the ethnicity variable for an individual was noted to be African-American, Hispanic, or Caucasian. Cases were excluded from analyses if (1) individuals were older than 7 years 11months; (2) the ethnic variable was not noted to be African-American, Hispanic, or Caucasian; and (3) the participant showed missing or inappropriate data for scores on any of the test items. From the original archival dataset, a total of 3,065 cases were excluded from analyses, resulting in a total sample size of N = 935 (n = 132 African-American individuals, n = 5,473 Hispanic individuals, and n = 665

Caucasian individuals) for the current study. More specific demographic information is displayed in Table 1.

Measures

The BG-II was created to measure visual-motor skills by using nine original designs from the Bender-Gestalt Test (Bender, 1938), in addition to seven new designs. Out of the new designs, four of the seven are used solely with individuals aged 4 to 7 years 11 months, while the remaining three new designs are used solely with individuals ages 8 to 85+ years. Test administration occurs in two stages, the Copy phase and the Recall phase. The Copy phase asks individuals to copy a series of designs onto a blank sheet of paper, while the Recall phase asks individuals to redraw the previously presented Copy phase designs from memory. Each phase consists of 16 items, scored on a scale from 0 (no resemblance to the stimulus card) to 4 (near-perfect resemblance). Item administration is dependent on age due to the developmental nature of visual-motor ability, where the level of difficulty increases with each subsequent item. Individuals aged 4 to 7 years 11 months are presented items 1 to 13, while individuals aged 8 years and older are presented items 5 to 16. Additional supplemental tests exist with the BG-II, but these tests are not a part of the standard battery. For the scope of this study, analyses only included data from the Copy phase of the standard battery for individuals aged 4 years to 7 years 11 months (i.e., items 1 through 13).

Interrater reliability was reported by the *Bender-Gestalt II Examiner's Manual* (Brannigan & Decker, 2003) to be .85 for the Copy phase and .92 for the Recall phase. The manual also reported that the BG-II has strong internal consistency with a split-half reliability coefficient of .91 (*SEM* = 4.55). Criterion validity for the Copy phase tests was reported by the manual via a correlation between the BG-II and the Beery-Buktenica Developmental Test of Visual-Motor Integration, Fourth Edition, Revised (VMI; Beery, 1997), where $r = .65$. Overall, the BG-II has been found to be a valid and reliable assessment instrument for measuring visual-motor ability.

Data Analyses

Item difficulty calibration. In order to investigate cultural invariance across diverse populations, differences in item difficulty across cultural-linguistic groups were examined; but first, a method of calibrating item difficulty was needed. Item difficulty calibration was conducted in WINSTEPS (Linacre, 2005) using the partial credit Rasch model. The Rasch model is a one-parameter Item Response Theory (IRT) model, wherein only item difficulty and person ability (and not guessing, etc.) are modeled to influence item scores. As in all IRT models, scores obtained on an assessment measure are used to compute the probability that a correct response will be provided for a particular item, based on a person's ability and the difficulty of the item (Nandakumar, Glutting, Oakland, & 1993). The premise behind Rasch modeling is that a test measures a single underlying dimension, and that items and persons can be arranged in order of difficulty and ability, respectively, on this dimension (Distefano & Morgan, 2010). The Rasch model calibrates item difficulty by converting ordinal-level data based on ranking items and persons into interval-level data using logarithmic transformation (Bond & Fox, 2001). Item difficulty and person ability are expressed in terms of logits, with a typical range of -2 to 2 logits (Bond & Fox, 2001). Higher logit values indicate more difficult items or persons with more ability on the measured dimension, while lower logit values, conversely indicate easier items or persons with less ability on the measured dimension. The partial credit Rasch model was used for the current study, because BG-II Copy phase items are scored on a scale of 0 to 4, rather than on a dichotomous (0 or 1) scale. The 13 BG-II Copy phase items were calibrated using scores from the study sample of $N = 935$ individuals.

Differential Item Functioning (DIF). After item difficulty values were calibrated, DIF analyses were conducted for the 13 BG-II Copy phase items using WINSTEPS software (Linacre, 2005). Again, DIF indicates that there is an interaction between persons and items, where on average, item difficulty significantly differs between groups impacting individuals' overall scores (Linacre, 2005). Two DIF comparison groups were created for analyses. The first comparison group ran DIF on BG-II Copy phase item performance between African-American and Caucasian children, while the second comparison group ran analyses between Hispanic and Caucasian children. In order to examine possible DIF across the cultural-linguistic groups, data for each BG-II Copy phase item was dummy coded (comparison group 1: African-American -1, Caucasian 1; comparison group 2: Hispanic -1 and Caucasian 1). Both comparison groups had DIF analyses run by age level (ages 4 through 7), which resulted in a total of 4 DIF analyses for each comparison group.

WINSTEPS uses the anchor theta method for calculating the magnitude of DIF across groups for each item, and the Mantel-Haenszel procedure for determining whether the difference in item difficulty across groups is statistically significant (Linacre, 2005). According to guidelines from previous research, DIF is indicated by a large magnitude in DIF contrast (> 0.50 logits) that is also statistically significant ($p < .003$, Bonferonni corrected for 13 comparisons; Lai, Cella, Chang, Bode, & Heinemann, 2003). Items that meet these criteria may be problematic, as DIF indicates that a factor beyond person ability or item difficulty (in this case, cultural-linguistic differences across groups) is influencing scores.

Because the BG-II is a nonverbal test and has significantly reduced linguistic demands, it was hypothesized that none of the 13 Copy phase items would demonstrate significant DIF across cultural-linguistic groups, suggesting the cultural invariance of the BG-II for African-American and Hispanic populations.

Table 2.1

Descriptive statistics (N =935).

Variable	[a]*M*	[a]*SD*	*n*	[b]*Percentage*
Gender				
Male	99.45	14.03	473	50.6
Female	103.13	13.16	462	49.4
Ethnicity				
African-American	99.31	12.65	132	14.1
Age 4	100.45	13.64	55	5.8
Age 5	103.64	9.34	25	2.7
Age 6	97.48	11.92	26	2.8
Age 7	94.77	13.06	26	2.8
Hispanic	102.39	12.04	138	14.8
Age 4	102.02	11.06	60	6.4
Age 5	102.28	11.42	25	2.7
Age 6	101.00	13.77	27	2.9
Age 7	104.58	12.97	26	2.8
Caucasian	101.47	14.21	665	71.1
Age 4	104.19	13.48	259	27.7
Age 5	100.70	14.26	138	14.8
Age 6	99.32	14.86	132	14.1
Age 7	99.81	14.14	136	14.5
Age				
4-year-olds	103.58	13.24	374	40.0
5-year-olds	101.69	13.33	188	20.1
6-year-olds	99.78	14.42	185	19.8
7-year-olds	100.11	14.09	188	20.1

[a] Mean and standard deviation values are based on standard scores for the copy phase of the BG-II.

[b] Percentage of the sample population is shown for reference to the 2000 U.S. Census data, which the normative sample of the BG-II was based on.

CHAPTER 3

RESULTS

Mean and standard deviations were calculated for the study sample based on standard scores (*M*= 100, *SD*= 15) for the BG-II copy phase. Results indicate a normally distributed, representative study sample across gender, ethnicity, and age where means and standard deviations approximately reach 100 and 15, respectively (see Table 1 for specifics). Using the current study's representative sample, the DIF analyses revealed only one BG-II Copy phase item to have both large and statistically significant differences in item difficulty across groups. The comparison between 4-year old African-American and Caucasian groups for item 3 of the BG-II Copy phase showed a DIF contrast magnitude of -0.85 logits, that was statistically significant at the Bonferroni-correct $p < 0.003$ level ($p = .001$). The item difficulty measure for African-American 4-year-olds was equal to -0.66 logits, while the Caucasian 4-year-olds was equal to -1.51 logits. The African-American group's logit value is higher, indicating that for some currently unknown reason African-American 4-year-olds appear to find item 3 of the BG-II Copy phase more difficulty than Caucasian 4-year-olds. No large and statistically significant differences in item difficulty were found across the second comparison group for Hispanic and Caucasian individuals.

Several items at different ages showed significant DIF contrast magnitudes, but they were not statistically significant at the Bonferroni-corrected p<.003 level.

Additionally, some items showed a statistically significant difference ($p < .003$), but the magnitude of the DIF contrast did not reach the criteria of being grater than 0.50 logits. Average item difficulty values for each group, DIF contrast magnitudes, and significance test results are shown for each comparison group by age in Tables 2 through 5.

Table 3.1

BG-II differential item functioning across group performance at age 4.

Copy Phase Item Number	Item Difficulty for Cultural Group	Item Difficulty for Caucasian	Differential Item Functioning Contrast	*p*-value
African-American vs. Caucasian				
1	-3.15	-3.15	0.00	1.000
2	-2.22	-2.57	0.35	0.182
3	-0.66	-1.51	0.85*	0.001**
4	-2.11	-2.58	0.47	0.073
5	-0.07	0.05	-0.13	0.640
6	0.05	0.51	-0.46	0.093
7	0.9	1.39	-0.48	0.100
8	2.02	2.29	-0.27	0.427
9	0.76	0.91	-0.15	0.600
10	0.76	0.74	0.02	0.932
11	0.63	0.89	-0.26	0.364
12	1.75	1.87	-0.12	0.714
13	0.97	1.25	-0.28	0.341
Hispanic vs. Caucasian				
1	-3.01	-3.05	0.03	0.892
2	-2.46	-2.49	0.03	0.903
3	-1.00	-1.47	0.47	0.061
4	-2.06	-2.51	0.45	0.072
5	0.46	0.05	0.41	0.120
6	0.40	0.47	-0.07	0.792
7	0.88	1.34	-0.46	0.098
8	2.53	2.23	0.31	0.370
9	0.64	0.88	-0.24	0.367
10	0.58	0.71	-0.13	0.624
11	0.35	0.86	-0.51*	0.057
12	1.55	1.82	-0.27	0.359
13	1.01	1.21	-0.21	0.560

Note. Significant differential item functioning contrast (>0.50 logi ts) are noted by a *; Significant bonferroni-corrected p-values (p< .003) are denoted by a **.

Table 3.2

BG-II differential item functioning across group performance at age 5.

Copy Phase Item Number	Item Difficulty for Cultural Group	Item Difficulty for Caucasian	Differential Item Functioning Contrast	*p*-value
African-American vs. Caucasian				
1	-4.17	-3.80	-0.37	0.349
2	-2.89	-2.93	0.04	0.915
3	-2.00	-2.20	0.20	0.605
4	-2.76	-2.81	0.05	0.900
5	0.45	0.24	0.22	0.581
6	0.20	0.94	-0.75*	0.063
7	1.34	1.49	-0.15	0.700
8	2.10	2.40	- 0.30	0.456
9	0.84	0.38	0.45	0.249
10	0.58	1.09	-0.51	0.197
11	1.59	1.55	0.04	0.911
12	2.76	2.37	0.40	0.335
13	1.97	1.31	0.65	0.100
Hispanic vs. Caucasian				
1	-4.28	-3.91	-0.37	0.358
2	-2.69	-3.02	0.33	0.406
3	-1.76	-2.26	0.50	0.211
4	-1.89	-2.90	1.01*	0.014
5	0.36	0.27	0.09	0.821
6	0.88	0.96	-0.08	0.841
7	1.92	1.54	0.38	0.337
8	2.05	2.45	-0.40	0.314
9	0.49	0.42	0.07	0.852
10	1.13	1.13	0.00	1.000
11	1.01	1.59	-0.57*	0.152
12	1.79	2.43	-0.64*	0.113
13	1.01	1.36	-0.34	0.385

Note. Significant differential item functioning contrsast are noted by a *; Significant bonferroni-corrected p-values are denoted by a **.

Table 3.3

BG-II differential item functioning across group performance at age 6.

Copy Phase Item Number	Item Difficulty for Cultural Group	Item Difficulty for Caucasian	Differential Item Functioning Contrast	*p*-value
African-American vs. Caucasian				
1	-4.87	-4.34	-0.53*	0.229
2	-3.94	-3.07	-0.86*	0.045
3	-1.21	-2.17	0.96*	0.028
4	-2.10	-2.25	0.15	0.710
5	0.58	0.23	0.35	0.402
6	0.15	0.80	-0.64*	0.128
7	1.14	1.52	-0.38	0.352
8	2.11	2.11	0.00	1.000
9	0.44	0.57	-0.13	0.749
10	0.44	1.02	-0.58*	0.165
11	1.84	1.50	0.34	0.407
12	3.27	2.60	0.67*	0.111
13	2.12	1.50	0.63*	0.135
Hispanic vs. Caucasian				
1	-3.96	-4.28	0.32	0.431
2	-2.88	-3.03	0.15	0.708
3	-2.49	-2.14	-0.35	0.378
4	-2.22	-2.22	0.00	1.000
5	-0.10	0.24	-0.34	0.409
6	0.44	0.79	-0.35	0.386
7	1.50	1.50	0.00	1.000
8	2.52	2.07	0.46	0.249
9	0.57	0.57	0.00	1.000
10	1.11	1.03	0.08	0.847
11	1.48	1.48	0.00	1.000
12	2.65	2.56	0.09	0.814
13	1.37	1.48	-0.11	0.783

Note. Significant differential item functioning contrsast are noted by a *; Significant bonferroni-corrected p-values are denoted by a **.

Table 3.4

BG-II differential item functioning across group performance at age 7.

Copy Phase Item Number	Item Difficulty for Cultural Group	Item Difficulty for Caucasian	Differential Item Functioning Contrast	*p*-value
African-American vs. Caucasian				
1	-3.95	-3.65	-0.29	0.499
2	-3.05	-2.75	-0.30	0.472
3	-1.67	-1.91	0.24	0.557
4	-1.94	-1.61	-0.33	0.421
5	-0.01	0.23	-0.25	0.549
6	0.81	0.39	0.43	0.294
7	1.21	1.26	-0.05	0.908
8	1.21	1.42	-0.21	0.597
9	0.54	0.61	-0.07	0.868
10	0.81	0.74	0.07	0.855
11	1.73	1.55	0.18	0.651
12	2.62	2.57	0.05	0.898
13	1.60	1.14	0.46	0.246
Hispanic vs. Caucasian				
1	-3.53	-3.65	0.12	0.784
2	-3.07	-2.75	-0.32	0.451
3	-2.03	-1.92	-0.11	0.793
4	-2.32	-1.61	-0.71*	0.094
5	0.32	0.25	0.07	0.865
6	0.60	0.40	0.21	0.616
7	1.70	1.27	0.43	0.283
8	2.23	1.42	0.81*	0.045
9	0.88	0.63	0.25	0.540
10	1.29	0.73	0.56*	0.170
11	1.16	1.55	0.39	0.340
12	1.97	2.53	0.56*	0.162
13	0.74	1.14	0.40	0.327

Note. Significant differential item functioning contrsast are noted by a *; Significant bonferroni-corrected p-values are denoted by a **.

CHAPTER 4

CONCLUSIONS

School psychologists frequently use norm-referenced measures when deciding children's eligibility for special education services or educational needs (Decker, 2008). Tests should accurately measure and represent an individual's ability, without factors like cultural background interfering with interpretation of performance. Measures that are influenced by cultural factors exhibit variance between groups, where scores significantly differ across different cultural groups, which has negative implications for interpretation of performance. However, establishing how much culture influences performance on standardized measures is complex (Ostrosky-Solis, Ramirez, & Ardila, 2004), where no test is "culture-free." Due to diverse populations significantly increasing within the schools, appropriate culturally sensitive assessment practices are needed, Test bias is especially concerning when it occurs as a result of cultural differences between groups of examinees, where the construct of interest determines what potential negative implications may occur.

Visual motor integration is one significantly studied construct that has been associated with school readiness, academic achievement, behavior, and neuropsychological assessment, and is believed to be culturally invariant. The current study sought to examine the cultural invariance of the Bender Gestalt-II (BG-II), which is one of the most commonly used measures (Archer & Newsom, 2000; Brannigan &

Decker, 2003; Decker, Allen, & Choca, 2006; Piotrowski, 1995; Sullivan & Bowden, 1997). The BG-II's low linguistic demands may make it ideal for individuals with diverse backgrounds. The cultural invariance of the BG-II was examined through assessing DIF on the copy phase items of the measure. It was hypothesized that due to the low linguistic demands of the measure, the BG-II would be found to be culturally invariant for African-American, Hispanic, and Caucasian groups with no significant DIF items across cultural-linguistic groups.

Results found a lack of significant DIF items on the BG-II Copy phase, supporting the hypothesis that the BG-II can be viewed as a culturally invariant measure due to minimal interactions between persons and items on the measure. However, results did find one significant DIF item on the BG-II. Specifically, item 3 of the BG-II for African-American 4 year olds was significant, indicating that caution should be used when interpreting visual-motor performance on the BG-II for item 3 with African-American 4 year olds. Overall, the BG-II was found to have virtually no DIF, indicating that this measure can be used with diverse populations in early age ranges as a screening measure for visual-motor skills. These findings are noteworthy because an analysis of the cultural invariance of the BG-II has not been previously conducted within the research literature.

At this point in time, there is no clear reason for the significant DIF in item 3 for the first comparison group. There is no evidence to support the notion that item 3 of the Copy phase holds some inherent cultural meaning that makes the item more difficult for African-American individuals and easier for Caucasian individuals. However, research indicates that socioeconomic status is related to exposure to fine and gross motor skills activities at early ages (Hoffman, 1966), indicating that African-American children may have slower development of visual motor skills in comparison to other ethnic groups as a result of their lower socioeconomic status (Koppitz, 1975; Sattler & Gwynne, 1989). One potential hypothesis is that due to slower development of visual motor skills in African-American children, the number of required motor movements for item 3 is too advanced, making this item more difficult when compared to other lower positioned items the BG-II. Item 3 of the Copy phase on the BG-II has 5 potential motor movements because it contains a total of 5 lines. In comparison, item 2 of the BG-II only requires one motor movement to copy an image that has a "U" shape, and item 4 appears to have 2 possible motor movements to copy an image that contains a smaller circle within a larger circle. The BG-II was designed to have items progress in their level of difficulty, so we would anticipate that item 2 should be easier than item 3, and item 3 should be easier than item 4. Results in difficulty levels for the African-American group show that this progression in difficulty does not occur in this sequence.

Item 2 of the Copy phase on the BG-II does in fact appear to be easier than item 3 across all age groups for the African-American group. However, even though differences in difficulty may not be statistically significant, item 4 appears to be easier than item 3 across all age groups for the African-American group. Differences in difficulty when comparing item 3 to item 2 and 4 are not found in the Hispanic or Caucasian groups, indicating a potentially unique issue with item 3 of the BG-II Copy phase for African-Americans.

Item 3 for African-American 4 year olds on the BG-II was found to have a difficulty value of -0.66 logits, while both item 2 and item 4 were easier at -2.22 logits and -2.11 logits, respectively. Item 3 for African-American 5 year olds had a difficulty value of -2.00 logits, while item 2 and item 4 were easier at -2.89 logits and -2.76 logits, respectively. Item 3 for African-American 6 year olds had a difficulty value of -1.21 logits, while item 2 and item 4 were easier at -3.94 logits and -2.10 logits, respectively. Finally, item 3 for African-American 7 year olds had a difficulty value of -1.67 logits, while item 2 and item 4 were easier at -3.05 logits and -1.94 logits, respectively.

The findings of this study are not without limitations. Due to the archival nature of working with a normative dataset, we were limited to the sample size originally corrected for each age group and ethnic group. Cases were selected from the original archival dataset of the BG-II, as previously discussed, to create the appropriate sample for the current study. Data for the BG-II was collected in the four U.S. census regions, Northeast, Midwest, South, and West, and data collection procedures were carefully designed to match to the percentages of the 2000 U.S. Census (U.S. Census Bureau, 2001). With that said, the normative data set for the BG-II is closely matched to the 2000 U.S. Census population percentages by race and ethnicity, but for the scope of this study it appears that the sample size for young children is somewhat limited in all ethnic groups except for Caucasian individuals. The analyses conducted for both African-American and Hispanic children at ages 5, 6, and 7 were completed on sample sizes that were less than 30 individuals.

Additionally, due to the smaller sample size of young children in the BG-II normative dataset, other ethnic categories could not be tested for cultural invariance. The normative data set for Asian individuals contained a total of 30 individuals (n = 13 Asian 4-year-olds, n = 7 Asian 5-year-olds, n = 5 Asian 6-year-olds, and n = 5 Asian 7-year-olds), while the dataset for the Other individuals contained a total of 34 individuals (*n* = 13 Other 4-year-olds, *n* = 5 Other 5-year-olds, *n* = 9 Other 6-year-olds, and *n* = 7 Other 7-year-olds). Knowing that the BG-II is commonly used for evaluation, it would be ideal to have a larger sample size in an attempt to replicate the true population.

Findings from the current study are able to inform clinical practice in an effort to promote nondiscriminatory assessment in the schools. Practitioner awareness that the BG-II has been found to be culturally invariant for young children will allow the measure to be confidently used as an appropriate method of assessment for individuals from diverse cultural backgrounds, which is significant due to the frequent use of the BG-II. Research on the cultural invariance of the BG-II should continue to facilitate future efforts towards the development and use of culturally appropriate measures.

One direction for future study would be to assess more young African-American children using the BG-II to replicate the findings within the current study for item 3. It is possible that with a larger sample size, significant DIF would be more likely at all ages and not just age 4 for African-American children. Another future direction could be for findings to be cross-validated with other measures of visual motor integration to determine if items with similar item difficulty levels exhibited any unique bias for African-American 3 year olds. Further, the current study could be expanded to include older individuals from the BG-II normative dataset to help advance our understanding of the cultural invariance of the BG-II. Having a full picture of the use of the BG-II with varying cultural-linguistic populations at all ages will support the current studies findings that the BG-II is a culturally invariant measure. Another direction could be to conduct DIF analyses using a two-parameter logistic model (2PL) model as opposed to the one-parameter Rasch model. A 2PL model looks at the probability of correctly answering an item based on person ability, item difficulty, and item discrimination (Furr & Bacharach, 2014). It is possible that the normative dataset may better fit a 2PL model, where items with higher item discrimination values impact the probability of correctly answering an item for individuals with varying ability level. Overall, continued research efforts are need within the field of psychology to expand knowledge on the influence of cultural factors on assessment, and to facilitate the development and use of culturally appropriate measures.

References

American Psychological Association (2007). *APA dictionary of psychology*. G. R. VandenBos (Ed.). Washington, DC: APA.

Andrich, D. (1982). An extension of the Rasch model for ratings providing both location and dispersion parameters. *Psychometrika, 47*(1), 105-113.

Archer, R. P., & Newsom, C. R. (2000). Psychological test usage with adolescent clients: Survey update. *Assessment 7*(3), 227-235.

Bart, O., Hajami, D., & Bar-Haim, Y. (2007). Predicting school adjustment from motor abilities in kindergarten. *Infant and Child Development*, 16, 597-615.

Becker, P., Grunwald, P., & Brazy, Y. (1999). Motor organization in low birth weight infants during caregiving: Effects of a developmental intervention. *Developmental and Behavioral Pediatrics, 20*(5), 344-354.

Beery, K. E. (1997). *The Beery-Buktenica Developmental Test of Visual-Motor Integration: 4th Edition, Revised.* Parsippany, NJ: Modern Curriculum Press.

Bender, L. (1938). *A visual motor gestalt test and its clinical use.* American Orthopsychiatric Association, Research Monographs (No. 3). New York: American Orthopsychiatric Association.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences.* Mahwah, NJ: Lawrence Erlbaum Associates.

Bracken, B. A., & McCallum, R. S. (1998). *Universal Nonverbal Intelligence Test*. Itasca, IL: Riverside Publishing.

Bradley, R. H., Corwyn, R. F. (2002). Socioeconomic status and child development. *Annual Review of Psychology, 53*, 371-999.

Brannigan, G.G., & Decker, S.L. (2003). *Bender-Gestalt II examiner's manual*. Itasca, IL: Riverside Publishing.

Brigham, C. C. (1923). *A study of American intelligence.* Princeton, NJ: Princeton University Press.

Brown, M. B., Holcombe, D. C., Bolen, L. M., & Thomson, W. S. (2006). Role function and job satisfaction of school psychologists practicing in an expanded role model. *Psychological Reports, 98*(2), 486-496.

Carlton, M. P., & Winsler, A. (1999). School readiness: The need for a paradigm shift. *School Psychology Review, 28*, 338-352.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor analytic studies*. New York, NY: Cambridge University Press.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54*, 1-22.

Center for Immigration Studies. (2012). Immigrants in the United States, 2010: A profile of America's foreign-born population. Retrieved May, 3, 2013 from http://www.cis.org/2012-profile-of-americas-foreign-born-population#execsum.

Cummins, A., Piek, J. P., Dyck, M. J. (2005). Motor coordination, empathy, and social behaviour in school-aged children. *Developmental Medicine and Child Neurology, 47*, 437-442.

Daly, C. J., Kelley, G. T., & Krauss, A. (2003). Relationship between visual-motor integration and handwriting skills of children in kindergarten: A modified replication study. *American Journal of Occupational Therapy, 57*, 459-462.

Dawson, G., & Watling, R. (2000). Interventions to facilitate auditory, visual, and motor integration in autism: A review of the evidence. *Journal of autism and developmental disorders*, *30*(5), 415-421.

Decker, S. L., Allen, R., & Choca, J. P. (2006). Construct validity of the Bender Gestalt II: Comparison with Wechsler Intelligence Scale for Children-III. *Perceptual and Motor Skills, 102*, 133-141.

Decker, S. L., Englund, J. A., Carboni, J. A., & Brooks, J. H. (2011). Cognitive and developmental influences in visual-motor integration skills in young children. *Psychological Assessment, 23*(4), 1010-1016.

DiStefano, C., & Morgan, G. B. (2010). Evaluation of the BESS TRS-CA using the Rasch rating scale model. *School Psychology Quarterly, 25*(4), 202-212.

Duncan, P. W., Bode, R. K., Min Lai, S., & Perera, S. (2003). Rasch analysis of a new stroke-specific outcome scale: the stroke impact scale1. *Archives of physical medicine and rehabilitation, 84*(7), 950-963.

Flanagan, D. P., & Ortiz, S. O. (2001). *Essentials of cross-battery assessment*. New York: Wiley.

Flanagan, D. P., Ortiz, S. O., Alfonso, V. C. (2013). *Essentials of cross-battery assessment* (3rd ed.). A. S. Kaufman, & N. L. Kaufman (Eds.). New York, NY: Wiley.

Fletcher-Janzen, E., & Ortiz, S. O. (2006). Cultural competence in the use of IQ tests with culturally and linguistically diverse children. *Gifted Education international*, *21*, 137-150.

Fox, C. M., & Jones, J. A. (1998). Uses of Rasch modeling in counseling psychology research. *Journal of Counseling Psychology, 45,* 30-45.

Furr, R. M., & Bacharach, V. R. (2014). *Psychometrics: An introduction* (2nd ed.). Los Angeles, CA: Sage Publications, Inc.

Goldberg, L. R. (1959). The effectiveness of clinicians' judgments: The diagnosis of organic brain damage form the bender-gestalt test. *Journal of Consulting Psychology, 23*(1), 25-33.

Harter. S. (1987). The determinants and meditational role of global self-worth in children. In N. Eisenberg, *Contemporary topics in developmental psychology* (pp. 219-242). New York, NY: Wiley.

Hebben, N., & Milberg, W. (2002). *Essentials of neuropsychological assessment*. A. S. Kaufman & N. L. Kaugman (Series Eds.) New York: Wiley.

Hoffman, S. D. (1966). The Bender Gestalt test (Doctoral dissertation). The University of Oklahoma, Norman, OK.

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

Koppitz, E. M. (1975). The bender-gestalt test for young children, volume II, research and application, 1963-1973. New York: Grune & Stratton, Inc.

Kulp, M. T. (1999). Relationship between visual motor integration skill and academic performance in kindergarten through third grade. *Optometry & Vision Science*, *76*(3), 159-163.

Kurdek, L. A., & Sinclair, R. J. (2000). Psychological, family, and peer predictors of academic outcomes in first- through fifth-grade children. *Journal of Educational Psychology*, 92, 449-457.

Lacks, P. (1999). *Bender Gestalt screening for brain dysfunction*. New York, NY: Wiley.

Lacks, P. & Newport, K. (1980). A comparison of scoring systems and level of scorer experience on the Bender Gestalt Test. *Journal of Personality Assessment, 44*, 351-357.

Lai, J. S., Cella, D., Chang, C. H., Bode, R. K., & Heinemann, A. W. (2003). Item banking to improve, shorten and computerize self-reported fatigue: an illustration of steps to create a core item bank from the FACIT-Fatigue Scale. *Quality of Life Research, 12*(5), 485-501.

Linacre, J. M. (2005). *A user's guide to WINSTEPS.* Chicago, IL: Mesa Press.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Eribaum.

McCallum, R. S., & Bracken, B. A. (1997). The Universal Nonverbal Intelligence Test. D.P. Flanagan, J. L. Genshaft, & P.L. Harrison (Eds.). *Contemporary intellectual* *assessment: Theories, tests, and issues* (pp. 268-280). New York, NY: Guilford Press.

McHale, K., & Cermak, S. (1992). Fine motor activities in elementary school: Preliminary findings and provisional implications for children with fine motor problems. *American Journal of Occupational Therapy, 46*, 898-903.

Nandakumar, R., Glutting, J. J., & Oakland, T. (1993). Mantel-haenszel methodology for detecting item bias: An introduction and example using the guide to the assessment of test session behavior. *Journal of Psychoeducational Assessment, 11*, 108-119.

National Center for Learning Disabilities. (2014). The state of learning disabilities (3rd ed.). Washington, DC: Author.

Ochoa, S. H., Rivera, B., & Ford, L. (1997). An investigation of school psychology training pertaining to bilingual psycho-educational assessment of primarily Hispanic students: Twenty-five years after Diana v. California. *Journal of School Psychology, 35*(4), 329-349.

Ortiz, S. O. (2004). Comprehensive assessment of culturally and linguistically diverse students: A systematic, practical approach for nondiscriminatory assessment. Retrieved from http://www.nasponline.org/resources/culturalcompetence/ortiz.pdf.

Ortiz, S. O. & Lella, S. A. (2005). Cross-cultural assessment. In Lee, S. W. (Ed.), *Encyclopedia of School Psychology* (pp. 136-139). Thousand Oaks, CA: SAGE Publications, Inc.

Ostrosky-Solis, F., Ramirez, M., & Ardila, A. (2004). Effects of culture and education on neuropsychological testing: A preliminary study with indigenous and nonindigenous populations. *Applied Neuropsychology, 11*, 186-193.

Piotrowski, C. (1995). A review of the clinical and research use of the Bender-Gestalt Test. *Perceptual and Motor Skills, 81*, 1272-1274.

Psychological Corporation. (2001). *Wechsler Individual Achievement Test-Second Edition: Examiner's manual*. San Antonio, TX: Author.

Reise, S. P., Widaman, K. F., Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*(3), 552-566.

Roid, G. H. (2003). *Stanford-Binet Intelligence Scales* (5th ed.). Itasca, IL: Riverside Publishing.

Rose, B., Larkin, D., & Berger, B. G. (1997). Coordination and gender influences on the perceived competence of children. *Adapted Physical Activity Quarterly, 14*, 210-221.

Sattler, J. M., & Gwynne (1982). Ethnicity and bender visual motor gestalt test performance. *Journal of School Psychology, 20*(1), 69-71.

Skinner, R., & Piek, J. (2001). Psychosocial implications of poor motor control coordination in children and adolescents. *Human Motivation Science, 20*, 73-94.

Sotelo-Dynega, M., Cuskley, T., Geddes, L., McSwiggan, K., & Soldano, A. (2011). *Cognitive assessment: A survey of current school psychologists' practices.* Research poster presentation at the annual conference of the National Association of School Psychologists, San Francisco, CA.

Stark, S., Chernyshenko, O. S., & F. Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91*(6), 1292-1306.

Sullivan, K., & Bowden, S. C. (1997). Which tests do neuropsychologists use? *Journal of Clinical Psychology, 53*, 657-661.

Sutton, G. P., Barchard, K. A., Bello, D. T., Thaler, N. S., Ringdahl, E., Mayfield, J., & Allen, D. N. (2011). Beery-Buktenica Developmental Test of Visual-Motor Integration performance in children with traumatic brain injury and attention-deficit/hyperactivity disorder. *Psychological assessment*, *23*(3), 805.

Tennant, A., Penta, M., Tesio, L., Grimby, G., Thonnard, J. L., Slade, A., ... & Phillips, S. (2004). Assessing and adjusting for cross-cultural validity of impairment and activity limitation scales through differential item functioning within the framework of the Rasch model: the PRO-ESOR project. *Medical care*, *42*(1), I-37.

Tolor, A. & Schulberg, H. C. (1963). *An evaluation of the Bender-Gestalt Test.* Springfield, IL: Thomas.

U.S. Census Bureau. (2001). *Census 2000 Summary File 1 United States*. Washington, DC: Author.

U.S. Census Bureau. (2011a). Population Distribution and Chance: 2000 to 2010. Retrieved April 4, 2014 from http://www.census.gov/prod/cen2010/briefs/c2010br-01.pdf.

U.S. Census Bureau. (2011b). The Black Population: 2010. Retrieved April 4, 2014 from http://www.census.gov/prod/cen2010/briefs/c2010br-06.pdf.

U.S. Census Bureau. (2011c). The Hispanic Population: 2010. Retrieved May 3, 2013 from http://www.census.gov/prod/cen2010/briefs/c2010br-04.pdf.

U.S. Census Bureau. (2011d). The White Population: 2010. Retrieved April 4, 2013 from http://www.census.gov/prod/cen2010/briefs/c2010br-05.pdf.

U.S. Census Bureau. (2012). Most children younger than age 1 are minorities, Census Bureau Reports. Retrieved May 17, 2013 from http://www.census.gov/newsroom/releases/ archives/population/cb12-90.html.

U.S. Census Bureau. (2013a). Language use in the United States: 2011. Retrieved, April 25, 2014 from http://www.census.gov/prod/2013pubs/acs-22.pdf.

U.S. Census Bureau. (2013b) Poverty rates for selected detailed race and Hispanic groups by state and place: 2007-2011. Retrieved April 25, 2014 from http://www.census.gov/

prod/2013pubs/acsbr11-17.pdf.

U.S. Department of Education (U.S.D.E.), National Center for Education Statistics (N.C.E.S.), Common Core of Data (C.C.D.; 2014). Racial/Ethnic Enrollment in

Public Schools. (2014). Racial/Ethnic Enrollment in Public Schools. Retrieved April 20, 2014 from https://nces.ed.gov /programs /coe/indicator_cge.asp.

Wechsler, D. (1991). *Wechsler Intelligences Scale for Children-Third Edition manual*. San Antonio, TX: Psychological Corporation.

Williams, J., Griebel, M. L., & Dykman, R. A. (1998). Neuropsychological patterns in pediatric epilepsy. *Seizure*, *7*(3), 223-228.

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