2-11-2004

Assessor Effects On The Evaluation Of The WISC-III

Sherecce A. Fields
University of South Florida

Assessor Effects On The Evaluation Of The WISC-III

by

Sherecce A. Fields

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Arts
Department of Psychology
College of Arts and sciences
University of South Florida

Major Professor: Vicky Phares, PhD
Cynthia Cimino, PhD
Carnot Nelson, PhD

Date of Approval:
February 11, 2004

Keywords: standardized test, examiner, matching, intelligence, ethnicity

Table of Contents

List of Tables

Assessor Effects On The Evaluation Of The WISC-III

Sherecce A. Fields

ABSTRACT

There have been many theories about cultural differences found between groups on intelligence test scores. The main debate has been between those in favor of a genetic explanation versus those in favor of a more environmental one. When considering environmental influences, one explanation has been that there could be differential effects due to the assessor. Although there have been several studies that have considered this possibility, the results are inconclusive. The current study attempted to tease apart the assessor effects by focusing on biases in the assessor alone and by eliminating effects from the test taker. The study is an experimental design where participants were randomly assigned a WISC-III protocol of members of different ethnic groups. It was hypothesized that different groups may score these IQ tests differentially depending on the race/ethnicity of the person who was assessed. Results showed that when given identical protocols, participants scored African American protocols lower than Caucasian American protocols in both high and average IQ conditions. Clinical implications of these results are discussed.

Assessor Effects of the Evaluation of the WISC-III

Over the course of research into intelligence tests, certain cultural differences have been evident. One major difference found between racial groups has been the scores obtained on intelligence tests, with African Americans scoring about one standard deviation lower than Caucasian Americans (Jones & Herndon, 1992). There have been many theories about cultural differences found between groups on intelligence test scores. The main debate has been between those in favor of a genetic explanation versus those in favor of a more environmental one. When considering environmental influences, one explanation has been that there could be differential effects due to the assessor. Although there have been several studies that have considered this possibility, the results are inconclusive. This study will attempt to tease apart the assessor effects by focusing on biases in the assessor alone and eliminating effects from the test taker.

History of Intelligence Testing

Intelligence testing is a relatively young endeavor. In the late 19th century, Sir Francis Galton developed the first comprehensive test of intelligence (Kaufman & Lichtenberger, 1999). Galton believed that because we get our knowledge from our senses, higher intelligence should be evidenced by better sensory discrimination ability which led him to develop his tests of sensory discrimination and motor coordination (Sattler, 2001).

All of Galton's tests had strong reliability and consistency, however, none proved to be valid measures of the construct of intelligence (Kaufman, 2000).

While Galton's worked focused on more sensory measures of intelligence, his counterpart in France, Alfred Binet, focused on more higher order abilities, especially language abilities. Binet's work on intelligence testing is the basis for the tests that we use today (Sattler, 2001). "Binet conceptualized intelligence as the ability to demonstrate memory, judgment, reasoning, and social comprehension" (Kaufman, 2000, p.446). Binet and his colleagues (Victor Henri and Theodore Simon), at the request of the French ministry of education, developed tests to measure the intelligence of children in order to identify those who could benefit from public instruction. Another major contribution of Binet's is the introduction of the idea of a mental age; his scale was developed to determine whether a child was performing at their appropriate age level (Sattler, 2001). The Binet scale was translated into English and adapted to American culture by Lewis Terman. Terman also completed a careful standardization of the scale on a sample of American children and adolescents (Kaufman & Lichtenberger, 1999).

Another major influence on the way intelligence is assessed today is David Weschler. Weschler borrowed from the Stanford-Binet and the Army Alpha and Beta tests to develop his scale of intelligence. Weschler insisted that everyone be evaluated on both performance and verbal scales and that profiles be provided to supplement the global measure of intelligence that is obtained (Kaufman, 2000). The development of his scales was not based on theory but on his view of IQ tests as a method to assess personality

(Kaufman & Lichtenberger, 1999). The first of these tests was the Weschler Bellevue, which was developed in 1939. Since that time there have been several revisions and the tests have been standardized and re-normed. There have also been tests created to assess intelligence in children. The Weschler Intelligence Scale for Children – Third Edition (WISC-III; Weschler, 1991), meant for children ages 6 years to 16 years and 11 months, is a direct descendent of the Weschler Bellevue.

The WISC-III contains 13 subtests, which yield three IQ scale scores: verbal, performance, and full scale. The WISC-III has been standardized on 2200 children from 6 to 16 years of age and has been stratified by gender, age, race-ethnicity, geographic region, and parental education (Kaufman, 2000). The WISC-III has strong reliability with ranges from .89 to .97. The lowest internal consistency value is r = .89 (Sattler, 2001). The WISC-III has been validated for criterion by comparison with other measures of intelligence like the Stanford-Binet and has a mean correlation with those tests of r = .72 (Zimmerman & Woo-Sam, 1997).

Controversies in Race and Ethnicity

Since before the use of the Army Alpha tests as screening devices for immigrants coming into the US, there has been controversy surrounding the use of intelligence tests to assess culturally diverse groups (Kamphaus, 2001). This controversy is especially evident in the use of different measures of intelligence for assessing minority groups in the United States. One example of this controversy is Samuel Morton's use of head size in the 1800s to theorize that blacks had lower intelligence because of smaller brain capacity (Jones & Herndon, 1992), a theory that still existed during civil war times. Even in recent years, there

have been theories about the intellectual inferiority of African Americans based on scores obtained from intelligence tests (Herstein & Murray, 1994; Jensen, 1969).

Cultural differences have consistently been found in standardized IQ test scores with African Americans scoring about fifteen points lower than Caucasians (Jones & Herndon, 1992; Lynn, 1995; Neisser, 1996). Mean score differences between other cultural groups have not been studied extensively. The mean scores for Hispanic Americans seem to lie between Caucasians and African Americans in the United States. Native Americans have been studied even less than other cultural groups and their scores tend to be similar to groups tested whose first language is not English (Neisser, 1996). The difference between African Americans and Caucasians has also been seen in other countries besides the United States (Skuy et al., 2000). Although this difference could be decreasing (Lynn, 1998; Vincent, 1991), the decrease has not been found to be statistically significant and seems to be very small.

Debates about the cause of IQ differences have gone on for years with most of the debate focusing on two camps: genetics and environment. This debate seems to have come in phases with the first phase springing from proponents of genetics explanations in the 1800s. The accepted view at that time was that the Caucasian race was the most intelligent and other races were inferior due to innate differences (Lynn, 1995).

In the 1930s, this view was challenged by environmentalists who felt that differences in environment were the primary cause of the gap, and that there was no proof that culturally different groups varied in innate ability (Lynn, 1995). Jensen consequently challenged the environmental view in the 1960s. Jensen's view is that intelligence is an innate ability and that African Americans have a deficit when compared to Caucasians. He argues that these deficits cannot be decreased because intellect is highly heritable (Jones & Herndon, 1992). The idea that heritability represents a test of innate ability has been challenged time and again. Angoff (1988) argues that heritability is only a population statistic and that it says

nothing about mean differences between groups or the changeability of the trait. Other environmental theories have been developed to explain the mean difference in scores (Dickens & Flynn, 2001). This debate continues today and there are various theories that support both sides although most believe that the difference in scores is due to a combination of environment and genes.

Possible Explanations of Racial Differences in Measured Intelligence

Genetic theories

There has been much research on the contribution of genes to intelligence. This line of research has also been used to explain the differences found between groups on intelligence test scores. The major argument in favor of a considerable genetic contribution is that of heritability. An estimate of heritability is a population statistic that describes the proportion of variance of a trait (for example intelligence) that is attributed to genetic differences (Grigorenko, 2000; Sattler, 2001). Studies of the heritability of intelligence have found wide ranging values, however, there is debate about whether the value is around 50% or 75% (Grigorenko, 2000; McGue et al., 1993). The major proponents of this view have been Jensen (1969), Hernstein and Murray (1994), and Rushton (1992).

The argument in favor of heritability is that intelligence is highly heritable and therefore genes play a major role. It has been shown that genetically similar family members have more highly correlated IQ and cognitive functioning than non-related individuals (McGue, 1993; Sattler, 2001). This pattern is the basis of how heritability is calculated in the first place.

The heritability index is determined by twin and adoption studies. Arguments against heritability as evidence for a genetic cause for the IQ differences between cultures are numerous. The biggest argument is that heritability applies to certain populations at certain times and does not give you an idea of the changeability of a trait or its genetic basis.

Another argument is that a trait can be heritable and have nothing to do with genetics. One example given by Grigorenko in 2000 was that of piloting knowledge. The example given is that "some time ago, when only men flew planes, the heritability of knowing how to fly a plane was high, because differences in whether a person had knowledge were accounted for by a chromosomal difference."

More recently, however, research on genetic differences has focused on finding actual biological correlates of intelligence. One major line of research is to find the genes related to intelligence. One obvious argument is that of some forms of mental retardation. For example, it is known that individuals with Downs Syndrome have a chromosomal anomaly, and that this disorder is completely genetic. Individuals with Downs Syndrome tend to experience severe mental retardation. The second most important cause of mental retardation after Downs Syndrome was identified as a gene on a chromosome called Fragile X (Plomin & Petrill, 1997). Individuals with this problem usually perform in the mild to moderate range of intelligence.

Another example of genetic effects on intelligence would be children with Williams Syndrome. The genetic underpinnings of Williams Syndrome involve a submicroscopic deletion of about 20 contiguous genes on chromosome 7, including the gene for elastin (Bellugi et. al., 2000). These individuals have intelligence quotient scores in the mild to moderately IQ range. Lastly, Phenylketonuria (PKU) is caused by a single gene on chromosome 12 and causes severe mental retardation (MacLullich et. al., 1998).

Not only have there been genes found to be associated with retardation, but there has also been some research on the genetic bases of the cognitive decline seen in Alzheimer's patients. Apolipoprotein E (ApoE) is a protein that is on chromosome 19 and has three common alleles (e2, e3, and e4). e4 has been associated with a lower age of onset and a higher risk for Alzheimer's Disease (MacLullich et. al, 1998).

Although few researchers dispute the effect of genes on intelligence, many find that this explanation cannot be the only explanation for the cultural differences found. They hypothesize that there may be other factors related to the tests themselves or the testing environment that could contribute to the difference in test scores.

Bias in measurement

There have been many arguments that standardized measures of intelligence such as the Weschler scales are culturally biased in that they favor the majority culture (i.e. Caucasian, western society). This argument is based mostly on the discrepancy of mean scores between cultural groups, and so it is a statistical definition of bias that is being considered and therefore psychometrics research has been used to solve this problem. Of the psychometrically based arguments of bias, there are a few common trends/sources of bias: inappropriate content, inappropriate standardization, measurement of different constructs, and differential predictive ability (Reynolds et al., 1999).

When considering the issue of bias in content validity, there are several methodologies that can be used. One of the most basic is to have expert judges review the items and select the ones that appear to be biased. This method is not empirically based. Another method uses the differential difficulty of the items in the test to determine bias. According to this method, if items are more difficult for one cultural group versus another, the items could be biased. This method has been tested with a variety of statistical procedures. One of the most common procedures before the 1980s was to use ANOVA and related procedures to examine the group by item interaction term. Also, methods derived from item response theory (IRT) have been used. In this technique, item characteristic curves (ICCs) are developed for each cultural group. These curves are based on three parameters (discrimination power, difficulty, and how well one would do by guessing alone). In order to use this technique, ICCs from different groups are compared based on

each question. There are other methods, but these are the most widely used. Based on a number of studies, content bias has not been found consistently (Anastasi, 1988; Hickman et al., 1986; Reynolds et al., 1999).

The question of appropriate standardization samples seems to have come to an end. Previous versions of most of the standardized measures had questionable standardization samples, however, the more recent revisions are well standardized across cultures (Sattler, 2001).

Another argument for bias in measures that assess intelligence is that they do not assess the same construct across cultural groups. Determining the construct validity of a measure can be done in different ways, one of the most popular being factor analysis (Reynolds, 1991). Factor analysis is a process by which items are grouped together based on how well they correlate. It is assumed that if the same items cluster together for each group then the test is measuring the same construct. There has been extensive research on the factor structure of the WISC and WISC-R and these results are thought to generalize to the WISC-III. This research has shown that WISC items factor the same for Caucasian, African American, Mexican American, and Native American children (Gutkin & Reynolds, 1981; Oakland & Feigenbaum, 1979; Reschly, 1978).

Another way to measure construct validity is through estimates of internal consistency (Reynolds, 1999). This process determines how well the items measure the same construct. This value should be the same between groups if a measure is to be considered unbiased. Studies looking at the Weschler measures have found no evidence of bias (Braden, 1999).

Lastly, bias in predictive ability of the tests has been considered. In order for a test to be biased in predictive ability, there must be a constant error in predicting scores of one group versus another. This bias is usually shown statistically by using regression techniques. Once lines are computed, they are compared on slope and intercept for

differences. If significant differences are found, the measure could be biased. This research, however, is also based on the reliability of the outcome that is being predicted. There have been few studies that have looked at the predictive validity of the WISC-III. One study by Weiss and Prifitera (1995), looked at prediction of the Weschler Individual Achievement Test (WIAT) based on WISC Full Scale IQ scores and found no bias. Most of the empirical evidence suggests that there is no bias in predictive ability (Clarizio, 1978; Reynolds, 1999). Because there is inconclusive psychometric evidence for bias in the intelligence tests themselves, other environmental theories of cultural differences have been considered.

Environmental Theories

Environment has been shown to have effects on IQ scores. Many variables have been explored, and they usually tend to co-vary among themselves and with genetic factors. Many factors have been explored such as birth weight, nutrition, family background, poverty, and family configuration.

Low birth weight and very low birth rate could be increased risk factors for lower intelligence. The intelligence test scores of very low birth rate children tend to be lower than those of normal birth weight children (Ramey et al., 1999; Sattler, 2001; Sigman, 1998). Within the low birth weight range, those who have lower weights have larger IQ deficits than those closer to normal weights. It has also been shown that the association between birth weight and IQ is evident in normal birth weight individuals (Matte et al., 2001).

Inadequate nutrition has also been shown to have a deleterious effect on intelligence. It has also been shown that if you give young children vitamins their non-verbal IQs can increase by as much as 9 points (Sattler, 2001; Sigman, 1998).

There are several family background variables that have been associated with intelligence. Family size has been shown to influence intellectual development. There tends

to be an inverse relationship between family size and cognition, the so-called dilution effect argued by Downey (2001). This theory suggests that families only have a certain amount of resources and that the more children born to that family, the less resources will be available. Zajonc (2001) posits another theory, which he calls the confluence model. According to this model, intellectual development is tied to how family members interact with each other, he argues that a first-born child may benefit at a certain age (around 11 years old) by having a sibling to "teach". Rodgers (2001) takes an entirely different stance on the family size issue. According to his admixture theory, different between-family processes (such as SES) are the actual causal explanations for the within family effects that are seen.

Poverty has been shown to have serious negative effects on intellectual development, and persistent poverty has a worse effect than transitory poverty (Sattler, 2001). It has also been shown through census data that three times as many black and Hispanic children live at or below the poverty level. This variable could be said to mediate many of the other family variables that appear to have an influence on intellectual development. Minority children tend to come from larger families, single-family households, or have parents with lower educational levels. Other variables that may co-vary with poverty could be low birth weight and poor nutrition. One study by Brooks-Gunn et al. (1996) found the usual 15-point difference between African Americans and Caucasians in IQ in their sample, however, after adjusting for family and neighborhood poverty, the difference was reduced to 8.5 points. Finally, after adjusting for variations in provisions of learning experiences and maternal warmth the difference was reduced to 3.4 points. This study is evidence that family factors may contribute substantially to the ethnic differences seen in IQ scores.

There are other environmental theories posited to explain the difference in mean IQ scores seen between cultures. One such theory is the difference theory presented by Segall, Dasen, Berry, and Poortinga in 1990. This theory hypothesizes that the tests are biased in favor of the majority culture that constructed the test. They believe that intelligence develops

differently in different cultures. One proponent of this theory found that African American children adopted into Caucasian American families had different cognitive styles and an IQ advantage compared to those reared in culturally similar adoptive families (Moore, 1986). It is argued that this IQ advantage is due to being socialized in a culturally dominant household.

Another environmental theory is the deficit theory. The argument for this theory is lead by Flynn and colleagues (Lynn, 1995). They do not believe that the tests themselves are biased. It is argued, however, that a number of environmental factors may each account for small decrements in IQ and that taken together these factors could account for the mean difference that has been found.

Differences in the testing environment

Motivation and Test Taking Strategies

One idea from achievement literature is that African American students have lower achievement motivation than their Caucasian American counterparts. It has been theorized that African American students have lower test taking motivation or less achievement motivation (Banks et al., 1995; Chan et al., 1997). There have been many lines of research to determine why there are achievement differences. One major line of research has been on locus of control. Students who believe that consequences (i.e. achievement outcomes) are a direct result of their actions are said to have an internal locus of control, whereas those who believe that consequences are not directly related to their behavior (i.e. they believe it is related to luck or task difficulty) are said to have an external locus of control.

Internal locus of control has been shown to be positively related to academic achievement and tends to be more prevalent in individuals in higher socio-economic brackets (Banks, 1988). There have been several studies on this topic, however, few of them have attempted some sort of cross-cultural analysis and the results of those studies are

inconclusive (Castenell, 1984). The general consensus is that African Americans have a more external locus of control (Graham, 1988).

There have been other theories about test taking strategies and locus of control. In one study by Banks et al (1995), White children were shown to reward effort for those with lower ability levels while Black children rewarded effort in high ability more than for low ability persons. It was argued that while White children could see high effort as a way to compensate for low ability, Black children may find this strategy ineffective. It was also hypothesized that Black children may believe that incentives for success are less reliable than the consequences of failure.

This idea has been argued before and has been discussed in a parenting framework. Epps (1969) argued that Black parents usually punish failure in a much harsher manner than they reward effort and success and therefore, Black children have a much stronger fear of failure than hope of success. If this theory is true, Black children would be much less likely to attempt to answer questions that they were unsure of and this strategy could be a disadvantage in test taking especially for tests like the WISC-III.

Although this area of research has potential for helping to explain the intelligence test score differences found between African Americans and Caucasians, there has been little research done in this area. In addition, it is unlikely that this theory alone would fully explain the results.

Examiner/Examinee Matching

Another environmental characteristic that could explain IQ differences is the issue of client-therapist matching, specifically the match in ethnicity between the examiner and examinee in an assessment situation. Client-Therapist matching has been researched and debated for many years. Most of the studies that have looked at this problem have considered intake procedures such as initial assessment and assignment of diagnostic labels. Few studies have examined outcomes or the treatment process itself. It has been

shown in the literature that African American clients are more likely to be referred for inpatient versus outpatient treatment, are more likely to be given pharmacotherapy versus psychotherapy, and have fewer sessions with their primary therapist (Whaley, 1998). These studies look at the overall status of minority populations, but do not take into consideration client-therapist matching.

Of those studies that do look at matching, some interesting patterns have been found. One study by Jenkins-Hall and Sacco (1991) found that when viewing a videotape of a white versus black depressed client with similar symptoms, the white therapists rated the black clients more severely. In another study by Geller (1988), researchers manipulated intelligence and race of client and asked therapists to evaluate ability to engage in psychotherapy. White psychiatrists were more likely to state that the less intelligent white client was more able to engage in psychotherapy than a more intelligent black client and the black client was more likely to be recommended for medication. In a more recent study by Orrell-Valente et al. (1999), the level of therapeutic engagement between parents of at-risk children and their intervention coordinators were examined. It was found that there was better family engagement with the intervention coordinators when clients were matched on race/ethnicity and socioeconomic status.

Similar evidence has been found with other cultural groups as well. Sue et al. (1991) found that ethnic match was a significant predictor of positive treatment outcome for Mexican Americans. Later, Sue (1998) found that Asian Americans fared better in a therapeutic situation when they were matched ethnically, linguistically, or both. Similar effects were found in this study with Mexican Americans. African Americans and Caucasian Americans in this study were shown to attend more sessions if matched culturally, although match was not associated with premature termination for African Americans. Sue also looked at clients who attended ethnic-specific programs and found that they had lower dropout rates and stayed in the program longer than those who used more

mainstream services. In a more recent study by Hall, Kaplan, Lee & Little (2002) ethnic, language and gender matched pairs had significantly higher treatment outcome and less dropout from treatment sessions. Finally, in a study by Malgady and Constantino (1998), Hispanic clients were matched with clinicians on variables of ethnicity and language of interview. They found that Hispanic clinicians rated patients as more severely impaired than Caucasian clinicians, especially in bilingual or Spanish interviews.

Not only have matching differences been found in the literature for therapy and psychiatry, but also in a study of physicians' perceptions of patients based on SES and ethnicity (van Ryn & Burke, 2000). The researchers found that physicians' perceptions of patients were influenced by race and SES. Black patients were more likely to be considered as engaging in noncompliant and risky behaviors and were rated as less intelligent than their white counterparts. One interesting thing about this study was the distinction made between ethnicity and SES. Ethnicity was associated with ratings of intelligence, feelings of affiliation toward patient, and likelihood of high-risk behaviors, whereas, SES was associated with broader perceptions in various domains.

As shown, most of the therapy literature suggests that it can be advantageous to match clients ethnically with therapists. Most of the studies shown suggest harsher ratings and less optimism for therapeutic outcomes by mismatched therapist-clients. If therapy research suggests that matching may be a good idea, assessment research is unclear to say the least.

Assessment literature on the match between examiner and examinee has been extremely contradictory and riddled with methodological problems. There were a few studies done on this issue before 1980, but little has been done since that time. Two review articles were published in 1982 that summarize these findings.

Graziano, Varca, and Levy (1982) defined the concept of an adequate and complete study for studying race of examiner effects. According to their criteria, a study

is considered adequate if it has at least two examiners of each ethnic group to avoid confound, has no systematic bias in the assignment of examiners to examinees, and has sufficient power to detect a difference. A study is considered complete if it is capable of detecting examiner by examinee interactions and so therefore must have examinees from all ethnic groups under consideration. They reviewed 28 articles that studied race of examiner effects on IQ scores. Of those 28 articles, only fifteen were adequate and complete. It was found that most of the adequate and complete studies (8 out of 15) found no difference. Of those studies, only five used some form of the WISC as the IQ measure, and three out of the five found a significant examiner by examinee effect. One study (Solkoff, 1972) found that both black and white children received higher scores from the black examiner. Another study (Solkoff, 1974) found that white students received higher scores with black examiners while the black children received lower scores with the black examiners. Lastly, a study by Savage (1971) found that white children scored higher except on Block Design where black children scored higher with an examiner of the same race.

Another study by Sattler and Gwynne (1982) reviewed 27 articles that studied race of examiner effects on IQ scores. This study came out the same year and of the 27 articles, not all were the same as those used by Graziano et al. They found that in 23 of the 27 studies, no significant relationship was found. In this review article, the criteria for an adequate and complete study were not used. When the studies from this paper are subjected to the same criteria, seven studies appeared to meet the criteria, and four of the seven had significant findings. Three of the four studies were the same as those mentioned

in the Graziano et al. paper (1982). The one study that differed was by Samuel et al. (1976) who found that white examiners obtained higher scores from both black and white subjects.

Analysis of the previous articles and inspection of the articles they reviewed have shown that the data are equivocal as to whether there could be an examiner by examinee racial effect. No articles could be found that discussed this topic directly, however, a few articles were found that considered this topic indirectly. One study by Mishra (1983) examined the effects of examiners' prior knowledge of subjects' ethnicity and IQ scores on scoring Stanford Binet protocols. They took 36 protocols and broke them into 4 groups of nine matched on IQ score. They found that there was no significant effect on the scoring. However, they did not report the ethnicity of those who scored the protocols and therefore could not determine a examiner by ethnicity interaction.

Another study by Terrell and Terrell (1983) examined the relationship between race of examiner, cultural mistrust, and IQ performance of black children. This study does not meet the criteria described by Graziano et al. for a complete study. Therefore, they were not able to determine an interaction. A main effect for race of examiner was not found. They did find a significant interaction between race of examiner and level of mistrust suggesting that there could be other factors in the testing situation that could attribute to the lower scores obtained by minorities.

A couple of studies considered examiner effects with Mexican Americans as well. A study done by Oakland and Glutting in 1990 investigated whether the test observations

of White psychologists were biased based as a function of the examinee's race (Mexican American, African American, or Caucasian American), gender, or SES. All tests were administered by white examiners, and therefore no interaction between race of examiner and race of examinee could be tested. The researchers only considered observers' ratings of the children's behavior. The results were correlational and found that the observations were significantly correlated with their test scores. Specifically, the observations of test behaviors indicate considerable intrasession validity for evaluating children's attentiveness, confidence, and cooperation relative to their WISC-R performance irrespective of children's race, SES, or gender.

Lastly, a study by Mishra (1980) looked at a 2 by 2 design between Mexican American and Caucasian American examiners and examinees. They found that Mexican American examinees scored lower when tested by members of a different cultural group than their own. However, the researchers did not look at the interaction effect.

The studies of ethnicity of examiners and examinees are confounded by other variables in the testing situation, for example, anxiety experienced by the examinee. No studies were found that considered the scoring practices of examiners of different ethnicities controlling for other testing situation variables. Specifically, what effects are inherent within the examiner themselves that could account for differential testing situations. One theory that could help explain this could be that of aversive racism.

Aversive Racism

Social psychologists have studied the effects of stereotyping on members of minority groups for years and there are many theories that account for stereotyping

behavior. Many of the theories, however, do not account for more covert forms of stereotyping and racism that could account for more well intentioned members of society having inherent biases toward members of minority groups. One theory posited by Gaertner and Dovidio (1986) describes how this phenomenon could occur. According to this theory, there is an "aversive" form of racism that is more subtle that characterizes many white Americans today who possess strong egalitarian values and who believe that they are not prejudiced.

Not only do aversive racists possess negative feelings and beliefs of which they are unaware, they also have a strong desire to be non-prejudiced and these two feelings together form the basis for the ambivalence that characterizes aversive racists (Dovidio & Gaertner, 1998). There are several patterns that seem to emerge. First, whether or not and when they discriminate is all dependent upon the appearance of being prejudiced. If it is obvious to themselves or others that a decision could be interpreted in a racial way, aversive racists are not likely to engage in discriminatory behavior. However, if there is a more ambivalent situation in which their behavior can be rationalized through some other means, they are quite likely to engage in more discriminatory behaviors. There are several studies that give evidence to this theory (Dovidio & Gaertner, 2000; Gaertner & Dovidio, 1977).

Second, these behaviors are often expressed as pro-white behaviors rather than anti-black behaviors (Nelson, 2002). For example, in one study by Dovidio and Gaertner in 1991, two sets of participants were given questionnaires. One group was given questions on a one-dimension scale from good to bad and asked to rate characteristics of whites and African Americans. The second group was given two sets of questions, one set negative and one set positive towards blacks and whites. The first group with the one dimension scale did not differ in their evaluation of blacks and whites, however, the second group, while not rating blacks more negatively, rated whites more positively. Kline and Dovidio (1982) also found that when students were asked to make admissions decisions for their university

between poorly, moderately, and highly qualified applicants based on race, there were no differences found in the poorly qualified and few differences found in the moderately qualified applicant pools. However, large differences were found in ratings for the highly qualified pool with the white applicants being rated more highly.

Despite trends toward more tolerance of society toward minority members, there do seem to be more subtle forms of racism that could disadvantage minority groups or unfairly benefit majority groups. If this theory is applied to intelligence testing, perhaps the examiners have some negative feelings toward the examinees and therefore could covertly "sabotage" the results of the intelligence test scores in some unconscious way if this process is not obviously related to racial issues.

Current Study

The current study examined scoring practices of examiners based on the child client's ethnicity. The current study attempted to address limitations in the literature in several ways. First, none of the previous studies controlled for differences in the testing environment itself. The present study involved using an already administered WISC-III protocol and therefore, controlled for client or environmental differences. Secondly, this study used an experimental design, which allows for causal interpretation. Lastly, in most previous studies, participants were professionals in the field. In this study, the participants were graduate students. The rationale for using graduate students is that they are still very aware of scoring decision rules and they are the professionals of tomorrow.

The present study attempted to address several important questions: 1) Do emerging professionals in the field score one ethnic group higher on Full Scale, Verbal or Performance IQ than others and does this vary as a function of how gifted they think the child is? 2) Of those subtests that have lower inter-rater reliabilities, would they be more

likely to show interaction effects than those with higher inter-rater reliability? 3) Will certain ethnic groups be more lenient scorers than others? Specific hypotheses are listed below:

1. It is expected that there will be higher IQ scores for gifted versus non-gifted protocols. Based on aversive racism theory (Kline and Dovidio, 1982), it is also expected that Caucasian gifted vignettes will be scored higher than any of the other three vignettes (Caucasian non-gifted, African American gifted and non-gifted).
2. Based on previous research (Sattler et al., 1978; Cuenot & Darbes, 1982), it is hypothesized that subtests with lower inter-rater reliability (Comprehension, Similarities, and Vocabulary) will show higher scores for Caucasian gifted vignettes than any of the other three vignettes (Caucasian non-gifted, African American gifted and non-gifted).
3. If an adequate number of subjects are obtained, based on previous research (Solkoff, 1972), it is hypothesized that African Americans will be more lenient (i.e., will provide higher scores) overall with their scoring practices than Caucasians.

Method

Participants

Based on a power analysis (Cohen, 1992; alpha level = .05, power = .80, and a large effect size), the minimum number of participants required per group is 18 for a minimum total number of subjects of 72. This number was based on four groups with the following make-up of the fictional vignette: African American boy – gifted, African American boy – non-gifted, Caucasian boy – gifted, Caucasian boy – non-gifted. The final make-up of the sample consisted of the following make-up of the fictional vignette: African American boy – gifted (N = 17), African American boy – non-gifted (N = 12), Caucasian boy – gifted (N = 24), Caucasian boy – non-gifted (N = 19) for a total of 72 participants. Although equal numbers of packets were distributed, a slightly unequal number of packets per cell were returned.

Participants in this study ranged in age from 22 to 53 with a mean age of 28.03, with 83.3% being female and 16.7% being male. The majority of respondents were Caucasian (59, 81.9%). The remaining sample consisted of African Americans (3, 4.2%), Hispanic Americans (6, 8.3%), Asian Americans (3, 4.2%) and others (1, 1.4%). The demographics of the sample are consistent with the make-up of doctoral programs in the United States based on the 2001 Annual Report from the American Psychological

Association (2002), which reports a make-up of 72.7% women and 22.8% ethnic minority.

The participants were graduate students in clinical and counseling psychology PhD (44, 61.1%), clinical and counseling PsyD (6, 8.3%) and school psychology PhD (22, 30.6%) programs. They had an average of 28.61 months experience with the WISC-III and an average of 27.76 months of direct experience with children and adolescents. Participants were selected at random by the directors of training programs at each individual institution. Follow up letters and packets were also sent one month after initial contact. Out of 700 distributed packets, only 72 were completed and returned. Packets were distributed to 108 institutions and the current sample represents participants from 31 institutions.

Measures

A. *WISC-III protocol (Appendix A)*

The Weschler Intelligence Scale for Children-Third Edition (WISC-III) is a standardized test of intelligence for children aged 6 years through 16 years 11months. The WISC-III is comprised of subtests that measure different types of intelligence and result in three types of scores (Verbal IQ, Performance IQ, and Full Scale IQ; Weschler, 1991). The psychometric properties of the WISC-III have been well established and include standardization on a sample that closely approximates the 1988 census data, as well as extensive reliability and validity data (Sattler, 2001).

The protocols that were administered were created to be somewhat ambiguous in interpretation (Appendices A and B). They only included answers to the questions in each subtest, and required the participant to interpret the answers and score them based on their interpretation. One protocol yielded scores in the normal range of intelligence, while the other protocol yielded scores in the gifted range. Instructions on how to administer and score WISC-III protocols were included for ease of completion.

B. Child Vignettes (Appendix C)

A short paragraph was created to be included with each of the WISC-III protocols (Appendix C). This paragraph included details about a fabricated male child of age 10 years, 4 months. The age of 10 years and 4 months was chosen because it was a median age of the standardization sample for the WISC-III. The majority of referrals for testing are for male children and therefore the gender was chosen to be male for each vignette. Each participant received similar paragraphs. The only thing that was different in each vignette was the ethnicity of the child "subject" and a sentence that stressed the gifted nature of the assessment. Two racial ethic groups were chosen: Caucasian and African American. These groups were chosen because they have been the most extensively researched on this topic, and they represent the two most populous groups in the United States.

Procedure

Letters (Appendix D) were mailed out to the directors of training programs at various universities in the United States along with 4 packets. The directors were asked to pick a diverse mixture of students and to distribute the packets. Each director received the same four packets. Included in each packet was a letter (Appendix E) to the participant with instructions on how to complete the protocol and a postage paid envelope. Participants were instructed to complete a brief demographics questionnaire (Appendix F) after scoring the WISC-III protocol, which contained a set of experimental check questions such as " Do you remember the gender of your subject, etc." Follow-up letters were mailed to directors of training programs after 4 months with instructions to distribute enclosed follow-up letters to the students that were given packets. A total of 700 packets were sent out and only 72 were received, for a response rate of 11.6%. Unfortunately, it is not clear whether all of the packets were distributed to the graduate students by their training directors.

Results

Participants appeared to be equally distributed between the four cells for this study. See Table 1 for sample characteristics for each vignette type. Chi-square analyses were conducted for gender, ethnicity, and program of participants for protocol type and child vignette. No significant differences were found for gender of participant on protocol type (χ^2 (2, N = 72) = 2.02; p = 0.36) or child vignette (χ^2 (1, N = 72) = 0.011; p = 0.91). No significant differences were found for ethnicity of participant on protocol type (χ^2 (4, N = 72) = 6.33; p = 0.17) or child vignette (χ^2 (4, N = 72) = 2.37; p = 0.67). No significant differences were found for program of participant (i.e. clinical, counseling or school psychology) on protocol type (χ^2 (2, N = 72) = 2.02; p = 0.36) or child vignette (χ^2 (2, N = 72) = 2.14; p = 0.34).

Each demographics questionnaire included a question about the ethnicity of the child to ensure that protocols were scored with ethnicity in mind. One hundred percent of the participants responded with the correct child ethnicity. The manipulation check for gender also showed 100% accuracy.

Means and standard deviations for FSIQ, VIQ, PIQ, index and subtest scores are given in Table 2 for gifted protocols and Table 3 for normal protocols. Full Scale IQ scores ranged from 127.0 – 144.0 for gifted protocols and 100.0 – 114.0 for normal protocols. The mean score for FSIQ for gifted protocols was 133.7, which is in the very superior range and would qualify a child for gifted services. Mean scores for Verbal IQ

(130.6) and Performance IQ (131.6) were also in the very superior range. The mean FSIQ score for the normal protocols was 108.3, which is in the average range. Mean scores for Verbal IQ (107.4) and Performance IQ (108.6) were also in the average range.

In order to test hypothesis 1, three 2 (Caucasian and African American) by 2 (gifted and non-gifted) ANOVAs were conducted to determine whether there was an interaction effect of child ethnicity and gifted status on FSIQ, VIQ, and PIQ scores. Child ethnicity was significant for FSIQ ($F(1,68) = 5.908$; $p < .05$), with Caucasian protocols ($M = 134.2$, $SD = 3.51$) being scored higher than African American protocols ($M = 133.0$, $SD = 2.03$), however no interaction effect of protocol type and child ethnicity was found for FSIQ ($F(1,68) = 0.930$; $p = .34$). Child ethnicity was also significant for VIQ ($F(1, 68) = 4.310$; $p < .05$), with Caucasian protocols ($M = 109.4$, $SD = 3.34$) being scored higher than African American protocols ($M = 106.7$, $SD = 4.03$), however no interaction effect of protocol type and child ethnicity was found for VIQ ($F(1,68) = 1.485$; $p = 0.23$). No interaction effect was found for PIQ ($F(1,68) = 0.244$; $p = 0.62$) and child ethnicity was also not significant for PIQ ($F(1,68) = 1.607$; $p = 0.21$). In other words, for FSIQ and VIQ, identical protocols for Caucasian Americans were scored higher than for African Americans.

In order to test hypothesis 2, thirteen (each subtest) 2 (gifted and non-gifted) by 2 (African American and Caucasian) ANOVAs were conducted to determine if there were interaction effects between ethnicity and protocol type on each of the subtest scores. Results of the ANOVAs can be found in Table 4. An interaction effect was found

between child ethnicity and protocol type for the arithmetic subtest only. Caucasian American protocols were scored significantly higher when the protocol was normal than if it were a gifted protocol. This finding is counter to what was predicted. Scores were significantly higher for Caucasian Americans than African Americans for the information subtest, which means that given identical protocols, Caucasian Americans were scored significantly higher than African Americans on this subtest. There were significant main effects for protocol for nearly all subtests. Only the symbol search subtest and the mazes subtest did not have a significant main effect for protocol.

Post Hoc analyses were done to determine if there was an interaction effect of child ethnicity and gifted status on index test scores. Four (Verbal Comprehension, Perceptual Organizational, Freedom from Distractibility, and Processing Speed) 2 (African American and Caucasian) by 2 (gifted and normal protocols) ANOVAs were conducted to determine if there were interaction effects of protocol type and child ethnicity among the index scores. Results of the analyses can be found in Table 5. An interaction effect was found between child ethnicity and protocol type for the freedom from distractibility and processing speed indexes. Depending on the index score, the results were a little different. For the freedom from distractibility index, scores were higher for Caucasian Americans than African Americans on the normal protocols. However, for the processing speed index, scores were higher for Caucasian Americans than African Americans on the gifted protocols. There were no main effects for ethnicity alone.

In order to test hypothesis 3, there needed to be adequate representations of both Caucasian and African American participants who scored protocols for the Caucasian and African American vignettes. There were only three African American participants. Due to the limited number of participants from ethnic groups other than Caucasian American, no one way ANOVA was conducted to determine overall ethnic differences in scoring.

Discussion

This study was conducted to examine scoring practices of examiners based on the child client's ethnicity. It was hypothesized that gifted protocols would be more differentially scored than normal protocols according to ethnicity. This result was not found. Participants consistently rated African American protocols lower on Full Scale and Verbal IQ than Caucasian protocols whether they were in the gifted or normal range of intelligence. Although this finding is not consistent with the findings from Kline and Dovidio (1982), it is consistent with a more recent finding by Hodson, Dovidio, and Gaertner (2002). They found that when credentials were consistently strong or weak, there were no perceived differences in response patterns for black versus white applicants for admission to a university. However, when credentials were mixed and ambiguous, differences in response patterns were found. This finding would suggest that because the gifted vignettes used in this study were strong, no differences between protocols would be found. However, if the nature of the referral were more ambiguous, perhaps differences in protocols would have been more evident. Therefore future researchers in this area may want to concentrate on using more ambiguous, "normal" protocols rather than ones in the extreme ranges.

It was also hypothesized that subtests that were shown previously to have lower inter-rater reliability (Similarities, Comprehension and Vocabulary; Cuenot & Darbes, 1982) would show higher scores for Caucasian gifted vignettes than the other three vignettes. The results of this study do not support that hypothesis. The comprehension, similarities, and vocabulary subtests did not show a significant difference between any of the vignette types. However, the information and arithmetic subtests showed a main effect for child ethnicity and the arithmetic subtest also showed an interaction effect between child ethnicity and protocol type. While these subtests have moderate inter-rater reliability, they

have lower test-retest reliability scores than the other subtests (Sattler, 2002). There were no apparent differences between cells that could potentially explain the differences found. One potential explanation for the differences found in these scores could be the area that these scores tap into. Information and Arithmetic are most consistent with school learning. Raters could have a preconceived notion of a difference in ability to achieve in school based on ethnicity.

Post hoc analyses were conducted to determine if there were interaction effects between child ethnicity and protocol type for the index scores (verbal comprehension, perceptual organizational, freedom from distractibility, and processing speed). Both freedom from distractibility and processing speed showed significant interaction effects. None of the indexes showed a significant main effect for ethnicity. These effects could suggest that raters could have preconceived notions of the speed with which African Americans process visually perceived nonverbal information and the ability to sustain attention, concentrate and exert mental control. Specifically, the significant interaction effects for freedom from distractibility and processing speed were consistent with aversive racism theory.

The results of this study indicated that protocols of perceived African Americans were scored significantly lower than those for perceived Caucasians. First of all, these results suggest that, when given the same protocol, the emerging professionals in the field on average will score the protocols differently based on the ethnicity of the child examinee. This finding is consistent with the theory by Dovidio and Gaertner (1986) who suggested that while more overt forms of racism are on the decline, there is a more covert form termed "aversive" which can be found in those with a strong desire to be non-prejudiced. Based on professional standards in Psychology, one would hope that graduate students and other professionals have a strong desire to be non-prejudiced. Although aversive racism was initially hypothesized to be reflected in higher ratings of gifted protocols for Caucasian

Americans, the current findings suggest a more global racism regardless of gifted status versus non-gifted status.

Although matching examiners and examinees was not considered in this study, it would be interesting to see if ethnic matching would play a role in scoring practices. This notion has been studied previously, however, the assessment literature on the match between examiner and examinee has been extremely contradictory and riddled with methodological problems (Sattler & Gwynn, 1982; Graziano, Varca, & Levy, 1982). Future studies examining this issue are recommended. Also, future studies could examine whether previous training in culture and ethnicity had an effect on the scoring practices of examiners. It could be that, with training in cultural diversity, biases in scoring might not appear. Training programs across the country could address culture and diversity in their curriculum more extensively and pay special attention to the training of standardized test scoring (Hertzsprung & Dobson, 2000).

The issue of statistical versus clinical significance is important to consider for the results of this study. Although the mean difference found between Caucasian protocols and African American protocols were significant, they were also small. One possible reason for such a small difference could be that the majority of participants in this study were Caucasian. If there were more African American participants, a larger discrepancy might have been found due to a possibility of more variance in scoring. However, if we consider the current results, what would be the implications of this difference in the real world? The difference obtained does fall into the range of the standard error of measurement for the WISC-III, which is on average about 3.20 (Sattler, 2002). However, this is a consistent discrepancy. This issue becomes extremely important when it comes to the issue of qualification for services. On the one hand, children who may be on the borderline for receiving gifted services could be denied those services because they did not quite meet

criteria based on scoring of the WISC-III. This pattern could lead to less African Americans receiving those services and possibly not being as academically enriched. Also, children who may be tested for a learning disability may not meet criteria because their scores are not high enough on the WISC to find a significant difference between intelligence test scores and achievement test scores (Shepard et al., 1983). These issues deserve further empirical attention.

This study had several limitations. The biggest limitation of this study is the sample size and response rate. Only 72 out of 700 of the protocols distributed were returned. There were several areas of potential breakdown in distribution. First of all, protocols were sent to directors of graduate programs instead of directly to graduate students themselves. If the directors did not distribute them, the graduate students could not have returned them. Also, graduate students themselves may not have had the time or inclination to participate in the study. Scoring of the WISC-III could take between fourty-five minutes to one hour. This commitment could have been too time intensive. There was no method of direct contact with the graduate students for follow up and so the study could have been forgotten easily. One method that could be attempted in future studies would be to contact the directors by phone to emphasize the importance of the study. Also, contacting graduate students through a listserv directly could increase response rate.

Another limitation of this study was the limited diversity of individuals who responded to the study. In future studies, more efforts could be given to recruit members of various ethnic groups to examine the match between ethnicity of examiner and examinee.

There have been many theories about cultural differences found between groups on intelligence test scores. One explanation has been that there could be differential effects due to the assessor. Although there have been several studies that have considered this possibility, the results of those studies are inconclusive. This study attempted to focus

on biases in the assessor alone and eliminate effects from the environment and test taker (such as genetic predispositions). It was found that Caucasian American protocols were scored higher than African American protocols on Full Scale IQ as well as Verbal IQ. Although the findings of this study are somewhat disturbing, they suggest that biases in assessment procedures may be an important topic for future study and remediation.

References

Anastasi, A. (1988). *Psychological Testing, 6th Ed*. (pp. 228-232). New York: Macmillan Publishing Company.

Angoff, W.H. (1988). The nature-nurture debate, aptitudes and group differences. *American Psychologist*, 43(9), 713-720.

Banks, J.A. (1988). Ethnicity, class, cognitive, and motivational styles: Research and teaching implications. *Journal of Negro Education*, 57(4), 452-466.

Banks, W.C., McQuater, G.V., and Sonne, J.L. (1995). A deconstructive look at the myth of race and motivation. *Journal of Negro Education*, 64(3), 307-325.

Bellugi**,** U.**,** Lichtenberger, L., Jones, W., Lai, Z., and St. George, M.I. (2000). The neurocognitive profile of Williams syndrome: A complex pattern of strengths and weaknesses. *Journal of Cognitive Neuroscience*, 12, 2000.

Braden, J. P. (1999). Straight talk about assessment and diversity: what do we know? *School Psychology Quarterly*, 14(3), 343-355.

Brooks-Gunn, J. and Klebanov, P.K. (1996). Ethnic Differences in children's intelligence test scores: role of economic deprivation, home environment, and maternal characteristics. *Child Development*, 67, 396-408.

Castenell, L. (1984). A cross-cultural look at achievement motivation research. *Journal of Negro Education*, 53(4), 435-443.

Chan, D., Schmitt, N., DeShon, R.P., Clause, C.S. et.al. (1997). Reactions to cognitive ability tests: The relationship between race, test performance, face validity perceptions, and test-taking motivation. *Journal of Applied Psychology*, 82(2), 300-310.

Clarizio, H. F. (1978). Nonbiased assessment of minority group children. *Measurement and Evaluation in Guidance*, 11(2), 106-113.

Cohen, J. (1992). A power primer. *Psychological Bulletin,* 112(1), 155-159.

Cuenot, R.G. & Darbes, A. (1982). A comparison of interscorer agreement for the comprehension, similarities, and vocabulary subtests of the WISC and WISC-R. *Educational and Psychological Measurement,* 42, 417-421.

Dickens, W.T., and Flynn, J.R. (2001). Heritability estimates versus large environmental effect: The IQ paradox resolved. *Psychological Review*, 108(2), 346-369.

Dovidio**,** J.F and Gaertner, S.L. (2000). Aversive racism and selection decisions: 1989 and 1999. *Psychological Science,* 11(4), 315-319.

Dovidio**,** J.F and Gaertner, S.L. (1998). On the nature of contemporary prejudice: The causes, consequences, and challenges of aversive racism. In Eberhardt, J.L. and Fiske, S.T. (Eds). *Confronting racism: The problem and the response. (pp. 3-32).* Thousand Oaks: Sage Publications, Inc.

Dovidio**,** J.F and Gaertner, S.L. (1991). Changes in the expression and assessment of racial prejudice. In Knopke, H.J., Norrell, R.J., et al. (Eds). *Opening doors: Perspectives on race relations in contemporary America. (pp. 119-148).* Birmingham: The University of Alabama Press.

Downey, D.B. (2001). Number of Siblings and Intellectual development: The resource dilution explanation. *American Psychologist*, 56(6/7), 497-504.

Gaertner, S.L and Dovidio**,** J.F. (1986). The aversive form of racism. In Dovidio**,** J.F. and Gaertner, S.L. (Eds). *Prejudice, discrimination, and racism. (pp. 61-89).* San Diego: Academic Press, Inc.

Geller**,** J.D. (1988). Racial bias in the evaluation of patients for psychotherapy. In L.Comas-Diaz and E. H. Griffith, (Eds). *Clinical guidelines in cross-cultural mental health. Wiley series in general and clinical psychiatry. (Pp. 112-134).* New York: John Wiley & Sons.

Graziano, W.G., Varca, P.E., and Levy, J.C. (1982). Race of examiner effects and the validity of Intelligence tests. *Review of Educational Research*, 52(4), 469-497.

Grigorenko**,** E.L. (2000). Heritability and intelligence. In R.J. Sternberg (Ed). *Handbook of Intelligence.* New York: Cambridge University Press.

Gutkin**,** T.B and Reynolds, C.R. (1981). Factorial similarity of the WISC-R for White and Black children from the standardization sample. *Journal of Educational Psychology*, 73(2), 227-23.

Hall, J., Guterman, D.K., Lee, H.B., & Little, S.G. (2002). Counselor-client matching on ethnicity, gender, and language: Implications for counseling school-aged children. *North American Journal of Psychology,* 4(3), 367-380.

Hernstein, R.J. and Murray, C. (1994). *The Bell Curve: Intelligence and class structure in American life*. New York: Simon & Schuster.

Hertzsprung, E.A. and Dobson, K.S. (2000). Diversity Training: Conceptual issues and practices for Canadian clinical psychology programs. *Canadian Psychology,* 41(3), 184-191.

Hickman, J.A. and Reynolds, C. R. (1986-87). Are race differences in mental test scores an artifact of psychometric methods? A test of Harrington's experimental model. *The Journal of Special Education*, 20(4), 409-430.

Hodson, G., Dovidio, J.F., and Gaertner, S.L. (2002). Processes in racial discrimination: Differential weighting of conflicting information. *Personality & Social Psychology Bulletin,* 28(4), 460-471.

Jenkins-Hall**,** K and Sacco, W.P. (1991). Effect of client race and depression on evaluations by White therapists. *Journal of Social & Clinical Psychology,* 10(3), 322-333.

Jensen, A.R. (1969). How much can we boost IQ and scholastic achievement? *Harvard Educational Review*, 39, 1-123.

Jones, R.T. and Herndon, C. (1992). The status of black children and adolescents in the academic setting: assessment and treatment issues. In C.E. Walker & M.C. Roberts (Eds.), *Handbook of Clinical Child Psychology, 2nd Ed*., Wiley Series on Personality Processes (pp. 901-917). New York: John Wiley & Sons.

Kamphaus, R.W. (2001). *Clinical Assessment of Child and Adolescent Intelligence, 2nd Ed*. (pp.81-85). Boston: Allyn & Bacon.

Kaufman, A.S. and Lichtenberger, E.O. (1999). *Essentials of WAIS-III Assessment.* New York: John Wiley & Sons.

Kaufman, A.S. (2000). Tests of Intelligence. In R. Sternberg (Ed.), *Handbook of Intelligence*. New York: Cambridge University Press.

Lynn, R. (1995). Cross-cultural differences. In D. Saklofske & M. Zeidner (Eds.), *International Handbook of Personality and Intelligence* (pp. 108-121). New York: Plenum Press.

Lynn, R. (1998). Has the black-white intelligence difference in the United States been narrowing over time? *Personality and Individual Differences*, 25, 999-1002.

MacLullich, A.M.J., Seckl, J.R., Starr, J.M., and Deary, I.J. (1998). The biology of intelligence: From Association to Mechanism. *Intelligence*, 26(2), 63-73.

Malgady, R.G. and Constantino, G. (1998). Symptom severity in bilingual Hispanics as a function of clinician ethnicity and language of interview. *Psychological Assessment*, 10(2), 120-127.

Matte, T.D., Bresnahan, M., Begg, M.D., and Susser, E. (2001). Influence of variation in birth weight within normal range and within sibships on IQ at age 7 years: cohort study. *BMJ*, 323, 310-314.

McGue, M., Bouchard, T.J., Iacono, W.G. and Lykken, D.T. (1993). Behavioral genetics of cognitive ability: A life-span perspective. In R.Plomin and G. McClearn (Eds.), *Nature, Nurture, and Psychology, 1st Ed*., Washington, DC: American Psychological Association.

Mishra, Shitala P. (1980). The influence of examiners' ethnic attributes on intelligence test scores. *Psychology in the Schools*, 17(1), 117-122.

Mishra, S.P. (1983). Effects of examiners' prior knowledge of subjects' ethnicity and intelligence on the coring of responses to the Stanford-Binet scale. Psychology in the Schools, 20, 133-136.

Moore, E.G.J. (1986). Family socialization and the IQ test performance of traditionally and transracially adopted black children. *Developmental Psychology*, 22(3), 317-326.

Neisser, U. (1996). Intelligence: Knowns and Unknowns. *American Psychologist*, 51(2), 77-101.

Oakland**,** T. and Feigenbaum, D. (1979). Multiple sources of test bias on the WISC-R and Bender-Gestalt Test. *Journal of Consulting & Clinical Psychology,* 47(5), 968-974.

Oakland, T., and Glutting, J.J. (1990). Examiner observations of children's WISC-R test-related behaviors: possible socioeconomic status, and gender effects. *Psychological Assessment*, 2(1), 86-90.

Plomin, R. and Petrill**,** S.A. (1997). Genetics and intelligence: What's new? *Intelligence,* 24(1), 53-77.

Ramey, C.T., Campbell, F.A. and Ramey, S.L. (1999). Early intervention: Successful pathways to improving intellectual development. *Developmental Neuropsychology*, 16(3), 385-392.

Reschly**,** D.J. (1978). WISC-R factor structures among Anglos, Blacks, Chicanos, and Native-American Papagos. *Journal of Consulting & Clinical Psychology,* 46(3), 417-422.

Reynolds, C.R., Lowe, P.A., and Saenz, A.L. (1999). The problem of bias in psychological assessment. In C. Reynolds & T. Gutkin (Eds.). *The Handbook of School Psychology, 3rd ed.*, New York: John Wiley & Sons.

Rodgers, J.L. (2001). What causes birth order intelligence patterns? *American psychologist*, 56 (6/7), 505-510.

Rushton**,** J. P. (1992). Contributions to the history of psychology: XC. Evolutionary biology and heritable traits (with reference to Orientalhite-sub (lack differences): The 1989 AAAS paper. *Psychological Reports*, 71(3, Pt 1), 811-821.

Sattler, J.M., Andres, J.R., Squire, L.S., Wisley, R., and Maloy, C.F. (1978). Examiner scoring of ambiguous WISC-R responses. *Psychology in Schools,* 15, 486-489.

Sattler, J.M., and Gwynne, J. (1982). White examiners generally do not impede the Intelligence test performance of black children: to debunk a myth. *Journal of Consulting and Clinical Psychology*, 50(2), 196-208.

Sattler, J.M. (2001). *Assessment of Children: Cognitive Applications* (4th ed.). San Diego: Sattler Publishing.

Segall, M.H., Dasen, P.R., Berry, J.W., and Poortinga, Y.H. (1990). Human behavior in the global perspective: An introduction to cross-cultural psychology. New York: Pergamon Press.

Shepard, L., Smith, M., and Vojir, C. (1983). Characteristics of pupils identified as learning disabled. *American Educational Research Journal,* 20(3), 309-331.

Sigman, M. and Whaley, S.E. (1998). The role of nutrition in the development of intelligence. . In Neisser, U. (Ed). *The rising curve: Long-term gains in IQ and related measures*. Washington, DC*:* American Psychological Association.

Skuy, M., Taylor, M., O'Carroll, S., Fridjhon, P., and Rosenthal, L. (2000). Performance of black and white South African children on the Weschler Intelligence Scale for

Children – Revised and the Kaufman Assessment Battery. *Psychological Reports*, 86, 727-737.

Sue, S., Fujino, D., Hu, L., Takeuchi, D., & Zane, N. (1991). Community mental health services for ethnic minority groups: A test of the cultural responsiveness hypothesis. *Journal of Clinical and Consulting Psychology,* 59, 533-540.

Sue**,** S. (1998). In search of cultural competence in psychotherapy and counseling. *American Psychologist,* 53(4), 440-448.

Terrell, F. and Terrell, S.L. (1983). The relationship between race of examiner, cultural mistrust, and the Intelligence test performance of black children. *Psychology in the Schools*, 20, 367-369.

van Ryn**,** M. and Burke, J. (2000). The effect of patient race and socio-economic status on physicians' perceptions of patients. *Social Science & Medicine,* 50(6), 813-828.

Vincent, K.R. (1991). Black/White differences: does age make the difference? *Journal of Clinical Psychology*, 47(2), 266-270.

Weiss, L.G and Prifitera**,** A**.** (1995). An evaluation of differential prediction of WIAT achievement scores from WISC-III FSIQ across ethnic and gender groups. *Journal of School Psychology,* 33(4), 297-304.

Weschler, D. (1991). *Weschler Intelligence Scale for Children* (3rd ed.). San Antonio, TX: Psychological Corporation.

Whaley, A.L. (1998). Racism in the provision of mental health services: a social-cognitive approach. *American journal of Orthopsychiatry*, 68(1), 47-57.

Zajonc, R.B. (2001). The family dynamics of intellectual development. *American psychologist*, 56 (6/7), 490-496.

Zimmerman, I. and Woo-Sam, J. (1997). Review of the Criterion-related validity of the WISC-III: The first 5 years. *Perceptual and Motor Skills*, 85, 531-546.

Appendices

Appendix A: Child Vignette

Jason is a 10 year, 4 month-old Caucasian boy who lives at home with his mother, father and younger sister. He attends 5th grade. Jason was brought for testing by his mother who reported that Jason has always been a little restless and inquisitive. She reported that by the time he entered kindergarten, Jason was able to read and write fairly well. She also noted that he has always been easily frustrated by puzzles and games that require sustained efforts.

Jason was referred for testing by the elementary school counselor due to his inconsistent academic performance. The counselor reports that Jason starts the school year earning all A's in his classes for the first 9-week assessment period. These grades fall to C's and D's for the remainder of the school year. This pattern has been present since second grade. Jason's teacher is not sure what is wrong and suggests that he be evaluated.

Jamal's teacher thinks that perhaps he is gifted and therefore gets bored during the year. She is suggesting that he be evaluated for the gifted program.

*__Note:__ This is the only sentence that will change in the vignette based on gifted versus non-gifted group.

Appendix B: Letter to Directors of Training Programs

Dear Director of Clinical Training:

I am conducting a study (for my Master's Thesis) to determine the possible variations in scoring and interpreting the WISC-III, and I am asking for your help. My goal is to send WISC-III protocols to graduate students in doctoral programs across the country (clinical PhD and PsyD, and school psychology programs). All returned protocols will be treated in a confidential manner.

Enclosed you will find 4 packets to distribute to graduate students in your program who have been trained in administration and scoring of the WISC-III and have had at least one year of experience. Each packet includes the following: an introductory letter to the student, a WISC-III protocol, a vignette describing the child that was administered the test, instructions for scoring the WISC-III, a brief demographic questionnaire, and a self-addressed stamped envelope. Although the method of distribution is completely at your discretion, it is my hope to have a good representation of diversity in my sample, and anything you could do to aid in this goal would be greatly appreciated. Students will send the protocols directly to me in the enclosed self-addressed, stamped envelope.

This study has been approved by the USF Institutional Review Board.

If you have any questions regarding this study, or would like further information, please do not hesitate to contact me.

Telephone: (813) 974-9222
Email: fieldss@helios.acomp.usf.edu
Address: Department of Psychology
University of South Florida
4202 E. Fowler Avenue
PCD 4118G
Tampa, FL 33620

Thank you for your cooperation.

Sherecce A. Fields
Graduate Student
Department of Psychology
University of South Florida

Appendix C: Letter to Participants

To Whom It May Concern:

I am conducting a study (for my master's Thesis) to determine the possible variations in scoring and interpreting the WISC-III, and I am asking for your participation. You are being asked to participate because you are a doctoral graduate student who has been trained in the administration and scoring of the WISC-III and have had at least one year of experience. My goal is to send WISC-III protocols to graduate students in doctoral programs across the country (clinical PhD and PsyD, and school psychology) to get a good representation of scorers and interpreters of the test. All returned materials will be treated in a confidential manner.

Enclosed you will find the following: a WISC-III protocol, a vignette describing the child that was administered the test, instructions for scoring the WISC-III, a brief demographic questionnaire, and a self-addressed stamped envelope. You are being asked to do the following: read the vignette about the child who was administered the test, score the WISC-III protocol, and complete the demographics questionnaire. The scored protocols and demographics questionnaire should be returned directly to me in the enclosed self-addressed, stamped envelope. Completion of all required items is estimated to take less than an hour.

All participants who return a completed packet will be entered into one of two drawings for $100. If you would like to be entered in the drawing, please include an index card with your name and address. In order to protect confidentiality, these cards will be separated from your completed measures immediately upon receipt.

Your privacy and research records will be kept confidential to the extent of the law. Authorized research personnel, employees of the Department of Health and Human Services and the USF Institutional Review Board may inspect the records from this research project.

The results of this study may be published. However, the data obtained from you will be combined with data from other people in the publication. The published results will not include your name or any other information that would in any way personally identify you. All records will be identified by numbers, and all access to the data will be restricted to students and faculty of the Psychology department at the University of South Florida.

Your decision to participate in this research study is completely voluntary. You are free to participate in this research study or to withdraw at any time.

By completing and returning the protocol and demographics questionnaire, you are giving your consent to participate in this research study.

This research project/study was reviewed and approved by the University of South Florida Institutional Review Board for the protection of human subjects. This approval is valid until the date provided below. The board may be contacted at (813) 974-5638.
If you have any questions regarding this study, or would like further information, please do not hesitate to contact me.

Telephone: (813) 974-9222
Email: fieldss@helios.acomp.usf.edu
Address: *Department of Psychology*
University of South Florida
4202 E. Fowler Avenue
PCD 4118G
Tampa, FL 33620

Thank you for your cooperation.

Sherecce A. Fields
Graduate Student
Department of Psychology
University of South Florida

Appendix D: Demographics Questionnaire

1. **Your Gender**: 1. Male
   2. Female

2. **Your Race**: 1. African American
   2. Caucasian
   3. Hispanic
   4. Asian
   5. Other

3. **Your Age**: _____

4. **Your Current Training Program**: 1. PhD Clinical Psychology
   2. PsyD Clinical Psychology
   3. PhD School Psychology

5. **Your Year in Graduate School**: _____

6. **Months of Experience you have had with the WISC-III**: _____

7. **Months of direct experience you have had with children's or adolescent's clinical issue (e.g. conducting assessments, conducting therapy, etc.).** _____

8. **Questions about the child protocol you've scored**:
   a. What was your child's gender? _____
   b. What was the race of your child? _____
   c. What was the age of your child? _____
   d. Would you think that this child could meet criteria for ADHD? **Yes**
      No
Don't Know

   e. Would you think that this child could meet criteria for giftedness? **Yes**
      No
Don't Know

   f. Would you think that this child could meet criteria for a Learning Disorder?
      Yes
      No
      Don't Know

Thank you very much for participating in this study. Because others in your program might be participating as well, I am asking that you please do not discuss the protocol until after they have completed it.

Tables

Table 1. Participant Characteristics by Vignette Type

	Gifted Caucasian (N = 24)	**Gifted African American (N = 17)**	**Non-gifted Caucasian (N = 19)**	**Non-gifted African American (N = 12)**
Gender				
Male	7	0	3	2
Female	17	17	16	10
Race/Ethnicity				
African American	1	0	2	0
Caucasian	21	14	16	8
Hispanic	1	2	1	2
Asian	1	0	0	2
Other	0	1	0	0
Mean Age	30.46	26.47	27.32	26.50
(Standard Deviation)	7.06	3.36	3.94	2.39
Mean Months of Experience	34.25	22.06	32.21	20.92
(Standard Deviation)	28.44	17.48	31.54	16.97

Table 2. Means and Standard Deviations of IQ, Index, and Subtest Scores for Gifted Protocols

Test	*Range*	**Overall Mean (Standard Deviation)**	**Mean for African American vignette**	**Mean for Caucasian Vignette**
IQ Scores				
Full Scale (FSIQ)	127.0 – 144.0	133.7 (3.01)	133.0(2.03)	134.2 (3.51)
Verbal (VIQ)	121.0 – 142.0	130.6 (3.90)	130.0 (2.18)	131.0 (4.76)
Performance (PIQ)	123.0 – 139.0	131.6 (2.53)	130.9 (2.37)	132.6 (2.53)
Index Scores				
Verbal Comprehension	114.0 – 140.0	125.5 (4.64)	124.9 (2.15)	126.0 (5.80)
Perceptual Organizational	126.0 – 141.0	131.8 (2.12)	131.6 (1.06)	131.8 (2.65)
Freedom from Distractibility	131.0 – 137.0	133.9 (0.83)	134.0 (0.00)	133.9 (1.08)
Processing Speed	101.0 – 126.0	116.5 (4.00)	114.8 (4.29)	117.6 (3.44)
Subtest Scores				
Picture Completion	12.0 – 14.0	13.05 (0.38)	13.06 (0.24)	13.04 (0.46)
Information	14.0 – 17.0	14.90 (0.74)	14.71 (0.69)	15.04 (0.75)
Coding	7.00– 17.0	12.73 (1.45)	12.24 (1.56)	13.08 (1.28)
Similarities	10.0 – 15.0	12.80 (1.05)	12.88 (0.78)	12.75 (1.22)
Picture Arrangement	18.0 – 19.0	18.90 (0.30)	18.89 (0.33)	18.92 (0.28)
Arithmetic	17.0 – 18.0	17.98 (0.16)	18.00 (0.00)	17.96 (0.20)
Block Design	15.0 – 18.0	17.76 (0.77)	17.82 (0.73)	17.71 (0.81)
Vocabulary	13.0 – 19.0	17.46 (1.23)	17.53 (1.01)	17.42 (1.38)
Object Assembly	10.0 – 17.0	11.10 (1.00)	10.94 (0.24)	11.21 (1.28)
Comprehension	10.0 – 18.0	12.88 (1.71)	12.47 (0.94)	13.17 (2.06)
Symbol Search	13.0 – 17.0	13.15 (0.70)	13.00 (0.00)	13.25 (0.90)
Digit Span	13.0 – 15.0	14.00 (0.23)	14.00 (0.00)	14.00 (0.29)
Mazes	1.00 – 19.0	10.66 (2.56)	10.47 (3.64)	10.78 (1.59)

Note: Standard deviations are in parenthesis. FSIQ, VIQ, PIQ, and index scores are standardized to have a mean of 100 and a standard deviation of 15. Subtest scores are standardized to have a mean of 10 and a standard deviation of 3.

Table 3. Means and Standard Deviations of IQ, Index, and Subtest Scores for Normal Protocols

Test	*Range*	**Overall Mean**	**Mean for African American vignette**	**Mean for Caucasian vignette**
IQ Scores				
Full Scale (FSIQ)	100.0 – 114.0	108.3 (3.80)	106.7 (4.03)	109.4 (3.34)
Verbal (VIQ)	87.00 – 115.0	107.4 (5.99)	105.0 (7.39)	108.8 (4.51)
Performance (PIQ)	102.0 – 116.0	108.6 (2.93)	108.3 (3.11)	108.8 (2.87)
Index Scores				
Verbal Comprehension	96.00 – 118.0	109.5 (5.77)	107.4 (6.44)	110.7 (5.05)
Perceptual Organizational	102.0 – 113.0	108.9 (2.55)	108.4 (1.78)	109.3 (2.94)
Freedom from Distractibility	69.00 – 101.0	94.65 (5.19)	92.67 (7.83)	95.89 (1.76)
Processing Speed	109.0 – 126.0	111.0 (2.95)	111.6 (4.86)	110.7 (0.75)
Subtest Scores				
Picture Completion	9.00 – 13.0	10.65 (0.95)	10.33 (0.98)	10.80 (0.90)
Information	10.0 – 14.0	12.23 (1.09)	11.67 (0.98)	12.58 (1.02)
Coding	10.0 – 17.0	10.87 (0.48)	11.00 (1.95)	10.79 (0.42)
Similarities	11.0 – 13.0	12.29 (0.74)	12.33 (0.78)	12.26 (0.73)
Picture Arrangement	7.00 – 12.0	10.58 (0.89)	10.50 (0.52)	10.63 (1.07)
Arithmetic	1.00 – 11.0	9.45 (1.65)	8.75 (2.49)	9.89 (0.46)
Block Design	7.00 – 12.0	11.00 (1.00)	10.92 (1.31)	11.05 (0.78)
Vocabulary	7.00 – 14.0	11.97 (1.68)	11.33 (2.06)	12.37 (1.30)
Object Assembly	10.0 – 16.0	13.10 (0.98)	13.17 (0.94)	13.05 (1.03)
Comprehension	5.00 – 13.0	10.16 (2.16)	9.83 (2.41)	10.37 (2.03)
Symbol Search	13.0 – 14.0	13.07 (0.25)	13.09 (0.30)	13.05 (0.23)
Digit Span	8.00 – 10.0	8.13 (0.50)	8.17 (0.58)	8.10 (0.46)
Mazes	6.00 – 11.0	10.67 (1.21)	10.64 (1.21)	10.69 (1.25)

Note: Standard deviations are in parenthesis. FSIQ, VIQ, PIQ, and index scores are standardized to have a mean of 100 and a standard deviation of 15. Subtest scores are standardized to have a mean of 10 and a standard deviation of 3.

Table 4. Analysis of Variance for Subtest Scores

Subtest	Source	df	F	p
Picture Completion				
	Child Ethnicity	1	2.235	0.14
	Protocol	1	224.3*	0.00
	Protocol X Child Ethnicity	1	2.558	0.11
	Error	68		
Information				
	Child Ethnicity	1	8.998*	0.004
	Protocol	1	174.9*	0.00
	Protocol X Child Ethnicity	1	1.920	0.17
	Error	68		
Coding				
	Child Ethnicity	1	0.958	0.33
	Protocol	1	29.35*	0.00
	Protocol X Child Ethnicity	1	2.640	0.11
	Error	68		
Similarities				
	Child Ethnicity	1	0.195	0.66
	Protocol	1	5.091*	0.03
	Protocol X Child Ethnicity	1	0.018	0.89
	Error	68		
Picture Arrangement				
	Child Ethnicity	1	0.293	0.59
	Protocol	1	2957*	0.00
	Protocol X Child Ethnicity	1	0.101	0.75
	Error	68		
Arithmetic				
	Child Ethnicity	1	4.794*	0.03
	Protocol	1	1181*	0.00
	Protocol X Child Ethnicity	1	5.545*	0.02
	Error	68		
Block Design				
	Child Ethnicity	1	0.002	0.96
	Protocol	1	992.6*	0.00
	Protocol X Child Ethnicity	1	0.340	0.56
	Error	68		
Vocabulary				
	Child Ethnicity	1	1.784	0.19
	Protocol	1	265.1*	0.00
	Protocol X Child Ethnicity	1	2.762	0.10
	Error	68		

* $p < .05$

Table 4 continued. Analysis of Variance of Subtest Scores

Subtest	Source	df	F	p
Object Assembly				
	Child Ethnicity	1	0.100	0.75
	Protocol	1	70.55*	0.00
	Protocol X Child Ethnicity	1	0.619	0.43
	Error	68		
Comprehension				
	Child Ethnicity	1	1.746	0.19
	Protocol	1	34.03*	0.00
	Protocol X Child Ethnicity	1	0.030	0.86
	Error	68		
Symbol Search				
	Child Ethnicity	1	0.587	0.45
	Protocol	1	0.148	0.70
	Protocol X Child Ethnicity	1	1.088	0.30
	Error	66		
Digit Span				
	Child Ethnicity	1	0.111	0.74
	Protocol	1	4058	0.00
	Protocol X Child Ethnicity	1	0.111	0.74
	Error	67		
Mazes				
	Child Ethnicity	1	0.112	0.74
	Protocol	1	0.005	0.95
	Protocol X Child Ethnicity	1	0.058	0.81
	Error	61		

* $p < .05$

Table 5. Analysis of Variance for Index Scores

Index	Source	df	F	p
Verbal Comprehension				
	Child Ethnicity	1	3.210	0.08
	Protocol	1	174.6*	0.00
	Protocol X Child Ethnicity	1	0.791	0.38
	Error	68		
Perceptual Organizational				
	Child Ethnicity	1	0.829	0.37
	Protocol	1	1631*	0.00
	Protocol X Child Ethnicity	1	0.339	0.56
	Error	68		
Freedom from Distactibility				
	Child Ethnicity	1	3.549	0.06
	Protocol	1	2318*	0.00
	Protocol X Child Ethnicity	1	4.144*	0.05
	Error	67		
Processing Speed				
	Child Ethnicity	1	1.160	0.29
	Protocol	1	34.32*	0.00
	Protocol X Child Ethnicity	1	4.752	0.03
	Error	66		

* $p < .05$