NRC G/T

THE NATIONAL RESEARCH CENTER ON THE GIFTED AND TALENTED

Senior Scholars Series

University of Connecticut
University of Virginia
Yale University

Intelligence Testing and Cultural Diversity: Concerns, Cautions, and Considerations

Donna Y. Ford
Vanderbilt University
Nashville, Tennessee

December 2004
RM04204

Intelligence Testing and Cultural Diversity: Concerns, Cautions, and Considerations

Donna Y. Ford
Vanderbilt University
Nashville, Tennessee

December 2004
RM04204

THE NATIONAL RESEARCH CENTER ON THE GIFTED AND TALENTED

The National Research Center on the Gifted and Talented (NRC/GT) is funded under the Jacob K. Javits Gifted and Talented Students Education Act, Institute of Education Sciences, United States Department of Education.

The Directorate of the NRC/GT serves as an administrative and a research unit and is located at the University of Connecticut.

The participating universities include the University of Virginia and Yale University, as well as a research unit at the University of Connecticut.

University of Connecticut
Dr. Joseph S. Renzulli, Director
Dr. E. Jean Gubbins, Associate Director
Dr. Sally M. Reis, Associate Director

University of Virginia
Dr. Carolyn M. Callahan, Associate Director

Yale University
Dr. Robert J. Sternberg, Associate Director

Copies of this report are available from:
NRC/GT
University of Connecticut
2131 Hillside Road Unit 3007
Storrs, CT 06269-3007

Visit us on the web at:
www.gifted.uconn.edu

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R206R000001, as administered by the Institute of Education Sciences, U.S. Department of Education. The findings and opinions expressed in this report do not reflect the position or policies of the Institute of Education Sciences or the U.S. Department of Education.

Note to Readers...

All papers by The National Research Center on the Gifted and Talented may be reproduced in their entirety or in sections. All reproductions, whether in part or whole, should include the following statement:

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R206R000001, as administered by the Institute of Education Sciences, U.S. Department of Education. The findings and opinions expressed in this report do not reflect the position or policies of the Institute of Education Sciences or the U.S. Department of Education.

This document has been reproduced with the permission of The National Research Center on the Gifted and Talented.

If sections of the papers are printed in other publications, please forward a copy to:

The National Research Center on the Gifted and Talented
University of Connecticut
2131 Hillside Road Unit 3007
Storrs, CT 06269-3007

Please Note: Papers may not be reproduced by means of electronic media.

Intelligence Testing and Cultural Diversity: Concerns, Cautions, and Considerations

Donna Y. Ford
Vanderbilt University
Nashville, Tennessee

ABSTRACT

At all levels of education, there is great concern about the low performance of racially and linguistically diverse students—African Americans, Hispanic Americans, and Native Americans—on standardized tests, as well as their under-representation in gifted education. While fewer concerns and criticisms target achievement tests, a wealth of controversy surrounds intelligence tests (also known as cognitive ability tests), specifically given the consistently lower performance of Black students on intelligence tests compared to White students. More so than with achievement tests, intelligence tests carry the burden of being associated with innate ability, particularly by laypersons and those unfamiliar with the purposes and limitations of tests; thus, to those unfamiliar with the purposes and limitations of tests, when one group performs lower than another group, the results, they believe, may be attributed to heredity or genetic inferiority. This simplistic explanation ignores the role of environment, including education and opportunity to learn, on students' test performance.

Issues regarding achievement tests and diverse students are less controversial than those regarding intelligence tests. Compared to intelligence tests, few publications have been written regarding biases in achievement tests. Performance on achievement tests is generally associated with the quality and quantity of students' educational or learning experiences at home and school. For the most part, low achievement test scores are associated with poor educational experiences, lack of motivation, and a host of other factors that tend to be environmental or social rather than inherited or genetic. Conversely, some people presume that intelligence tests measure unlearned abilities—abilities less dependent on instruction and education—and they interpret low performance on intelligence tests with low cognitive ability and potential. This belief is particularly relevant among: (a) individuals who are untrained in testing and assessment; (b) individuals who believe that intelligence is fixed, innate, and unchangeable, and (c) individuals who believe that intelligence tests are comprehensive, exact, and precise measures of intelligence (see discussion in Groth-Marnat, 1997, 2003). Whatever position one holds regarding the nature of intelligence (and achievement) as measured by tests, these tests: (a) measure only a sample of the construct being measured; (b) measure present behavior, namely students' attainment of skills at the time of assessment; and (c) intelligence test scores are an estimate of a person's current level of functioning as measured by the various tasks required.

Attempts to develop an accurate definition and measure of "intelligence" have been fraught with difficulty and controversy. Nowhere are the debates and controversies

surrounding intelligence more prevalent than in gifted education and special education. These two educational fields rely extensively on tests to make educational and placement decisions. In gifted education, low test scores often prevent diverse students from being identified as gifted and receiving services; in special education, low test scores often result in identifications such as learning disabled, mentally retarded, and so forth. Racially and linguistically diverse students (African Americans, Hispanic Americans, and Native Americans) are under-represented in gifted education and over-represented in special education (see Council of State Directors of Programs for the Gifted and National Association for Gifted Children, 2003; U.S. Department of Education, 2003). Ford (1998), Frasier, García, and Passow (1995), and others reported that Black, Hispanic, and Native American students have always been under-represented in gifted education programs.

There are two persistent, major debates or controversies surrounding minority students' intelligence test performance. In one camp, scholars argue that the low test performance of minority students can be attributed to cultural deprivation or disadvantage(s); connotatively, this refers to the notion of diverse students being inferior to other students (see Rushton, 2003). Unfortunately, deficit thinking orientations are present even today (e.g., Ford, Harris, Tyson, & Frazier Trotman, 2002). For instance, Frasier, García, and Passow (1995) and Harmon (2002) argued that teachers tend not to refer racially and culturally diverse students to gifted programs because of their deficit thinking and stereotypes about diverse students. When the focus is on what diverse students cannot do rather than what they can do, then they are not likely to be referred for gifted education services. In a different camp, scholars argue that minority students are culturally different, but not culturally disadvantaged or deficient (e.g., Boykin, 1984; Delpit, 1995; Erickson, 2004; Hale, 2001; Nieto, 1999; Rodriguez & Bellanca, 1996; Shade, Kelly, & Oberg, 1997). These individuals acknowledge that culture impacts test performance, but they do not equate or associate low performance with inferiority.

Beyond the ongoing debates about the source in intelligence, there are equally spirited and rigorous debates about the use of standardized tests with diverse groups, with the greatest attention to issues of test bias (Armour-Thomas, 1992; Helms, 1992). Publications on test bias seem to have waned in the last decade, although the *Bell Curve* (Herrnstein & Murray, 1994) generated renewed debates and controversy. Many test developers have gone to great length to decrease or eliminate (if this is possible) culturally biased (or culturally-loaded) test items (Johnsen, 2004). Accordingly, some scholars contend that test bias no longer exists (e.g., Jensen, 1980, 1998; 2000; Rushton, 2003; also see discussion by Fancher, 1985). Others contend that tests can be culturally-reduced, that bias can be decreased; still others contend that tests can never be bias free or culturally neutral because they are developed by people, they reflect the culture of the test developer, and absolute fairness to every examinee is impossible to attain, for no other reasons than the fact that tests have imperfect reliability and that validity in any particular context is a matter of degree (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, hereafter referred to as "Joint Standards," 1999, p. 73).

In sum, there is little consensus in education (and psychology) about the reasons diverse students score lower on standardized tests of intelligence than do White students. Further, there is little consensus regarding the definition of intelligence, the definition of test bias, the existence of test bias, the types of test biases, the impact of test bias on diverse students, and the nature and extent of test bias in contemporary or newly re-normed tests.

With so many unanswered questions and controversies regarding intelligence, testing in general, and testing diverse students in particular, what can educators in gifted education do to ensure that these students have access to and are represented in gifted education programs and services?

In this monograph, I examine test bias by first reviewing seminal publications and research. For example, no discussion of test bias can take place without attending to Jensen's (1969) *Bias in Mental Testing*. This discussion provides the historical context for the monograph. Next, I discuss intelligence tests, paying specific attention to interpretations of and explanations for the comparatively low performance of racially and culturally diverse students on cognitive ability tests. Most of the research has targeted Black students' test performance and Black-White IQ differences. Next, I explore definitions of and strategies for determining the nature and extent of test bias. Finally, I summarize the findings and draw implications for the field of gifted education. Central questions of this monograph are:

1. What is test bias and how is a test determined to be biased, biased reduced, or bias free?
2. What efforts have been made to reduce bias in standardized intelligence tests?
3. Which intelligence tests (e.g., WISC-III, Binet-IV, Cognitive Ability Test, etc.) and types of intelligence tests (e.g., verbal vs. non-verbal) appear to be less culturally-loaded?
4. What explanations are given for the low performance of minority students on standardized tests, that is, for Black-White differences in intelligence test performance? How do assumptions about intelligence tests affect diverse students and who those who work with the tests and students?
5. What proposals have been advanced to increase the performance of diverse students on intelligence tests? For instance, what precautions do test developers give about the purpose and appropriate uses of their particular test, particularly when used with diverse students?
6. What are the implications of testing issues and diversity for gifted education? That is, what direction(s) should educators take in terms of using standardized tests to assess the intelligence of diverse students?
7. What resources and professional standards exist to help the field of gifted education in adopting equitable instruments and assessment practices and policies (e.g., American Educational Research Association, American Psychological Association, Association of School Psychologists)?

8. What other considerations must gifted education address as we seek to increase the representation of minority students in gifted education? For example, what other measures and procedures might be used to assist us in increasing the representation of diverse students in gifted education?

Table of Contents

ABSTRACT v

Background: Confusion and Controversy 1

Testing Issues and Diverse Populations: Beyond Historical Issues 5

Test Bias: Definitions, Types, and Measurement 7
- Statistical Biases 7
  - Mean Differences in Scores 8
  - Single-group or Differential Validity (Also Referred to as Predictive Validity Bias) 8
  - Differential Construct Validity 9
  - Content Validity 9
  - Reliability Issues 10
- Non-statistical Biases: Additional Sources of Validity Error 11
  - Selection Bias and Cutoffs 11
  - Norming Bias 12
  - Testing Environment or Atmosphere Bias 13
  - Examiner Effects 14
  - Interpretation Bias (Attribution of Cause Bias) 15
  - Implementation Bias (or Different Treatment) 16
  - Construct Irrelevance and Construct Under-representation Issues 17
- Equivalence Issues 18

The Influence of Culture on Test Performance: African-American Students as a Case in Point 21

Beyond Traditional, Culturally-loaded Tests: Alternative Tests and Promising Practices 26
- Non-verbal Tests as Alternative Measures 27
- Promising Practices and Considerations 31
  - Culturally Sensitive Assumptions 31
  - Test Interpretation 33
  - From Testing to Assessment: Multi-factored and Collateral Data Collection 33
  - Diversity Training of Test Developers, Administrators, and Users 34
- Adopt Contemporary Definitions and Theories of Intelligence and Giftedness 35

Summary: Guiding Principles for Equitable and Culturally Responsive Assessment 35
- A Word on Test Fairness 37

Conclusion 38

Table of Contents (continued)

References 41

List of Tables

Table 1	Projected Demographics of Gifted Education Programs Nationally	2
Table 2	Average Number of Books and Bookstores in Three California Communities	17
Table 3	African Cultural Components in Cognitive Ability Testing: Hypothesized Effects of African-centered Values and Beliefs	23
Table 4	Culturally-loaded Versus Culturally-reduced Dimensions of Tests	27
Table 5	Numbers and Percentages of Children Who Earned Varying NNAT Standard Scores by Group	31

List of Figures

Figure 1 Vocabulary Scores for Black and White 3- and 4-Year-Olds, 1986-94 13

Intelligence Testing and Cultural Diversity: Concerns, Cautions, and Considerations

Donna Y. Ford
Vanderbilt University
Nashville, Tennessee

Background: Confusion and Controversy

The ambiguity in the term "intelligence" has also enabled it to become influenced by and framed within the context of different philosophical assumptions, political agendas, social issues, and legal restrictions (Groth-Marnat, 1997).

At all levels of education, there is great concern about the low performance of racially and linguistically diverse students—African Americans, Hispanic Americans, and Native Americans—on standardized tests, as well as their under-representation in gifted education. While fewer concerns and criticisms target achievement tests, a wealth of controversy surrounds intelligence tests (also known as cognitive ability tests), specifically given the consistently lower performance of Black students on intelligence tests compared to White students. More so than with achievement tests, intelligence tests carry the burden of being associated with innate ability, particularly by laypersons and those unfamiliar with the purposes and limitations of tests; thus, to those unfamiliar with the purposes and limitations of tests, when one group performs lower than another group, the results, they believe, may be attributed to heredity or genetic inferiority. This simplistic explanation ignores the role of environment, including education and opportunity to learn, on students' test performance.

Issues regarding achievement tests and diverse students are less controversial than those regarding intelligence tests. Compared to intelligence tests , few publications have been written regarding biases in achievement tests. Performance on achievement tests is generally associated with the quality and quantity of students' educational or learning experiences at home and school. For the most part, low achievement test scores are associated with poor educational experiences, lack of motivation, and a host of other factors that tend to be environmental or social rather than inherited or genetic. Conversely, some people presume that intelligence tests measure unlearned abilities—abilities less dependent on instruction and education—and they interpret low performance on intelligence tests with low cognitive ability and potential. This belief is particularly relevant among: (a) individuals who are untrained in testing and assessment; (b) individuals who believe that intelligence is fixed, innate, and unchangeable; and (c) individuals who believe that intelligence tests are comprehensive, exact, and precise measures of intelligence (see discussion in Groth-Marnat, 1997, 2003). Whatever position one holds regarding the nature of intelligence (and achievement) as measured by tests, these tests: (a) measure only a sample of the construct being measured; (b) measure present behavior, namely students' attainment of skills at the time of assessment; and (c) intelligence test scores are an estimate of a person's current level of functioning as

measured by the various tasks required. For instance, as stated by Groth-Marnat (1997), the Wechsler scales, like other tests of intelligence, are limited in the scope of what they can measure. They do not assess such important factors as need for achievement, motivation, creativity, or success in dealing with people.

As discussed later, attempts to develop an accurate definition and measure of "intelligence" have been fraught with difficulty and controversy. This is because intelligence is an abstract concept and has no actual basis in concrete, objective, and physical reality (Groth-Marnat, 1997). While it is possible to observe problem-solving techniques, for example, and to measure the results of these techniques objectively, the intelligence assumed to produce these techniques cannot be observed or measured directly. Groth-Marnat goes on to state that this concept is akin to the term "force" in physics: it can be known by its effects, yet its presence must be inferred.

Nowhere are the debates and controversies surrounding intelligence more prevalent than in gifted education and special education. These two educational fields rely extensively on tests to make educational and placement decisions. In gifted education, low test scores often prevent diverse students from being identified as gifted and receiving services; in special education, low test scores often result in identifications such as learning disabled, mentally retarded, and so forth. Racially and linguistically diverse students (African Americans, Hispanic Americans, and Native Americans) are under-represented in gifted education and over-represented in special education (see Council of State Directors of Programs for the Gifted and National Association for Gifted Children, 2003; U.S. Department of Education, 2003). Ford (1998), Frasier, García, and Passow (1995), and others reported that Black, Hispanic and Native American students have always been under-represented in gifted education programs. In 1993, the U.S. Department of Education noted that Black and Hispanic students were under-represented by 50% in gifted programs, and Native American students were under-represented by 70%. More recent data (Office for Civil Rights [OCR], 1998) indicate that, in 1997, Black students' under-representation *increased* to 60%! (see Table 1)

Table 1

Projected Demographics of Gifted Education Programs Nationally

	% School Population	**% Gifted Education**
White American	63.7	76.6
African American	17.0	7.3
Hispanic American	14.3	8.6
Native American/American Indian	1.1	0.9
Asian American	4.0	6.6

Note. From OCR Elementary and Secondary Civil Rights Compliance Report: 1998, U.S. Department of Education.

There are two persistent, major debates or controversies surrounding minority students' intelligence test performance. In one camp, scholars argue that the low test performance of minority students can be attributed to cultural deprivation or disadvantage(s); connotatively, this refers to the notion of diverse students being inferior to other students (see Rushton, 2003). Unfortunately, deficit thinking orientations are present even today (e.g., Ford, Harris, Tyson, & Frazier Trotman, 2002). For instance, Frasier, García, and Passow (1995), and Harmon (2002) argued that teachers tend not to refer racially and culturally diverse students to gifted programs because of their deficit thinking and stereotypes about diverse students. When the focus is on what diverse students cannot do rather than what they can do, then they are not likely to be referred for gifted education services.

In a different camp, scholars argue that minority students are culturally different, but not culturally disadvantaged or deficient (e.g., Boykin, 1986; Delpit, 1995; Erickson, 2004; Hale, 2001; Nieto, 1999; Rodriguez & Bellanca, 1996; Shade, Kelly, & Oberg, 1997). These individuals acknowledge that culture impacts test performance, but they do not equate or associate low performance with inferiority. Beyond the ongoing debates about the source in intelligence, there are equally spirited and rigorous debates about the use of standardized tests with diverse groups, with the greatest attention to issues of test bias (Armour-Thomas, 1992; Helms, 1992).

Publications on test bias seem to have waned in the last decade, although the *Bell Curve* (Herrnstein & Murray, 1994) generated renewed debates and controversy. Many test developers have gone to great length to decrease or eliminate (if this is possible) culturally biased (or culturally-loaded) test items (Johnsen, 2004). Accordingly, some scholars contend that test bias no longer exists (e.g., Jensen, 1980, 1998; 2000; Rushton, 2003; also see discussion by Fancher, 1995). Others contend that tests can be culturally-reduced, that bias can be decreased; still others contend that tests can never be bias free or culturally neutral because they are developed by people, they reflect the culture of the test developer, and absolute fairness to every examinee is impossible to attain, for no other reasons than the fact that tests have imperfect reliability and that validity in any particular context is a matter of degree (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, hereafter referred to as "Joint Standards," 1999).

In sum, there is little consensus in education (and psychology) about the reasons diverse students score lower on standardized tests of intelligence than do White students. Further, there is little consensus regarding the definition of intelligence, the definition of test bias, the existence of test bias, the types of test biases, the impact of test bias on diverse students, and the nature and extent of test bias in contemporary or newly re-normed tests.

With so many unanswered questions and controversies regarding intelligence, testing in general, and testing diverse students in particular, what can educators in gifted education do to ensure that these students have access to and are represented in gifted education programs and services?

In this monograph, I examine test bias by first reviewing seminal publications and research. For example, no discussion of test bias can take place without attending to Jensen's (1969) *Bias in Mental Testing*. This discussion provides the historical context for the monograph. Next, I discuss intelligence tests, paying specific attention to interpretations of and explanations for the comparatively low performance of racially and culturally diverse students on cognitive ability tests. Most of the research has targeted Black students' test performance and Black-White IQ differences. Next, I explore definitions of and strategies for determining the nature and extent of test bias. Finally, I summarize the findings and draw implications for the field of gifted education. Central questions of this monograph are:

1. What is test bias and how is a test determined to be biased, biased reduced, or bias free?
2. What efforts have been made to reduce bias in standardized intelligence tests?
3. Which intelligence tests (e.g., WISC-III, Binet-IV, Cognitive Ability Test, etc.) and types of intelligence tests (e.g., verbal vs. non-verbal) appear to be less culturally-loaded?
4. What explanations are given for the low performance of minority students on standardized tests, that is, for Black-White differences in intelligence test performance? How do assumptions about intelligence tests affect diverse students and those who work with the tests and students?
5. What proposals have been advanced to increase the performance of diverse students on intelligence tests? For instance, what precautions do test developers give about the purpose and appropriate uses of their particular test, particularly when used with diverse students?
6. What are the implications of testing issues and diversity for gifted education? That is, what direction(s) should educators take in terms of using standardized tests to assess the intelligence of diverse students?
7. What resources and professional standards exist to help the field of gifted education in adopting equitable instruments and assessment practices and policies (e.g., American Educational Research Association, American Psychological Association, Association of School Psychologists)?
8. What other considerations must gifted education address as we seek to increase the representation of minority students in gifted education? For example, what other measures and procedures might be used to assist us in increasing the representation of diverse students in gifted education?[1]

[1] It is currently unrealistic to recommend that tests be completely eliminated from the decision-making process in education. However, some school districts have decreased the heavy or exclusive reliance on test scores to identify gifted students and, accordingly, have increased the percentage of diverse students identified as gifted (see Tomlinson, Ford, Reis, Briggs, & Strickland, 2004). These schools adopted multifaceted, non-discriminatory assessment policies and procedures, as well as developed instruction-based intervention plans for low-performing and potentially gifted diverse students.

Testing Issues and Diverse Populations: Beyond Historical Issues

There is a longstanding and persistent debate regarding the equitable use of tests and assessment strategies with diverse populations. This debate and related concerns are especially prevalent in cases of high-stakes testing, where tests are used to make important and long-term educational decisions about students. As Lamb (1993) observed, once test scores become numbers in students' files, they provide the basis for high-stakes decisions concerning placement, selection, certification, and promotion that are made without consideration of the inequities surrounding testing in general and testing culturally diverse students in particular.

Psychological and psychoeducational assessment is an area that has been heavily subjected to complaints about the differential treatment of diverse groups. Hilliard (1991), Korchin (1980), Olmedo (1981), and others contend that standardized tests have contributed to the perpetuation of social, economic, and political barriers confronting diverse groups (Padilla & Medina, 1996; Suzuki, Meller, & Ponterotto, 1996). Specifically, questions have been raised regarding whether standardized intelligence tests are biased. Tests can be biased in terms of impact (e.g., how they are used) and statistically. Tests can be biased if they treat groups unfairly or discriminate against diverse groups by, for example, "underestimating their potential or over-pathologizing their symptoms" (Suzuki et al., 1996, p. xiii). This concept is referred to as disparate impact (OCR, 2000) and may not be associated with statistical biases, defined next. The Joint Standards (1999) defined statistical bias as a systematic error in a test score. In discussing test fairness, statistical bias may refer to construct under-representation or construct-irrelevant components of test scores that differentially affect the performance of different groups of test takers. Thus, it is important to note that when tests are used for selecting and screening, the potential for denying diverse groups access to educational opportunities, such as gifted education programs, due to bias is great.

The long history of testing diverse groups and related debates and controversies are not described here; they are described in depth in most assessment and intelligence books (e.g., Aiken, 2000; Gould, 1995; Gregory, 2004; Groth-Marnat, 2003; Jones, 1996; Kaufman, 1994; Mensh & Mensh,1991; Montagu, 1999; Sattler, 1992; Sternberg, 1982; Sternberg & Grigorenko, 1997; Suzuki et al., 1996; Valencia & Suzuki, 2001). The focus of this monograph is not to dwell on past injustices, but rather to discuss contemporary problems associated with testing diverse groups, and to make recommendations to decrease or eliminate these assessment problems. With improved and equitable assessment instruments, it is possible that more diverse students will have access to gifted education services.

Past and contemporary arguments against the use of intelligence tests in assessing diverse students fall into two major categories:

1. Intelligence tests contain cultural bias—they contain a strong bias that is in favor White, middle class groups; for example: (a) the tests measure knowledge and content that are more familiar to White, middle class

students than to diverse students; (b) the language on these tests is more familiar to White, middle class students; and (c) the examples used in questions are more familiar to White, middle class students. In this case, it is argued that such tests are measuring what diverse groups have not been exposed to and their differential experiences rather than their intelligence—the construct being assessed (see Fagan & Holland, 2002; Groth-Marnat, 1997). As a result of these issues regarding intelligence tests, those who believe tests contain cultural bias argue that diverse groups are denied access to high-quality, challenging educational opportunities, such as gifted education programs, and they are often misplaced in special education programs and low ability groups.

2. National norms based on middle class Whites may be inappropriate for use with diverse groups. The primary argument is that minority students are not well-represented in the norming or standardization sample; thus, the validity of using such tests with diverse groups is questionable and impedes generalizability due to their low representation in the norming group. For example, if Native Americans represent 1% of the U.S. population and they represent 1% of a test normed on 2,000 students, then only 20 Native Americans will be represented in the test![2] The test performance of these small—minute—number of students cannot possibly be generalized to the larger group of Native Americans, so many of whom come from different nations (once referred to as "tribes"), speak so many different languages, come from diverse socio-economic levels, and have different cultural values and traditions. In essence, how representative are these 20 scores of the Native American population?

These two criticisms relate to the effects of disparate impact associated with testing instruments and policies. Federal government regulations provide that schools that receive federal funds may not "utilize criteria or methods of administration which have the effect of subjecting individuals to discrimination" (OCR, 2000, p. 17). Thus, educators must be vigilant about examining instruments and policies for bias and, thus, disparate impact. In general, bias—intended and unintended—relates to the validity and reliability of the test, and the extent to which the test content and format favor one group over another group.

As described below, just as the nature of intelligence has been extensively debated, so has the issue of test bias. These debates have surrounded such questions as: (a) What is test bias? (b) How can test bias be measured? (c) Does test bias continue to exist in contemporary intelligence tests and, if so, how? (d) Can test bias be eliminated?

---

[2] One example of a widely used test that is normed on a small sample size is the Wechsler Intelligence Scale for Children—Fourth Edition. This test was normed on 2,200 children, 5% (110 children) of whom were gifted, and racial percentages based on 1988 U.S. Census (Gregory, 2004). Further, the Stanford Binet Intelligence Scales—Fifth Edition was normed on 4,800 children (see Groth-Marnat, 1997 and Johnsen, 2004).

And (e) does test bias result in intended or unintended consequences that have a disparate impact on diverse students, namely their participation in gifted education?

Test Bias: Definitions, Types, and Measurement

> An intelligence test is a neutral, inconsequential tool until someone assigns significance to the results derived from it. Once meaning is attached to a person's test scores, that individual will experience many repercussions, ranging from superficial to life-changing. These repercussions will be fair or prejudiced, helpful or harmful, appropriate or misguided—depending on the meaning attached to the test score (Gregory, 2004, p. 240).

Some criticisms about intelligence test and bias were summarized by Gregory (2004) as follows:

1. Intelligence tests are misnamed because they were never intended to measure intelligence and might have been more aptly called CB (cultural background) tests;
2. Persons from backgrounds other than the culture in which the test was developed will always be penalized;
3. There are enormous social class differences in a child's access to experiences necessary to acquire the valid intellectual skills;
4. The poor performance of African American children on conventional tests is due to the biased content of the test; the test material is drawn from outside the African American culture.

The topic of test bias has received wide attention from measurement psychologists, test developers, test critics, educators, legislators, and the courts. The test-bias controversy has its origins in the observed differences in average IQ scores among various racial and ethnic groups compared to White populations (Gregory, 2004), and concerns about and efforts to eliminate bias in tests stem from the belief that biased tests could perpetuate a legacy of racial discrimination.

There are numerous definitions of test bias, each of which has some value in explaining the properties of tests and their uses (Sattler, 1992). However, it is generally agreed that test bias is a technical concept amenable to impartial analysis (Gregory, 2004). Two broad categories of bias are discussed below, statistical bias and non-statistical bias.

Statistical Biases

Tests can be evaluated using various statistical techniques to determine if they are biased against diverse groups. Evidence about the intended and unintended consequences of tests, including disparate impact as possibly reflected by the under-representation of diverse groups in gifted education programs, provide important information about the

validity of the tests and of the inferences to be drawn from the test results. Likewise, such evidence can be used to raise concerns about the inappropriate use of a test. Five statistical biases are described: mean differences in scores; single-group validity; differential construct validity; differential validity; and reliability issues.

Mean Differences in Scores

A test may be biased when it yields lower scores for one group than for another group, specifically if the groups are similar in terms of ability or similar status (e.g., SES) (Joint Standards, 1999). The reasons for the different scores must be analyzed. For example, data indicate that Black students tend to score 15 points lower on intelligence tests than White students (e.g., Jencks & Phillips, 1998). Despite the fact that intelligence tests such as the Wechsler Intelligence Scale for Children—Third Edition (WISC-III) (Wechsler, 1991) and Stanford-Binet Intelligence Scale—Fourth Edition (Thorndike, Hagen, & Sattler, 1986) yield lower scores for minority children (see Kaufman, 1994, for discussion of the WISC-III), they have been widely used for gifted identification. Wasserman and Becker (2000) provided a summary of recent research on the WISC-III (Wechsler, 1991), Stanford-Binet IV (Thorndike et al., 1986), and Woodcock-Johnson Tests of Cognitive Ability (WJ-R; Woodcock & Johnson, 1989) that used samples matched on key demographic variables. They found that the average *differences*, in favor of Whites, between standard scores for matched samples of Black and White groups were as follows: WISC-III=11.0; Stanford-Binet IV=8.1; and Woodcock-Johnson Tests of Cognitive Ability=11.7. These sizable mean score differences suggest that fewer minority children might be identified when such tests are used for determination of giftedness (Naglieri & Ford, 2003).

For some educators, as stated earlier, the presence of mean differences in scores of matched groups of students may indicate that the test is biased. When the use of a test results in outcomes that affect the life chances or educational opportunities of examinees, evidence of mean test score differences between groups should be examined. Where mean differences are found, an investigation should be undertaken to determine that such differences are not attributable to a source of construct under-representation or construct-irrelevant variance, discussed later (Joint Standards, 1999) and are not attributable to lack of access to high-quality curriculum and instruction, thereby reflecting differences in achievement rather than ability (OCR, 2000). As discussed later, one question worth investigating is whether the group differences are evidence of a biased test or evidence of differences in life circumstances.

Single-group or Differential Validity (Also Referred to as Predictive Validity Bias)

A second perspective on test bias is whether a test is an equally good predictor for different groups. A test is considered biased with respect to predictive or differential validity if the inference drawn from the test score is not made with the smallest feasible random error or if there is constant error in an inference or prediction as a function of membership in a particular group (Reynolds, 1998). Test bias is present when a validity

coefficient is significantly different from zero (0) for one group but not for another group (see Gregory, 2004).

Differential Construct Validity

A third way to assess bias in tests is to study the extent to which the test measures the same construct in various minority groups. The central question here is whether Blacks, Whites, Hispanic Americans, Asian Americans and Native Americans are performing differently on the same intelligence measure, and whether the factor structures of the tests are similar to those of White students.

Content Validity

Related to the notion of construct validity is content validity or bias. Concerns regarding content validity fall into at least three categories: (a) the items asks for information that minority persons have not had equally opportunity to learn; (b) the scoring of the items is improper, since the test author has arbitrarily decided on the only correct answer, minority groups are inappropriately penalized for giving answers that would be correct in their own culture; and (c) the wording of questions is unfamiliar, therefore, a minority person who may "know" the correct answer may not be able to respond because he/she does not understand the question (Reynolds, 1998). Reynolds (1998) stated:

> An item or subscale of a test is considered to be biased in content when it is demonstrated to be relatively more difficult for members of one group than another when the general ability level of the groups compared is held constant and no reasonable theoretical rationale exists to explain group differences on the item (or subscale) in question. (cited in Gregory, 2004, p. 243)

A useful approach to measuring construct validity is to examine each test item to determine if groups perform differently. The central question here is whether a test is measuring the same construct in different groups. Do the tests items have the same meaning to each individual taking the test or is the test measuring a different construct for each group (e.g., Groth-Marnat, 1997; Helms, 1992)? If a test is non-biased, then comparisons across relevant subpopulations should reveal a high degree of similarity for: (a) the factorial structure of the test; and (b) the rank order of item difficulties within the test (Gregory, 2004). That is to say, an essential criterion of non-bias is that the factor structure of test scores should remain invariant across relevant subpopulations.

As discussed later, in addition to employing statistical measures of bias, it is also necessary to examine factors that affect students' performance on tests. Stated differently, measures of validity can provide *statistical* evidence that a test does not have an inherent bias against any ethnic group. Whether use of a particular test in a particular situation results in discrimination, however, will depend on such factors as the purposes to which the results are put, how the results are interpreted, and how the test is administered (Sattler, 1992). Put another way, a test may accurately measure differences

in the level of students' academic achievement; low scores may indicate that students do not know the content. However, educators should ensure that they interpret test scores with this information in mind. It should not be assumed that low test scores reflect a lack of ability (low intelligence), or an inability to master the content, or an inability to achieve in actual educational settings, including gifted education classes.

Reliability Issues

Undoubtedly, the most important aspect of a test is the degree to which it is valid. Thus, concerns regarding test validity are given greater attention in discussions of test bias than are concerns regarding reliability. However, the potential problems and subsequent influences of reliability problems cannot be ignored. Validity *and* reliability are two necessary conditions for the existence of a test as a viable tool of measurement—for what use is a test that does not fulfill its purpose? Of what use is a test that does not provide consistent results (Samuda, 1998)? Reliability refers to the consistency of scores obtained by the same individual when re-examined with the same test on different occasions, or with different sets of equivalent items, or under other variable examining conditions (Anastasi, 1968). No test can be valid if it is not reliable; conversely, a test can be reliable but not valid, as the following example indicates:

> As with a witness testifying in a courtroom trial—the fact that he consistently tells the same story does not guarantee that he is telling the truth. The truthfulness of his statements can be determined only by comparing them with some other evidence. Similarly, with evaluation results, consistency is an important quality but only if it is accompanied by truthfulness, and truthfulness, or validity, must be determined independently. Little is accomplished if evaluation results consistently provide the wrong information. (Gronlund, 1981, p. 76)

A number of resources discuss the various types of reliability coefficients (see Anastasi, 1968; Kazdin, 1992) and, as such, are not discussed here. Rather, the focus is on factors that affect a test's reliability. According to Samuda (1998), test length (longer tests are more reliable than shorter tests, and lower scores are more reliable than higher scores), item difficulty, group heterogeneity (the more heterogeneous the group, the higher the reliability), and spread of scores (the wider the spread, the more reliable the test) affect reliability. The final factor, spread of scores, carries important implications for diverse students. Data indicate that diverse groups tend to have a narrow spread of scores, and the scores tend to cluster at the lower end of the scale with smaller differences among them. As such, Samuda (1998) concluded that minority children are assessed by means of tests that do not indicate the value of the reliability coefficient for their group. When a group differs from the sample on whom reliability was established, the actual effectiveness of the test will tend to be lower for that group. Stated another way:

> The sensitive test user should be alert to the reliability considerations in regard to the particular group involved and the intended use of the test. . . . He will try to determine whether the standard error of measurement varies with score levels and whether his testing conditions are similar to those of the persons and purposes

> with which he is concerned. He will know that high reliability does not guarantee validity of the measures for the purpose at hand, but he will realize that low reliability may destroy validity. (Fishman, Deutsch, Kogan, North, & Whiteman, 1964, p. 133)

Non-statistical Biases: Additional Sources of Validity Error

According to the Joint Standards (1999), in addition to statistically examining validity and reliability issues (discussed earlier), educators must examine the *sources* of validity errors, namely such factors as selection bias and cutoff scores, norming, test interpretation, construct under-representation and construct irrelevance, educational opportunity, examiner effects, and their influence on diverse students' test performance. Essentially, the validity and reliability of test scores is called into question when the test scores are substantially affected by irrelevant factors—factors that are not related to the knowledge and skills that the test is supposed to measure (Joint Standards, 1999).

Selection Bias and Cutoffs

The selection of a test refers to the extent to which the test has a differential effect on the number of examinees from various groups who enter certain programs, including gifted programs. On what basis is a test selected? How is the test used? How are scores on tests interpreted and used for placement and services decisions?

A discussion of disparate impact is relevant to the topic of cutoff scores. Districts often use an IQ score of 130 or higher to identify students as gifted (Colangelo & Davis, 2003; Davis & Rimm, 2004); cutoff cuts are used, in many cases, to control or limit the number of students being identified. This may be due to the limited number of students who can be served in gifted education classes because of funding or other issues.

Cutoff scores are often arbitrarily chosen; they are specific points on the test where results are used to divide levels of knowledge, skill, and ability. The point at which to divide the levels should not be taken lightly, particularly when diverse students are involved. The primary question here is, given that diverse groups tend to score lower on conventional standardized tests than White, middle class students, should there be the same or different cutoff scores for all groups and/or should some groups be given extra points? If so, this would result in schools adopting a quota system based on race, which is not feasible legally. However, as already noted, within the same school district, school building P may have an average IQ of 90; in school building G, the average IQ may be 115. In building P, unchallenged students may be those with an IQ of 120 or higher; in building G, unchallenged students may be those with an IQ of 145 or higher. Must all gifted children have an IQ of 130 or higher? Can selections for gifted education services be made at the school building level, particularly when districts have such different demographics within their school district?

Like all scores associated with tests, cutoff scores must be accurate representations of the knowledge and skills of students. Levels and categories must be

distinctively different and based on sound empirical data (Joint Standards, 1999). In other words, validity evidence should be able to demonstrate that students above the cutoff score represent or demonstrate a qualitatively greater degree or different types of skills and knowledge than those below the cutoff scores. It is essential to examine the validity of the inferences that underlie the specific decisions being made on the basis of the cutoff scores. What must be validated is the specific use of the test based on how the scores of students above and below the cutoff score are being interpreted.

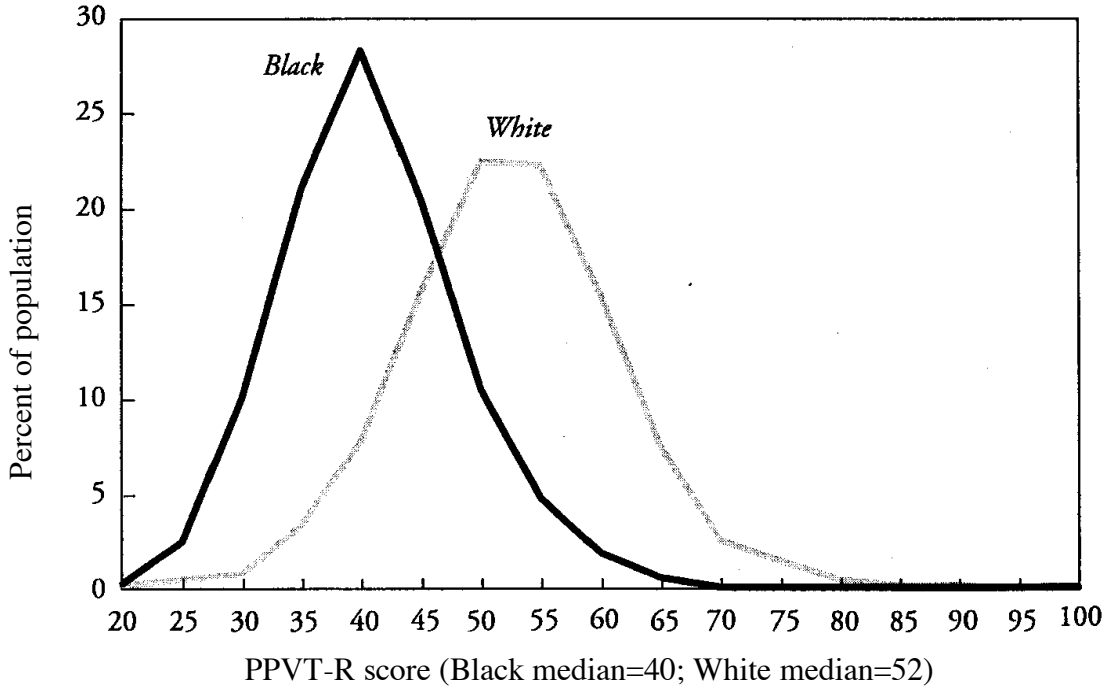
Reliability considerations (less often discussed in testing) must be also addressed with cutoff scores. When cutoff scores are specified for placement in gifted education, the degree of measurement error around each cutoff score should be reported (Joint Standards, 1999). Are the measurement errors different for different groups? Equally important, evidence should be collected on the misclassification rates (false positives and false negatives) that are likely to occur among students with comparable knowledge and skills, and between White and diverse students.

Norming Bias

In the earliest versions of intelligence tests, diverse students were not included in the norming or standardization sample; nonetheless, diverse students were assessed with the instruments. Not surprising, diverse students did not do well (or as well as) White students on the tests. In the more recent versions of tests, publishers have sought to ensure that diverse groups are included in the norming sample in proportion to their percentage of the U.S. or school-age population. For example, if Blacks comprised 15% of the school population when the test was developed, they would comprise 15% of the norming sample; if Native Americans comprised 1% of the U.S. population, they would comprise 1% of the norming sample, and so forth (see footnote 1 for previous discussion on this issue).

Using proportions in this way helps to ensure representation and diversity and, thus, to improve generalizability. It is worth noting that when the sample is small, sampling using these proportions can still be problematic. Specifically, when a test is normed on 3,000 students, 1% of the norming sample represents only 30 Native American students. How generalizable are these small number of test scores (see Sandoval, Frisby, Geisinger, Scheuneman, & Grenier, 1998)? How representative are these test scores for Native Americans, particularly given that there are over 400 different tribal groups in the U.S.? How many of these 30 students come from higher SES backgrounds? How many are males? How many have learning disabilities? How many are limited in English proficiency? Are these student characteristics represented in the 30 students? If two of the 30 Native American students are Navajo, can the results be generalized to other Native American groups? If four of these Native American students are from higher (or lower) SES backgrounds, can their test scores be generalized to low socio-economic status (low SES) Native Americans? These same questions, and others, are relevant to all diverse groups.

Most schools identify students as gifted if they have an IQ score that is two standard deviations above the test's mean or the national norm. When we state that "gifted students score two standard deviations about the norm" on intelligence tests, the over-arching question is, *whose* norm? Are we referring to the average U.S. IQ of 100 or the lower average IQ norms (means) for the different diverse groups? Figure 1, showing two different normal curves or bell curves for Black and White students illustrates this point. In short, cutoff scores must be considered carefully and with diverse students in mind.



Source: National Longitudinal Survey of Youth Child Data, 1986-94. Black N=1,134; White N=2,071. Figure is based on Black and White 3- and 4-year-olds in the Children of the National Longitudinal Survey of Youth (CNLSY) data set who took the Peabody Picture Vocabulary Test-Revised (PPVT-R). The test is the standardized residual, coded to a mean of 50 and a standard deviation of 10, from a weighted regression of children's raw scores on their age in months, age in months squared, and year-of-testing dummies. See chapter 4 for details on the CNLSY and the PPVT-R. (As cited in Jencks, C., & Phillips, M. (Eds.). (1998). *The Black-White Test Score Gap*. Washington, DC: The Brookings Institute.) [Reproduced with the permission of the publisher.]

Figure 1. Vocabulary scores for Black and White 3- and 4-year-olds, 1986-94.

Testing Environment or Atmosphere Bias

Testing environments affect students' test performance. If students taking the test feel out of place or unwelcome, they will not put forth their best effort. When their performance is inhibited in this way, the test environment is biased and the students' test performance or score is questionable. Research by Steele (1999) and Steele and Aronson

(1998) on "stereotype threat" sheds light on this issue. Their works show that some Black students suffer from "stereotype threat" during evaluative situations. These students become anxious and over-preoccupied with concerns about how their test scores, especially if low, will be used to (mis)interpret the abilities and potential of other Black students. This anxiety hinders or undermines their test performance.

Examiner Effects

The Joint Standards (1999) listed several factors associated with test examiners that may affect the performance of students while taking tests, including: (a) the cultural and linguistic background of the examiner; (b) the gender of the examiner; (c) the testing style of the examiner; and (d) the level of acculturation of the examiner. A related issue related to the testing environment is race examiner effects. Just as gender effects were noted by the Joint Standards, we must also consider the potential effects of race. How does the race of the examiner affect the performance of diverse students during test administration?

Those who believe that the examiners' race matters when testing diverse students believe the anxiety, insecurity, latent prejudice, and other reactions to diverse children that are experienced by White examiners may be transmitted to the children in several ways (see Groth-Marnat, 1997; Sattler, 1992). Examiners may exhibit paternalism, over-identification, over-concern, excessive sympathy, reactive fear and suspicion, and stereotypes again diverse groups, and these reactions and perceptions can affect diverse students in negative ways. In particular, stereotypes can interfere with rapport between the diverse child and White examiner. When the test is administered individually, examiners effects may be more important to consider. For instance, if the examiner has a distain for non-standard English, which many Black students speak, he/she may communicate this to the student in subtle ways, and interpret test scores with this distain in mind.

Although Sattler (1992) notes that most research indicates that the examiner's race (and stereotypes and values) does not jeopardize the test performance of diverse students, he nonetheless cautions:

> . . . examiners cannot be indifferent to the examinee's ethnicity. They must be alert to any nuances in the test situation that suggest an invalid performance. Testing children from different cultures is a demanding task. At times it may be difficult to understand children's responses, and every effort must be used to enlist their best efforts. (p. 571)

Further, in an interestingly titled book, *Intelligent Testing with the WISC-III*, Kaufman (1994) states:

> The value of the scores increases when the examiner functions as a true experimenter and tries to determine *why* the child earned the particular profile revealed on the record form; the IQs become harmful when they are

> unquestionably interpreted as valid indicators of intellectual functioning and are misconstrued as evidence of the child's maximum or even typical performance. (p. 9) (also see Groth-Marnat, 1997)

Interpretation Bias (Attribution of Cause Bias)

The extent to which a test and its results are inappropriate or biased also depends on whether test scores are an accurate reflection of a student's knowledge or skills, or whether the scores are influenced by extraneous factors unrelated to the specific skills being tested. Central to this discussion is whether all students have had an equal opportunity to acquire the knowledge and skills that are being tested (Fagan & Holland, 2002; Joint Standards, 1999; OCR, 2000). Accordingly, when professionals interpret the test scores of diverse groups, they must use caution to not under-interpret or over-interpret the scores. For example, a low test score does not always mean that a student is not gifted; a high test score does not mean that a student is gifted (false negative and false positive, respectively). Further, a low test score on a vocabulary test or subscale does not necessarily mean that a student has poor verbal skills when communicating verbally/orally. Speaking non-Standard English during test administration is not an indication that a student cannot or does not speak Standard English; many Black students know how to code switch, to speak both non-Standard and Standard English but may be more confident and comfortable speaking non-standard English (e.g., Labov, 1972, 1982; Smitherman, 1977, 1999; Taylor, 1990).

The consequences of interpretation bias are grave. For instance, because many school districts rely on a single test score to place students in gifted education program,[3] and given the lower performance of diverse groups on tests, this practice serves as an effective gate-keeping mechanism. Interpreting test performance—high or low—based on one test or measure must be avoided due to the limited data provided from a single score. Joint Standards (1999), NAGC (1997), and OCR (2000) have noted the serious limitations and negative consequences (e.g., disparate impact) of using one test score to identify students as gifted and to determine their need for placement in gifted education programs. In other words:

> Tests are not perfect. Test questions are a sample of possible questions that could be asked in a given area. Moreover, a test score is not an exact measure of a student's knowledge or skills. A student's scores can be expected to vary across different versions of a test—within a margin of error determined by the reliability of the test, and as a function of the particular sample of questions asked and/or transitory factors, such as the student's health on the day of the tests. Thus, no single test score can be considered a definitive measure of a student's knowledge. (OCR, 2000, p. 14)

---

[3] According to the most recent report by the Council of State Directors of Programs for the Gifted and the National Association for Gifted Children (2003), in 2001-2002, only 24 states mandate non-discriminatory testing in their gifted education policies and procedures, while 18 report no such mandate. Further, several states report using one score to make placement decisions (e.g., Arizona, Oregon, Ohio).

Test scores can mislead just as easily as they can lead (Kaufman, 1994). For many reasons students get high test scores (e.g., coaching, tutoring, practice effects) and for many reasons, students can score low (poor test taking skills, test anxiety, poor listening skills, poor reading skills, poor attention span). Thus, if an interpretive approach relies strictly on one view of the world, no matter how theoretically or psychometrically defensible that view may be, it is doomed to fail for some children (Kaufman, 1994).

Implementation Bias (or Different Treatment)

Implementation bias occurs when diverse students earn test scores that qualify them for placement in gifted education programs and services, but they are denied access. To what extent do test scores lead to misplacement in low level courses or deny access to high-quality educational experiences, including access to gifted education programs and services? An insidious example of implementation bias or different treatment is that of a school district intentionally treating students differently on the basis of race where minority students have scores that qualify them for high level classes, but they are placed in lower level classes (e.g., *People Who Care v. Rockford Board of Education,* 1994, cited in OCR, 2000).

Our basic obligation as educators is to meet the needs of students as they come to us—with their different learning styles, economic backgrounds, cultural backgrounds, and academic skills. In *Larry P. v. Riles* (1979), the court argued: If a test predicts that a person is going to be a poor employee, the employer can legitimately deny the person the job. On the other hand, if a test suggest that a child is probably going to be a poor student, a school cannot deny the child the opportunity to improve and develop the academic skills necessary to succeed in our society. The school cannot use one test, one piece of data, to deny opportunities to children. Stated differently, gifted education must not only teach gifted students who demonstrate their gifts and talents, they must also address student potential and, thus, create talent development models (Callahan & McIntyre, 1994; USDE, 1993, 1998).

The philosophy of developing (and nurturing) talent in students is especially essential when students live in poverty. Like diverse students, low SES students are under-represented in gifted programs (USDE, 1993). Giftedness is present in all SES groups, but those who live in low SES homes and communities (e.g., Watts and Compton in California) often have fewer educationally enriched experiences than those in higher SES situations (e.g., Beverly Hills). Table 2 illustrates this point. Children in Beverly Hills come from higher SES backgrounds than those from Watts and Compton. The exposure of Beverly Hills children to books, to literature and literacy—at home *and* at school—is greater than exposure for the other two groups of students. Therefore, children in Beverly Hills are more likely to have higher reading skills and literacy rates than the other children, and they are more likely to perform better on intelligence (and achievement) tests. Further, as reported by Hart and Risley (1995) and more recently reported by the National Center for Education Statistics (NCES) (U.S. Department of Education, 2003), Black students tend to have less exposure to words and literature than

White students. More specifically, Hart and Risley found that the amount of exposure to language predicted the vocabulary and IQ scores of children at age 3, and that the children of professionals (typically, Whites) were exposed to five times the amount of words than were children of parents on welfare (typically, Blacks). These collective data tell us that such differences in life in exposure to information on the part of Blacks and Whites are an empirical fact that must be dealt with. Exposure to language and literacy, and other learning opportunities are environmental factors—due to exposure and experience—and this reality must be considered when comparing test scores across groups who are different relative to race and SES. Stated another way, we cannot deny the heritability of intelligence; neither can we deny the effects of literacy and educational experience and opportunity on one's intelligence test performance.

Table 2

Average Number of Books and Bookstores in Three California Communities

	Books in home	***Books in classroom libraries***	***Books in school libraries***	***Books in public libraries***	***Number of bookstores***
Beverly Hills (high SES & White)	199	392	60,000	200,600	5
Watts (low SES & Black)	.4	54	23,000	111,000	0
Compton (low SES & Black)	2.7	47	16,000	90,000	1

Note. From "Differences in Print Environment for Children in Beverly Hills, Compton, and Watts" by C. Smith, R. Constantino, & S. Krashen, 1997, *Emergency Librarian, 24,* pp. 8-9.

Construct Irrelevance and Construct Under-representation Issues

Construct irrelevance is another source of validity error that influences students' test performance. It occurs when a test measures material that is extraneous to the intended construct, thereby confounding the ability of the test to measure the construct that it is intended to measure (OCR, 2000). How well a child reads may influence his/her test score in mathematics computation. As such, the student's *reading* skills may be irrelevant when the skill of mathematics computation is what is purported to be measured by the test. In a lengthy discussion, Groth-Marnat (1997) describes how various subscales on the Wechsler tests are influenced by culture or socio-cultural factors. Specifically, using the intelligence theory of Horn and Cattell, he reports that the verbal subtest is a measure of crystallized intelligence that is more influenced by cultural and

social factors than the performance subtest, a measure of fluid intelligence. In brief, crystallized intelligence is learned intelligence or acquired knowledge, while fluid intelligence is considered a measure of raw, unlearned intelligence (also see Gregory, 2004). Two findings are worth noting: (a) the performance scores bear a weaker relationship to school achievement than the verbal scores (Gregory, 2004); and (b) individuals from low SES backgrounds and culturally diverse groups tend to score higher on the performance subtest than the verbal subtest (e.g., see Groth-Marnat, 1997; Naglieri & Ford, 2003). Along these lines, three subtests on the Wechsler verbal scale will be discussed relative to how they both measure and are influenced by social and cultural factors. Groth-Marnat (1997) notes that:

1. the *information subtest* assesses: old learning or schooling; alertness to daily world; long-term memory; intellectual curiosity or urge to collect knowledge; and range of general factual knowledge;
2. the *vocabulary subscale*[4] assesses: educational background; range of ideas, experiences, or interests that a subject has acquired; language development; word knowledge; language usage and accumulated verbal learning ability; and more; and
3. the *comprehension subscale* assesses: social judgment or common sense; grasps of one's social milieu (e.g., information and knowledge of moral codes, social rules, and regulations); knowledge of conventional standards of behavior; social maturity, and more.

Joint Standards (1999) and OCR (2000) contend that construct under-representation occurs when some important aspects of the intended construct being tested are omitted or irrelevant to the test itself. One example of construct under-representation would be a test that is being used to measure English language proficiency, defined as specific skills in listening, speaking, reading, and writing the English language, but the test only measures reading skills (OCR, 2000). With both construct irrelevance and under-representation, the central concern is validity, as well as the confounding effects of factors that are hindering students' test scores.

Equivalence Issues

According to Helms (1992), Armour-Thomas (1992), and Armour-Thomas and Gopaul-McNicol (1998), functional equivalence, conceptual equivalence, and linguistic equivalence are potential sources of bias when testing diverse groups. Functional equivalence is the extent to which test scores have the same meaning in different cultural groups and measure psychological characteristics that occur with equal frequency within these groups. Conceptual equivalence concerns whether groups are equally familiar or unfamiliar with the content of the test items and, therefore, attribute the same (or

[4] Studies indicate that vocabulary is the best single predictor of general intelligence, with 86% of its variance accounted for by g on the WAIS-R and 80% of its variance accounted for by g on the WISC-III (see Groth-Marnat, 1997). Vocabulary generally reflects the nature and level of sophistication of one's schooling and cultural learning; vocabulary is primarily dependent on the wealth of early educational environment, but it is susceptible to improvement by later experience or schooling.

different) meaning to them. Linguistic equivalence concerns the extent to which the test developer has equalized the language used in the test so that it signifies the same thing to different cultural groups. Here are a few examples of test items that illustrate problems associated with functional, conceptual, or linguistic equivalence:

1. If an intelligence test item asks students about items to take the beach, and the students have never been to the beach or had discussions about the beach, they will be at a disadvantage in responding correctly.
2. If an intelligence test asks students to describe the purpose of wearing a life jacket when in a boat, they may not know the answer if they have never been on a boat, fishing trip, etc.
3. If an intelligence test item asks students why people read the newspaper, students whose families do not read or subscribe to the newspaper may not be able to answer this item correctly. The same issue arises with the question, why do we have books? (particularly given the data in Table 2).
4. If an intelligence test item asks students to describe the best part of going on a vacation, some students who have never taken a vacation will have difficulty responding to this item.
5. If an intelligence test items asks students to identify Santa Claus or Elvis Presley from a picture, some students may respond incorrectly because they do not celebrate Christmas or their family does not listen to rock 'n roll.
6. If an intelligence test item asks students "what is a crib?" they may get the answer incorrect if they use the word "crib" in a different way from the person who developed this item (e.g., Is a crib a baby bed versus a "house" in slang vs. a place to keep corn for pigs on a farm?).
7. If an intelligence test item asks students to find a synonym for the word "good," and they choose, "large," then they will not get credit for the item (Note the use of "good" in this sentence by a British person: "A good part of the looting came after the fighting ended." Translation: A "large" part of the looting . . .).

Many of the above examples assess comprehension; however, clichés, analogies, and face/object recognition items can contain potentially biased content (e.g., Analogies—How are a car and boat alike? In what ways are a piano and guitar alike? How are an apple and banana alike? How are a mango and banana alike?); (e.g., General information—What are the four seasons of the year?); (e.g., Face or object recognition—students are asked to look at a picture/object, and to identify). Kaufman (1994) noted that the most "culturally-loaded" tasks on the K-ABC are the faces and places items, contributing to differential Black and White IQ test scores. All of these items assume familiarity with the content, context, and language, and that test takers have had an equal opportunity to learn the test's content. For example, all students are likely to have been taught that there are four seasons of the year. However, in "What are the four seasons of the year?" a child from India might give "Monsoon season" as one response. Further, in sharing this item with males, I have found that many respond, "Football season, baseball season, basketball season, and golf season."

A final example of how test content can be biased, and otherwise influenced by different educational, cultural, and social experiences, is taken from Helms (1992) and Walsh and Betz (1985). On a sample GRE question, test takers were presented with this item:

> *Old Mother Hubbard went to the cupboard*
> *To get her poor dog a bone*
> *But when she got there her cupboard was bare*
> *And so the poor dog got none.*
>
> *If the above is an accurate report of an event, which of the following headline versions gives an account that does not add to the given facts?*
> *A Mother Hubbard refuses bone to hungry dog.*
> *B Mealtime brings only bare cupboard for Mrs. Hubbard and dog.*
> *C Mother Hubbard seeks bone for dog. Finds empty cupboard.*
> *D Dog lover unable to continue support of pet.*
> *E Bone missing from Hubbard cupboard—Mystery unsolved.*
> (Walsh & Betz, 1985, p. 174).

The "correct" answer is "C." Yet one can make an argument for at least one of the other alternatives. For example, "E" is a viable alternative if one assumes that Mother Hubbard was expecting to find a bone in her cupboard (otherwise she would not have gone there looking for one), then indeed the whereabouts of the bone is (without going beyond the facts) and unsolved mystery to her (and probably the dog as well). If Black students are not socialized to believe that authority figures reward obvious answers but, instead, reward expansive or creative answers, then they might choose an alternative other than "C."

These examples illustrate that diverse students (in fact, all students) bring different experiences, different ideas, and different vocabularies to the test-taking situation, and this reality should not be dismissed or taken lightly. Thus, the potential for diverse students to misunderstand the items, to be confused by the content of the test, to be unfamiliar with the format of the test, and to lack the experiences to respond to the items are present.

Thus, to address problems associated with sources of errors, all potential sources of error must be considered and explored; we must collect evidence about what a test measures for particular groups of students, and help ensure that the responses by and, thus, scores of culturally diverse students are not unduly influenced by extraneous sources of error.

When examining construct under-representation, construct irrelevance, equivalence issues, and environmental effects, content validity and construct validity are of concern, as is the legal notion of "disparate impact." Disparate impact occurs when educational decisions based on test scores reflect significant disparities based on race, national origin, sex, or disability in the kinds of educational benefits afforded to students.

When this happens, questions about the educational practices (including testing practices) should be thoroughly examined to ensure that they are in fact non-discriminatory and educationally sound (OCR, 2000).

When courts have examined disparate impact, three questions have been explored to determine if the practice or instrument at issue is discriminatory: (a) Does the practice or procedure in question result in significant differences in the award of benefits or services based on race, national origin, or sex? (b) Is the practice or procedure educationally justified? (c) Is there an equally effective alternative that can accomplish the institution's goal with less disparity? (OCR, 2000, p. 18).

Given the concerns and issues just presented, educators and other decision makers must consider the extent to which the tests have a disparate impact on diverse students doing well and, therefore, having access to gifted education programs and services. Essentially, how useful are the test results?

The Influence of Culture on Test Performance: African-American Students as a Case in Point

The body is the hardware; culture is the software.
— Hofstede

Culture can be defined as the collective beliefs, attitudes, traditions, customs, and behaviors that serve as a filter through which a group of people view and respond to the world (Erickson, 2004; Ford & Harris, 1999; Ford et al., 2002; Hall, 1959, 1976). Culture is a way of life, a way of looking at and interpreting life, and a way of responding to life. This definition becomes clearer when one thinks of the "the terrible twos," the teen or adolescent culture, the culture of poverty, and so forth. Members of these groups have in common beliefs, attitudes, traditions, customs, and behaviors (e.g., Storti, 1998).

In a thoughtful and compelling monograph entitled *A New Window for Looking at Gifted Children,* Frasier, Martin, et al. (1995) state, "Manifestation of characteristics associated with giftedness may be different in minority children, yet educators are seldom trained in identifying those behaviors in ways other than the way they are observed in the majority culture." This statement was confirmed in a study that included teachers' perceptions of giftedness among diverse students (Frasier, Hunsaker, Lee, Finely, García, et al. (1995), Frasier, Hunsaker, Lee, Mitchell, et al. (1995) and also discussed in Frasier, Hunsaker, Lee, Finely, Frank, et al. (1995)).

Helms (1992) provides another thoughtful, conceptual treatise on the issue of how culture impacts test performance and, thereby, raises questions about the validity of tests when used with diverse groups (see Groth-Marnat, 1997, 2003; Miller, 1996; Sattler, 1992; Sternberg, 1982 for other definitions of culture and the impact of culture on test performance).

As stated earlier, Helms maintains that the notion of cultural or functional equivalence must be considered when diverse students are being tested or assessed. Using Boykin's (1986) research on the modal characteristics of Blacks, Helms (1992) hypothesized how these Afrocentric cultural dimensions or characteristics can and do influence the test performance of Black students (see Table 3). The Afrocentric styles gleaned from Boykin's research include: spirituality; verve and movement; harmony; communalism; orality; time perspective; affect; and expressiveness[5].

- *Spirituality* is a belief that nonmaterial forces have governing powers in one's everyday affairs. It is a conviction that all of life is governed by a power greater than oneself.
- *Harmony* is seeing oneself as one with environmental surroundings; the aim is to blend in with the setting, to be a member of the setting; harmony is also an enhanced ability to read the environment and to read non-verbal behaviors well.
- *Movement* is a preference for being active, mobile, and physically engaged or involved; it is a rhythmic orientation to life, as seen in music and dance, and an ability to express oneself non-verbally.
- *Verve* is a propensity for high levels of energy and stimulation; it denotes a disdain for routine and doing things in a rigid, sequential fashion; verve entails a preference for doing things simultaneously and instantaneously.
- *Affect* involves a propensity to be feeling oriented, to engage in or avoid activities and people for whom one has strong positive or negative feelings, respectively. Students strong in affect can be impulsive and very sensitive or emotional. Like harmony, affect is also a keen ability to read the emotional cues of others.
- *Communalism* refers to social interdependence and connectedness; this is a social orientation, that is often accompanied by a strong need for affiliation; group affiliation is important as denoted by an other-centeredness rather than self-centeredness. Students with this orientation prefer to work in groups and make group decisions rather than work independently or alone.
- *Expressiveness* (also known as expressive individualism) refers to an orientation of being creative and a risk taker. There is concern about style, being spontaneous and original in dress, music, speech, and other forms of expression. Life is approached in artistic and creative ways.

[5] Boykin's (1994) data-based model of Afrocentric cultural styles has been examined and discussed in hundreds of publications; for a discussion of the cultural styles of Blacks, Hispanic Americans, Asian Americans and Native Americans, see, for example, Baldwin and Vialle (1999), Banks (1995), Callahan and McIntyre (1994), Castellano (2003), Cline and Schwartz (1999), Maker and Schiever (1989), Shade, Kelly, and Oberg (1997), Storti (1998), and USDE (1998).

Table 3

African Cultural Components in Cognitive Ability Testing: Hypothesized Effects of African-centered Values and Beliefs

Dimension	***General Description***	***Influence on Test Responses***
Spirituality	Greater validity of the power of immaterial forces in everyday life over linear and factual thinking	It may be difficult to separate relevant aspects of the test stimuli from factors caused by luck or circumstances.
Harmony	The self and one's surroundings are interconnected; individual reads environment and non-verbal and body language well.	The ambience in which one takes the test may influence one's responses; the test taker may be distracted by events taking place during the test.
Movement and verve	Personal conduct is organized through movement. Students are spontaneous, active, energetic, and lively.	Active test-taking strategies may result in better performance than sedentary ones; test taker may have difficulty sitting through and concentrating during lengthy tests.
Affect	Integration of feelings with thoughts and actions; sensitive and emotional.	Feelings may facilitate or hinder test performance; test taker may find it difficult to "understand" persons in test stimuli who act without feeling.
Communalism	Valuing of one's group(s) more than outsiders or other individuals; social; interdependent.	Performance may be influenced when test taker is anxious about the test scores being reflective of his/her cultural group and having negative consequences for them.
Expressiveness	Unique personality is expressed through one's behavioral styles; creative, risk taker, spontaneous.	Test taker may choose the more imaginative response alternative; may be impulsive in choosing responses.
Orality (oral traditional)	Knowledge may be gained and transmitted orally and aurally; a preference to talk and explain verbally.	Test performance may differ when the test taker is tested orally and aurally; test taker may be frustrated by paper-pencil test.
Social Time (polychronicity)	Time is measured by socially meaningful events and customs; person is able to do more than one thing simultaneously.	The belief that obtaining a "good" answer is more important than finishing on time may lead the test taker to "waste" or mismanage time; he/she may not begin responding immediately to the test.

Note. Adapted from "Why Is There No Study of Equivalence in Standardized Cognitive-ability Testing?" by J. Helms, 1992, *American Psychologist, 47*, p. 1096 (copyright © 1992 by the American Psychological Association. Adapted with permission.); also see Boykin (1986).

- *Orality* (also referred to as oral tradition) refers to the emphasis and preference placed on communicating by word of mouth; it is a special sensitivity to aural modes of communication, and an ability to use words to convey meaning and feelings in expressive ways. To speak is to perform, which also entails playing with words and language (e.g., being blunt, and using humor, puns, riddles, proverbs). To speak is to affirm, as denoted by the "call-and-response" mode of communicating.
- *Social time perspective* (also called polychronicity) indicates that time is social and circular; time is not a limited commodity so there is plenty of it; time is to be spent having fun and enjoying others, not worrying about appointments, deadlines, the future, and so forth. Social time perspective is also the realization that nothing in life is guaranteed, so enjoy the moment, the here and now.

These dimensions affect students' communication styles, learning styles, thinking styles, and test-taking styles and skills (Ford & Harris, 1999; Frasier, Hunsaker, Lee, Finely, Frank, et al., 1995, Frasier, Martin, et al., 1995; Helms, 1992). For instance, students for whom movement and verve are strong will be spontaneous, active and energetic; they may have a difficult time sitting through lengthy tests. Students for whom orality is strong may prefer to explain their answers or write essays rather than respond to multiple-choice items; students for whom time perspective is predominant may have difficulty managing their time when taking tests. When students have a communal orientation, they may prefer to work in groups rather than alone; this preference may be interpreted as immaturity, laziness, or cheating by educators who are unfamiliar with cultural diversity (Delgado-Galtan & Trueba, 1985).

The dimensions described by Boykin (1986) are research-based, and they describe many, but not all, Black students. The point here is that culture matters not only when students are learning, but also when they are taking tests; this reality should not be ignored, negated, or minimized with examining the test scores of diverse students. Accordingly, Helms (1992) asks:

1. Is there evidence that the culturally conditioned intellectual skills used by Blacks and Whites generally differ and that these differences have been equivalently incorporated into the measurement procedures?
2. Do Blacks and Whites use the same test-taking strategies when ostensibly responding to the same material, and do these strategies have equivalent meaning?
3. If different strategies are used by the racial groups, to what extent are these differences an aspect of test predictors and test criteria?
4. How does one measure the cultural characteristics of intelligence tests? (Helms, 1992, p. 1097)

The implications of these questions for educators is that, when differences in performance on intelligence tests are attributed to racial or ethnic differences, educators

must recognize this explanation for the non sequitur that it is. Instead of continuing to use such measures until something better comes along, educators must challenge the scientists on whose work their test usage is based to find culturally defined psychological explanations (e.g., culture-specific attitudes, feelings, and behaviors) for why such racial and ethnic differences exist (Helms, 1992, p. 1097).

Sattler (1992) negates the influence of culture on test performance; he states: "Items on intelligence tests represent important aspects of competence in the *common* culture" (p. 568). This statement begs the question: Items on intelligence tests represent important aspects of competence common in *whose* culture? Relative to socio-economic status, children in poverty live in a different culture than children in middle class families. One has only to look at the enriched educational experiences—mainly due to economic opportunity and higher educational backgrounds—that middle class families provide their children compared to families that live in poverty (refer to Figure 1). Further, children who are limited English proficient may not be able to respond to the questions if they do not understand the vocabulary words, if they use the vocabulary words in different ways, and the word does not exist or does not translate in their language. Zigler and Butterfield (1968) concluded: Although ethnic minority children may have an adequate storage and retrieval system to answer questions correctly, they may fail in practice because they have not been exposed to the material (cited in Sattler, 1992).

Lam (1993) discussed five assumptions (or misassumptions) that summarize the many concerns that persist relative to intelligence testing and diverse groups:

1. Test developers assume that test takers have no linguistic barriers (or differences) that inhibit their performance on tests.
2. Test developers assume that the content of the test at any particular level is suitable and of nearly equal difficulty for test takers.
3. Test developers assume that test takers are familiar with or have the test sophistication for taking standardized tests.
4. Test developers assume that test takers are properly motivated to do well on the test.
5. Test developers assume that test takers do not have strong negative psychological reactions to testing.

Although not discussed by Lam (1993) and Gregory (2004), another erroneous but prominent assumption among laypersons and those not familiar with standardized tests is that intelligence tests measure innate ability, and that the tests are not measuring such variables as achievement and the impact of educational experiences and exposure (Fagan & Holland, 2002; Groth-Marnat, 1997; Sternberg, 1982), including instructional quality. According to Sattler (1992), at least six other assumptions about intelligence need to be avoided, as indicated by the counter responses:

Assumption 1: Intelligence tests measure innate intelligence. Counter-assumption—IQs are always based on the individual's interactions with the environment.

Assumption 2: Intelligence tests measure capacity or potential. Counter-assumption—intelligence tests provide information about the individual's repertoire of cognitive abilities and knowledge at a given point in time.

Assumption 3: IQs or IQ scores are fixed and immutable and never change. Counter-assumption—IQ scores change in the course of development, including educational experiences (e.g., as discussed by the notion of the Flynn Effect).

Assumption 4: Intelligence tests provide perfectly reliable scores. Counter-assumption—No intelligence test is perfectly reliable; test scores are only estimates of a person's ability. Every test score should be reported as a statement of probability or odds.

Assumption 5: Intelligence tests measure all we need to know about a person's intelligence. Counter-assumption—No one intelligence test can measure the entire spectrum of abilities related to intellectual behavior. Some tests measure verbal and non-verbal abilities, but do not adequately measure other areas. Any test only *samples* the individual's repertoire of skills.

Assumption 6: IQs obtained from a variety of tests are interchangeable. Counter-assumption—Although there is some overlap in intelligence tests, IQs may not be interchangeable, especially when the standard deviations of the tests are different.

Assumption 7: A battery of tests can tell us everything that we need to know to make judgments about a person's competence. Counter-assumption—No battery of tests can give us a complete picture of any person's abilities.

Having explored statistical and non-statistical biases and concerns, I now turn to alternative tests and other practices that hold promise for assessing diverse students and, ideally, increasing their participation in gifted education programs.

Although many cautions have been shared regarding the limitations of traditional intelligence tests, it is not the intent of this monograph to argue that traditional intelligence tests should not be used with diverse students. Instead, the major premise is that tests should be used responsibly, with diversity and equity in mind. When diversity and equity are considered, tests can provide useful information.

Beyond Traditional, Culturally-loaded Tests: Alternative Tests and Promising Practices

Several scholars have contended that all tests are culturally biased or culturally-loaded in some respects; that no test will be free of culture and, thus, bias (e.g., Joint Standards, 1999; Miller, 1996; Sternberg, 1982). Interestingly, Jensen (1980) distinguished between culturally-loaded and culturally-reduced tests, as illustrated in Table 4. Some characteristics or dimensions of culturally-loaded tests are: paper-pencil tests, printed instructions, oral instructions, reading required, specific factual knowledge,

and more. Conversely, characteristics of culturally-reduced tests include: performance tests; abstract figural items; non-verbal content; and non-scholastic skills; and more. Noted earlier, scholars have also found that when comparing the performance (i.e., non-verbal) and verbal subtest scores of culturally diverse students, many of these students have higher performance scores (Groth-Marnat, 1997). Given these findings, a discussion of non-verbal tests is in order.

Table 4

Culturally-loaded Versus Culturally-reduced Dimensions of Tests

Culture Loaded	***Culture Reduced***
Paper-pencil tests	Performance tests
Printed instructions	Oral instructions
Oral instructions	Pantomime instructions
Reading required	Purely pictorial items
Pictorial (objects)	Abstract figural items
Written response	Oral response
Language	Non-language
Speed test	Power tests
Verbal content	Non-verbal content
Specific factual knowledge	Abstract reasoning
Scholastic skills	Non-scholastic skills
No practice items	Practice items
Recall of past-learned information	Solving novel problems
Content graded from familiar to rare	All time content highly familiar
Difficulty based on rarity of content	Difficulty based on *complexity* of relations to education

Source: Jensen (1980, p. 637).

Non-verbal Tests as Alternative Measures

> Non-verbal tests may provide the only available window into the examinee's verbal reasoning in his or her native language. (Harris, Reynolds, & Koegel, 1998, p. 227)

Traditional standardized intelligence tests are here to stay. However, there is room in assessment practices to consider other types of intelligence tests. That is to say, in addition to the recommendations just described, professionals need to consider alternatives to traditional standardized tests, namely non-verbal tests. Much discussion in the educational and psychological fields has focused on finding culture-free or culture-reduced tests. As stated earlier and illustrated in Table 4, all tests, regardless of their format, will contain content that is influenced by culture. Non-verbal tests are no exception; they are not culture-free. Rather, more than traditional intelligence tests, non-verbal tests are culture-reduced or less culturally-loaded tests.

There are many culturally and linguistically diverse children in our country who may not be considered gifted because they lack the reading, writing, and arithmetic skills typically seen in gifted children and they are identified, in part, by tests of ability that demand school-related knowledge and skills. These students may be gifted students who are low achievers or underachievers. As our 1993 federal definition noted, we must develop talent in students who come from different, less advantaged backgrounds, and we must recognize that giftedness may be demonstrated in some students, while there is potential and promise in other students. School ability tests that have verbal and quantitative sections may put at a disadvantage minority children with limited educational skills and, therefore, these children are more likely to earn lower IQ scores. This problem has led some educators to suggest the use of alternative means of assessment that may have limited validity or reliability (Naglieri & Ford, 2003).

Typically, when the term "non-verbal" is used, it refers to the conditions required for administering the test (e.g., language not required), what the test purports to measure (e.g., non-verbal reasoning), or both (Harris et al., 1998). In its strictest definition, a non-verbal test is any test that does not require the examinee to be literate, nor require written or spoken language from the examinee (Anastasi, 1988; Naglieri & Prewert, 1990). Hence, orally administered tests of vocabulary and comprehension would not be considered non-verbal. The principle guiding non-verbal tests is that no reading or other language variable should influence the individual's score (Naglieri & Prewert, 1990), and recall the earlier discussion of construct irrelevance.

Anastasi (1988) distinguishes between non-reading, non-language, and non-verbal (performance) tests. *Non-reading tests* require no reading or writing by the examinee, although the examiner tends to use oral instructions. *Non-language tests* tend not to require language on the part of either the examiner or examinee. The instructions can be demonstrated, gestured, or pantomimed. *Non-verbal performance tests* require the examinee to perform some action or manipulation of concrete objects, although the intent of the performance is not to measure manipulative skill or manual dexterity per se. Test items often include mazes, copying geometric figures, drawing human figures, and identifying missing pieces or rotated shapes. Directions are typically non-verbal (can be demonstrated).

Non-verbal tests rely less on learning or acquired knowledge than traditional intelligence tests. As discussed in previous pages, using Horn's theory, Sattler (1992) and

Kaufman (1990) noted that non-verbal tests are measures of fluid intelligence rather than crystallized intelligence. Crystallized intelligence is taught or learned. This is an important distinction because students who do well on non-verbal tests (tests of fluid intelligence) may be abstract thinkers who lack academic skills Groth-Marnat (1997). They may be students who have a high IQ score on a non-verbal test, but also have poor reading skills, math skills, and other school-related skills. They are, nonetheless, still "intelligent;"[6] they may be gifted underachievers (see Colangelo & Davis, 2003; Davis & Rimm, 2004; Ford, 1996; Whitmore, 1980 for more discussion on gifted underachievers).

Harris et al. (1998) discussed several potential benefits of non-verbal tests. One essential benefit is that non-verbal assessment can provide a useful cross-check for traditional verbal assessments. Whenever there is a question about the role of language in the assessment process, best practices call for professionals to test the hypothesis that the verbal aspect of the traditional assessment is depressing the examinee's test performance and masking the examinee's potential. Therefore, while the verbal assessment might be an accurate reflection of the examinee's current verbal performance, the scores are invalid for making inferences or generalizations about the examinee's potential. For a more in-depth discussion of characteristics and types of non-verbal tests and subtests, see Harris et al. (1998).

Beyond the types and characteristics of non-verbal tests, an interesting finding is that the 15-point IQ score gap that exists on traditional intelligence tests is not present on non-verbal tests (Saccuzzo, Johnson, & Guertin, 1994). While many non-verbal tests exist, only two have been systemically studied with gifted students (Ravens' Progressive Matrices and Naglieri Non-Verbal Ability Test); for this reason, I focus only on these two tests.

The Raven's Progressive Matrices (Raven, 1947) is the oldest and most widely used non-verbal test. This test has been studied in many countries around the world and with a substantial variety of individuals, including gifted students. Despite its widespread use in the United States, the test has been consistently criticized for its poor psychometric qualities, including the lack of a well-constructed norm group, uneven gradients of item difficulty, inadequate numbers of items, and the need for better documentation of psychometric qualities in the test manual (Jensen, 1980; Nicholson, 1989). Most importantly, however, the difficulty with Raven's Progressive Matrices most relevant to this discussion is findings of higher mean score differences between White and minority children (see Mills & Tissot, 1995; Vincent, 1991).

---

[6] This is not to say that students who are admitted to gifted programs using non-verbal tests do not do as well as those admitted with traditional tests. I am, instead, stating that we need more studies on this issue. I also recognize that some readers will be concerned about admitting students into gifted education programs/classes using non-verbal tests because students may lack academic skills deemed important for doing well (e.g., reading and writing skills). Our 1993 federal definition of giftedness encourages educators to work with students who show gifted potential. This will be many students who do not perform well on traditional intelligence tests, but who do well on non-verbal intelligence tests.

Another non-verbal test is the Naglieri Nonverbal Ability Test (NNAT). It uses the same progressive matrix format as Raven's tests, but there are some important differences between the tests. The NNAT is well standardized on a sample of more than 89,000 students in grades K through 12. The psychometric properties of the test are amply documented (Naglieri, 1997a). Finally, there is a research base on the NNAT and its earlier versions (the MAT-EF and MAT-SF) that support the test's use for diverse populations of children.

Naglieri's progressive matrices tests have a history of yielding small differences between White and minority groups. Naglieri (1985a) summarized the results of two studies involving minority children conducted using the original versions of the NNAT, the MAT-SF and MAT-EF standardization sample. White (n=336) and Black (n=336) children matched on school, gender, and age in years performed similarly (effect size=0.17 or about 2.6 standard score points) on the MAT-SF. Results for the MAT-EF were similar—matched samples of White (n=55) and Black (n=55) children earned standard scores (mean of 100, *SD* of 15) of 90.6 and 90.0, respectively. In other research, the MAT correlated significantly with the Wechsler Intelligence Scale for Children—Revised (WISC-R; Wechsler, 1974) Performance IQ Scale (r=.43, p<.001) and Raven's Progressive Matrices (r=.64, p<.001) for a sample of 114 Native American students (Naglieri, 1985a).

In addition to these initial studies conducted on the first editions of progressive matrices tests by Naglieri (1985a & 1985b), there has been one published study that examined differences between matched samples of White with Black, Hispanic, and Asian American children on the second edition (NNAT; Naglieri, 1997b). In this study, Naglieri and Ronning (2000) examined differences between three matched samples of White (n=2,306) and Black (n=2,306); White (n=1,176) and Hispanic (n=1,176); and White (n=466) and Asian (n=466) children on the NNAT. They found only small differences between the NNAT mean scores for the White and Black samples (*d*-ratio=.25 or about 4 standard score points), and minimal differences between the White and Hispanic (*d*-ratio=.17 or about 2.5 standard score points), as well as White and Asian groups (*d*-ratio=.02 or less than one standard score point). The results also suggested that the NNAT scores had utility for assessment of White and minority children, and that should the NNAT be used for identification of gifted children, similar numbers of each population might be identified.

In a more recent study, Naglieri and Ford (2003) found that, again, the NNAT holds much promise for increasing the representation of diverse students in gifted programs. As Table 5 illustrates, unlike with many traditional intelligence tests, a comparable percentage of diverse students scores at the IQ highest levels on the NNAT, thereby qualifying for gifted programs. The reader is also referred to studies by Saccuzzo et al. (1994), which also demonstrate that the inclusion of non-verbal tests in the assessment process tends to be effective at increasing the representation of diverse students in gifted programs.

Table 5

Numbers and Percentages of Children Who Earned Varying NNAT Standard Scores by Group

	White		Black		Hispanic		Expected
	n	%	*n*	%	*n*	%	%
120 & above	1,571	10.3	269	9.4	190	9.5	9.0
125 & above	906	5.6	145	5.1	88	4.4	5.0
130 & above	467	2.5	75	2.6	46	2.3	2.0
135 & above	190	1.1	42	1.5	18	0.9	1.0
140 & above	90	0.6	19	0.6	9	0.4	0.4
Total Sample *n*	14,141		2,863		1,991		

Note: Expected percentage values are those associated with normal curve probabilities.
From "Addressing the Underrepresentation of Gifted Minority Children Using the Naglieri Nonverbal Ability Test (NNAT)," by J. A. Naglieri and D. Y. Ford, 2003, *Gifted Child Quarterly, 47,* p. 159.

While non-verbal tests have shown to be helpful at increasing the representation or placement of diverse students in gifted programs, the predictive validity of such tests has not been shown. In other words, I have yet to find published data indicating that students placed in such programs do well academically and/or do as well as students placed using traditional intelligence tests. Such published data are certainly needed and long overdue.

Promising Practices and Considerations

In this section, promising practices as considerations are discussed, including culturally sensitive assumptions, diversity training for professionals, test interpretation considerations, and comprehensive data collection.

Culturally Sensitive Assumptions

The accuracy and appropriateness of the intellectual assessment process is based on a number of assumptions, a few of which were discussed earlier. Kaufman (1990, 1994) suggested alternative assumptions worthy of adoption because they offer promise in making testing more culturally sensitive:

1. *The focus on an assessment is the person being assessed, not the test* (Kaufman, 1990). Stated differently, professionals should not become

preoccupied with the IQ scores to the detriment of the individual being assessed. An individual is not best represented by a sum of scores (Suzuki, Vraniak, & Kugler, 1996). Thus, useful information for interpreting and using test scores can be gained when professionals observe students during the assessment process.

2. *The goal of any examiner is to be better than the tests he/she uses* (Kaufman, 1990). It requires knowledge, skills, and cultural competence to make a complete and comprehensive assessment of diverse groups. Professionals should be familiar with the culturally diverse individuals being assessed, be sensitive to culturally diversity, and incorporate this information into the selection of tests and the interpretation of the scores and assessment information.
3. *Intelligence tests measure what the individual has learned* (Kaufman, 1990). The content of all tasks, whether verbal or non-verbal, is learned within a culture (Miller, 1996). Therefore, all tests are culturally-loaded. Individuals bring learning from home, school, and the community to the test taking situation. When interpreting test scores, educators and decision makers must consider the influence of educational background and opportunity to learn the content.
4. *The tasks composing intelligence tests are illustrative samples of behavior and are not meant to be exhaustive* (Kaufman, 1996). Collateral information (e.g., learning styles, motivation, interests, health) must be collected to develop a profile of an individual's strengths and weaknesses and to develop educational interventions and opportunities.
5. *Intelligence tests measure mental functioning under fixed experimental conditions* (Kaufman, 1990). As such, how individuals will demonstrate their intelligence in other settings cannot be accurately predicted without gathering extensive information—test information and non-test information—on individuals in other settings. Essentially, test scores simply assist educators in making conditional probability statements on the basis of the particular test (Frisby, 1998).
6. *IQ tests must be interpreted on an individual basis by a "shrewd and flexible detective"* (Kaufman, 1990). Professionals must investigate all information collected on students to provide a comprehensive picture of the individual in his/her cultural context.
7. *Intelligence tests are best used to generate hypotheses of potential help to the person; they are misused when the results lead to harmful outcomes* (Kaufman, 1990). Too often, data obtained from intelligence tests have been used to indicate the inferiority of culturally diverse groups (see lengthy discussions on this topic by Gould, 1995 and Fancher, 1995). Professionals need to move beyond deficit thinking when assessing diverse populations (Ford et al., 2002; Samuda, 1998). Such thinking is counter-productive, seldom offering constructive information that can be used to guide educational and instructional interventions.

Test Interpretation

Validity and reliability are not only established by test developers, they are also established by test users and interpreters.

Sandoval et al. (1998) offered the following recommendations relative to promoting equitable assessments with diverse groups; these recommendations focus primarily on ways to improve interpretations of diverse students' scores.

Identify preconceptions—Self-awareness is the first step in gaining the capacity to understand others. Professionals must identify their conceptions and viewpoints—negative and positive—about diverse groups, and recognize that these perceptions influence their assessment of diverse groups.

Develop complex schemas or conceptions of groups—A major problem with interpreting the test scores of diverse groups is that results are examined with little regard to the many factors that affect the lives and performance of these groups. Simplistic interpretations of scores are insufficient when complex factors affect test performance.

Actively search for disconfirmatory evidence—When using and interpreting test scores, especially low test scores, of diverse groups, professionals must constantly search for alternative explanations. For example, central questions are: "Did the individual have the opportunity to learn the information or to express it on the test?" "How does the individual's culture affect his/her test performance?" More generically, the question becomes, "Did A cause B or did B cause A, and what is the role of C?"

Resist a rush to judgment—Professionals must be reflective, thoughtful, inquisitive in their practice of interpreting and using test scores with diverse groups. To avoid rushing to judgment, Kaufman (1994) recommended that professionals spend time interacting in the neighborhoods that are serviced by their schools as a firsthand means of learning local cultural values, traditions, and customs.

From Testing to Assessment: Multi-factored and Collateral Data Collection

> Test scores can mislead just as they can lead . . . we must use multiple sources of information and diverse pieces of data. (Kaufman, 1994, p. 13)

> When decisions are made affecting students' educational opportunities and benefits, it is important that they be made accurately, fairly, and comprehensively. When tests are used in making educational decisions for individual students, it is important that they accurately measure students' abilities, knowledge, skills, or needs, and that they do not discriminate . . . . (OCR, 2000, p. 1)

Evaluating the validity of the hypotheses with multiple sources of information, before accepting them as gospel, is the goal of meaningful psychoeducational and clinical assessment (Kaufman, 1994). A minimum, assessment must include scores from

intelligence tests, achievement tests, and criterion-referenced tests, as well as observational and contextual (e.g., cultural) information about diverse students during the assessment process. Also needed is information on school performance (GPA), motivation, interests, values, biographical data, test-taking skills, educational level, thinking and learning styles, communication styles, and language background or dominance, skills and proficiency (Samuda, Feuerstein, Kaufman, Lewis, & Sternberg, 1998).

When multiple types and sources of information are collected and used, we move from identification, which is narrow and limiting, to assessment, which is broad and comprehensive. Too frequently, limited and de-contextualized information is gathered on students, and there is an over-reliance or exclusive reliance on testing; these issues make test interpretation less than effective. Stated differently, because of the appearance of objectivity and numerical precision, test data are sometimes allowed to totally override other sources of evidence about test takers (Joint Standards, 1998). This practice is indefensible.

Diversity Training of Test Developers, Administrators, and Users

"We have become proficient at training testers . . . we need to focus on making them good clinical assessors" (Suzuki et al., 1996, p. 162). Few professionals who develop, use and interpret tests have multicultural training. Thus, they run the risk of misunderstanding, misinterpreting, and misusing test results (Samuda, 1998).

Diversity training among those who develop, administer, and use tests can increase their effectiveness at interpreting and using the test results of diverse students. To repeat, if an interpretive approach relies strictly on one view of the world, no matter how theoretically nor psychometrically defensible that view may be, it is doomed to fail for some children (Kaufman, 1994). Thus, the burden of responsibility for fairness shifts from the test itself to the test user (Samuda, 1998).

Examiners need to be less dependent on the specific scores earned by students and come to the interpretive task armed with research knowledge, theoretical sophistication, and clinical acumen. More bluntly, examiners who are weak in any of these areas are not supposed to *give* the WISC-III, much less interpret it (Kaufman, 1994). In sum, all participants in the testing process must possess the knowledge, skills, and abilities relevant to their role in the testing process, as well as have an awareness of personal and contextual factors that may influence the testing process (Joint Standards, 1999). Several publications, too many to list here, provide frameworks to help educators gain cross-cultural competence and offer some guidance to professionals (see, for example, Banks & Banks, 2004; Ford & Harris, 1999; Nieto, 1999; Storti, 1998).

Adopt Contemporary Definitions and Theories of Intelligence and Giftedness

When all is said and done, the intelligence tests that decision makers choose to adopt will be influenced, in part, by the definitions and theories of intelligence they

espouse. With this in mind, one definition and two theories of intelligence and/or giftedness are presented.

As discussed earlier, in 1993, the federal government, recognizing that our schools are filled with potentially gifted students, proposed a new definition of gifted, one that relies heavily on the notion of talent development:

> Children and youth with outstanding talent perform or show the potential for performing at remarkably high levels of accomplishment when compared with others of their age, experience, or environment.
>
> These children and youth exhibit high performance capacity in intellectual, creative, and/or artistic areas, possess an unusual leadership capacity, or excel in specific academic fields. They require services or activities not ordinarily provided by the schools.
>
> Outstanding talents are present in children and youth from all cultural groups, across all economic strata, and in all areas of human endeavor. (U.S. Department of Education, 1993, p. 3)

This definition, albeit philosophical in nature, urges us to identify and develop students' potential, as well as to recognize gifts and talents in all groups. Related, a number of theories of intelligence and giftedness exist, but two appear to capture the strengths, abilities, and promise of gifted diverse learners, particularly Sternberg's (1985) Triarchic Theory of Intelligence and Gardner's (1983) Theory of Multiple Intelligences. These two comprehensive, flexible, and inclusive theories contend that giftedness is a social construct that manifests itself in many ways and means different things to different cultural groups. The theorists acknowledge the multifaceted, complex nature of intelligence and how current tests (which are too simplistic and static) fail to do justice to this construct.

Summary: Guiding Principles for Equitable and Culturally Responsive Assessment

> Ways must be found to put measurement at the service of diversity. (Samuda, 1998, p. xv)

Regardless of whether one is using traditional intelligence tests or tests considered to be less culturally-loaded, testing, assessment, test interpretation, and test use must be guided by sound, defensible, and equitable principles and practices. Based on the issues described throughout this monograph, the following guiding principles are offered for consideration:

1. Every school system must be committed to equity in finding potentially gifted students; this goal is non-negotiable (Frasier, Martin, et al., 1995).

2. While there are arguments to be made for a purely technical definition of bias and validity, there are strong arguments to be made for the inclusion of politics, values, and culture in considering the full context of test interpretation and test use in which test bias arises (Messick, 1989, cited in Padilla & Medina, 1996). No discussion of test bias is complete when it focuses only on statistical or technical bias.
3. In addition to examining test bias, we must examine test fairness (Gregory, 2004). We must not become complacent in the belief that finding a test to be unbiased means that the test is fair—an unbiased test can still be unfair (Gregoary, 2004). Test bias *and* test fairness should be explored.
4. The effects of threats to a test's validity and reliability must be examined and considered when interpreting and using test scores (Joint Standards, 1999).
5. A given pattern of test performances represents a cross-sectional view of the individual being assessed within a particular context (i.e., ethnic, cultural, familial, social) (Joint Standards, 1999).
6. There is no test score that can tell, ex post facto, the native potential that a student may have had at birth (Samuda, 1998); do not overvalue IQs or treat them as a magical manifestation of a child's inborn potential (Kaufman, 1994); do not over-interpret test scores by assigning them undue power.
7. Test scores should not be allowed to override other sources of evidence about test takers (Joint Standards, 1999).
8. In educational settings, a decision or characterization that will have major impact on a student should not be made on the basis of a single test score (National Association for Gifted Children, 1997). Other relevant information should be taken into account if it will enhance the overall validity of the decision (Joint Standards, 1999); a test taker's score should not be interpreted in isolation; collateral information that may lead to alternative explanations for the examinee's test performance should be considered (Joint Standards, 1999).
9. Comprehensive assessment, the gathering of a wide range of information about test takers, helps to place test scores into a socio-cultural context by considering how an examinee's performance is influenced by acculturation, language proficiency, socioeconomic background, and ethnic/racial identity (Samuda et al., 1998) . . . . Comprehensive assessment is a continuous process and the assessor must learn as much as possible about the test taker's culture . . . and level of acculturation.
10. In educational settings, reports of group differences in test scores should be accompanied by relevant contextual information, where possible, to enable meaningful interpretation of these differences. Where appropriate, contextual information is not available, users should be cautioned against misinterpretation (Joint Standards, 1999).
11. It is the responsibility of those who mandate the use of tests to identify and monitor their impact and to minimize potential negative consequences.

Consequences resulting from the uses of the test, both intended and unintended, should also be examined by the test user (Joint Standards, 1999).

12. In cases where a language-oriented test is inappropriate due to the test takers' limited proficiency in that language, a non-verbal test may be a suitable alternative (Joint Standards, 1999); in situations where linguistic or reading ability is not part of the interested nor targeted construct, the linguistic or reading demands of the test should be kept to the minimum necessary for the valid assessment of the intended construct (Joint Standards, 1999). Thus, both verbal and non-verbal tests can provide balanced and important information about diverse students (Samuda et al., 1998).
13. When interpreting test scores, the examiner or tester must take into account that many traditional tests have not been normed adequately with various cultural groups (Samuda et al., 1998); test users must be constantly aware of the limitations of standardized tests (Kaufman, 1994).
14. Validation is the joint responsibility of the test developer and test user (Joint Standards, 1999).
15. The ultimate responsibility for appropriate test use and interpretation lies predominantly with test users (Joint Standards, 1999, p. 112); they must gain experience in working with culturally diverse groups to improve their ability to interpret and effectively use test scores (Kaufman, 1994).
16. Tests selected should be suitable for the characteristics and background of the test taker (Joint Standards, 1999). Test scores must not be interpreted and used in a color-blind or culture-blind fashion (Ford, 1996).
17. Every effort must be made to eliminate prejudice, racism, and inequities and to provide accurate and meaningful scores linked to appropriate intervention strategies (Samuda et al., 1998). Essentially, test scores should be used to help students, not to hurt them. Tests can be helpful to diverse students if they do not serve as gatekeepers and/or the first barrier to keeping diverse students out of gifted education classes.

A Word on Test Fairness

In sections entitled "A Reprise on Test Bias" and "A Reprise on Test Fairness," Gregory (2004) suggested that scholars should focus not only on the concept of "test bias," but also on the concept of "test fairness." He argued that the understanding of test bias is made difficult by the implicit and often emotional assumptions that may lead people to view the same information in different ways. Further, disagreements about test bias are perpetuated because adversaries in the debate fail to clarify essential terminology. Accordingly, to many people, terms such as "test bias" and "test fairness" are considered interchangeable and used without regard to definition.

In contemporary psychology and related fields, test bias refers to objective statistical indices that examine the patterning of test scores for relevant subpopulations. While experts might disagree about nuances, on the whole, there is, according to Gregory

(2004), consensus about the statistical criteria that indicate when at test is biased. Gregory goes on to clarify that test fairness is a broad concept that recognizes importance of social values in test usage. Thus, even a test that is unbiased according to technical criteria might still be deemed unfair because of the social consequences of using it for selection or placement decisions.

Conclusion

Selecting, interpreting, and using tests are complicated endeavors. When one adds student differences, including cultural diversity, to the situation, the complexity increases. A discussion on the nature-nurture debate was presented briefly. Little attention was given to this controversy because the discussion is convoluted—for every publication that convincingly argues for the heredity position, an equally compelling publication argues for the environmental position. Likewise, for every publication that argues persuasively against the existence of test bias, a counterargument convincingly contends that tests continue to be biased against diverse groups.

There is no debate, however, that culturally and linguistically diverse students are consistently under-represented in gifted programs. In this monograph, it was argued that under-representation exists primarily because of diverse students' performance on traditional intelligence tests. These tests have served as gatekeepers for diverse students. Accordingly, this monograph focused only on intelligence tests because of the assumptions, misassumptions, perceptions, and misperceptions about the origins of intelligence and the debates surrounding test fairness and appropriateness with diverse groups.

Since its inception, gifted education has had an under-representation of diverse students in its programs and services. With tests being used so extensively in the decision-making process, it seems impossible that one would (or could) ignore their role as gatekeepers. Wiggins (1989) stated: "When an educational problem persists despite the well-intentioned efforts of many people to solve it, it's a safe bet that the problem hasn't been properly frames" (p. 703). Given the array of unresolved assessment issues regarding the identification of talent potential among minority students, the probability is raised that the questions being asked need reframing (Frasier, García, & Passow, 1995).

Too often, tests have been touted to be objective and color-blind. In what ways does ignoring and minimizing the role of social variables and culture in testing contribute to the under-representation of diverse students in gifted education? Suggestions for ensuring equitable, culturally responsive assessment practices were provided, along with attention to alternative tests—non-verbal ability tests.

Professionals must be vigilant about finding and solving factors that hinder the test performance of diverse students. Tests are tools. The ultimate responsibility for equitable assessment rests with those who develop, administer, interpret, and use tests. Tests in and of themselves are harmless; they become harmful when misunderstood and

misused. Historically, diverse students have been harmed educationally by test misuse. The pedagogical clock is ticking. What better time than today to be more responsible in eliminating barriers to the representation of diverse students in gifted education. A mind is a terrible thing to waste; a mind is a terrible thing to erase (Ford & Harris, 1999).

References

Aiken, L. R. (2000). *Psychological testing and assessment* (10th ed.). Boston: Allyn and Bacon.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

Anastasi, A. (1968). *Psychological testing* (2nd ed.). New York: Macmillan.

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

Armour-Thomas, E. (1992). Intellectual assessment of children from culturally diverse backgrounds. *School Psychology Review, 21,* 552-565.

Armour-Thomas, E., & Gopaul-McNicol, S. (1998). *Assessing intelligence: Applying a bio-cultural model*. Thousand Oaks, CA: Sage.

Baldwin, A. Y., & Vialle, W. (1999). *The many faces of giftedness: Lifting the masks*. Belmont, CA: Wadsworth.

Banks, J. A. (1995). Multicultural education: Historical development, dimensions, and practice. In J. A. Banks & C. A. M. Banks (Eds.), *Handbook of research on multicultural education* (pp. 3-24). New York: Macmillan.

Banks, J. A., & Banks, C. A. M. (Eds.). (2004). *Multicultural education: Issues and perspectives*. Hoboken, NJ: John Wiley and Sons.

Boykin, A. W. (1986). The triple quandary and the schooling of Afro-American children. In U. Neisser (Ed.), *The school achievement of minority children* (pp. 57-91). Hillsdale, NJ: Lawrence Erlbaum.

Boykin, A. W. (1994). Afrocultural expression and its implications for schooling. In E. R. Hollins, J. E. Kings, & W. C. Hayman (Eds.). *Teaching diverse populations: Formulating a knowledge base* (pp. 225-273). New York: Teachers College Press.

Callahan, C. M., & McIntyre, J. A. (1994). *Identifying outstanding talent in American Indian and Alaska Native students*. Washington, DC: U.S. Department of Education.

Castellano, J. A. (2003). *Special populations in gifted education: Working with diverse gifted learners*. Boston: Allyn and Bacon.

Cline, S., & Schwartz, D. (Eds.). (1999). *Diverse populations of gifted children: Meeting their needs in the regular classroom and beyond.* Columbus, OH: Merrill/Prentice Hall.

Colangelo, N., & Davis, G. A. (2003). *Handbook of gifted education* (3rd ed.). Boston: Allyn and Bacon.

Council of State Directors of Program for the Gifted and National Association for Gifted Children. (2003). *State of the states gifted and talented education report*, 2001-2002. Washington, DC: National Association for Gifted Children.

Davis, G. A., & Rimm, S. B. (2004). *Education of the gifted and talented* (5th ed.). Boston: Allyn and Bacon.

Delgado-Galtan, C., & Trueba, H. T. (1985). Ethnographic study of the participant structures in task completion: Reinterpretation of "handicaps" in Mexican children. *Learning Disability Quarterly, 8*, 67-75.

Delpit, L. (1995). *Other people's children: Cultural conflict in the classroom.* New York: The New Press.

Erickson, F. (2004). Culture in society and in educational practices. In J. A. Banks & C. A. M. Banks (Eds.), *Multicultural education: Issues and perspectives* (5th ed., pp. 31-55). Hoboken, NJ: John Wiley and Sons.

Fagan, J. F., & Holland, C. R. (2002). Equal opportunity and racial differences in IQ. *Intelligence, 30*, 361-387.

Fancher, R. E. (1995). *The intelligence men: Makers of the IQ controversy.* New York: W. W. Norton.

Fishman, J. A., Deutsch, M., Kogan, L., North, R., & Whiteman, M. (1964). Guidelines for testing minority group children. *Journal of Social Issues Supplement, 20,* 129-145.

Ford, D. Y. (1996). *Reversing underachievement among gifted Black students: Promising practices and programs.* New York: Teachers College Press.

Ford, D. Y. (1998). The under-representation of minority students in gifted education: Problems and promises in recruitment and retention. *The Journal of Special Education, 32*(1), 4-14.

Ford, D. Y., & Harris, J. J., III. (1999). *Multicultural gifted education.* New York: Teachers College Press.

Ford, D. Y., Harris, J. J., III, Tyson, C. A., & Frazier Trotman, M. (2002). Beyond deficit thinking: Providing access for gifted African American students. *Roeper Review, 24,* 52-58.

Frasier, M. M., García, J. H., & Passow, A. H. (1995). *A review of assessment issues in gifted education and their implications for identifying gifted minority students* (RM95204). Storrs, CT: The National Research Center on the Gifted and Talented, University of Connecticut.

Frasier, M. M., Hunsaker, S. L., Lee, J., Finley, V. S., Frank, E., García, J. H., & Martin, D. (1995). *Educators' perceptions of barriers to the identification of gifted children from economically disadvantaged and limited English proficient backgrounds* (RM95216). Storrs, CT: The National Research Center on the Gifted and Talented, University of Connecticut.

Frasier, M. M., Hunsaker, S. L., Lee, J., Finely, V. S., García, J. H., Martin, D., & Frank, E. (1995). *An exploratory study of the effectiveness of the staff development model and the research-based assessment plan in improving the identification of gifted economically disadvantaged students* (RM95224). Storrs, CT: The National Research Center on the Gifted and Talented, University of Connecticut.

Frasier, M. M., Hunsaker, S. L., Lee, J., Mitchell, S., Cramond, B., Krisel, S., García, J. H., Martin, D., Frank, E., & Finley, V. S. (1995). *Core attributes of giftedness: A foundation for recognizing the gifted potential of minority and economically disadvantaged students* (RM95210). Storrs, CT: The National Research Center on the Gifted and Talented, University of Connecticut.

Frasier, M. M., Martin, D., García, J. H., Finley, V. S., Frank, E., Krisel, S., & King, L. L. (1995). *A new window for looking at gifted children* (RM95222). Storrs, CT: The National Research Center on the Gifted and Talented, University of Connecticut.

Frisby, C. L. (1998). Culture and cultural differences. In J. Sandoval, C. L. Frisby, F. K. Geisinger, J. D. Scheuneman, & J. R. Grenier (Eds.), *Test interpretation and diversity: Achieving equity in assessment* (pp. 51-73). Washington, DC: American Psychological Association.

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences.* New York: Basic Books.

Gould, S. J. (1995). *The mismeasure of man.* New York: Norton.

Gregory, R. J. (2004). *Psychological testing: History, principles and applications* (3rd ed.). Boston: Allyn and Bacon.

Gronlund, N. E. (1981). *Measurement and evaluation in teaching*. New York: MacMillan.

Groth-Marnat, G. (1997). *Handbook of psychological assessment* (3rd ed.). New York: John Wiley & Sons.

Groth-Marnat, G. (2003). *Handbook of psychological assessment* (4th ed.). New York: John Wiley & Sons.

Hale, J. E. (2001). *Learning while Black: Creating educational excellence for African American children*. Baltimore: Johns Hopkins.

Hall, E. T. (1959). *The silent language*. New York: Doubleday.

Hall, E. T. (1976). *Beyond culture*. New York: Doubleday.

Harmon, D. (2002). They won't teach me: The voices of gifted African American inner-city students. *Roeper Review, 24,* 68-75.

Harris, A. M., Reynolds, M. A., & Koegel, H. M. (1998). Nonverbal assessment: Multicultural perspectives. In L. A. Suzuki, P. J. Meller, & J. G. Ponterotto (Eds.), *Handbook of multicultural assessment: Clinical, psychological, and educational applications* (pp. 223-252). San Francisco: Jossey-Bass Publishers.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Baltimore: Paul H. Brookes.

Helms, J. (1992). Why is there no study of equivalence in standardized cognitive-ability testing? *American Psychologist, 47*, 1083-1101.

Herrnstein, R. J., & Murray, C. (1994). *The bell curve: Intelligence and class structure in American life*. New York: Free Press.

Hilliard, A. G., III. (Ed.). (1991). *Testing African American students*. Morristown, NJ: Aaron Press.

Jencks, C., & Phillips, M. (Eds.). (1998). *The Black-White test score gap*. Washington, DC: The Brookings Institute.

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

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

Jensen, A. R. (1998). *The g factor*. Westport, CT: Praeger.

Jensen, A. R. (2000). Testing: the dilemma of group differences. *Psychology, Public Policy, and Law, 6*, 121-127.

Johnsen, S. K. (2004). *Identifying gifted students: A practical guide*. Waco, TX: Prufrock Press.

Jones, R. L. (Ed.). (1996). *Handbook of tests and measurements for Black populations* (Vols. 1 and 2). Hampton, VA: Cobb & Henry Publishers.

Kaufman, A. S. (1990). *Assessing adolescent and adult intelligence*. Needham Heights, MA: Allyn and Bacon.

Kaufman, A. S. (1994). *Intelligent testing with the WISC-III*. New York: John Wiley & Sons.

Kazdin, A. E. (1992). *Research design in clinical psychology* (2nd ed.). Needham Heights, MA: Simon & Schuster.

Korchin, S. J. (1980). Clinical psychology and minority populations. *American Psychologist, 35*, 262-269.

Labov, W. (1972). *Language in the inner city*. Philadelphia: University of Pennsylvania.

Labov, W. (1982). Objectivity and commitment in linguistic science: The case of the Black English trial in Ann Arbor. *Language in Society, 11*, 165-201.

Lam, T. C. M. (1993). Testability: A critical issue in testing language minority students with standardized achievement tests. *Measurement and Evaluation in Counseling and Development, 26*, 179-191.

Larry P. v. Riles (1979, October). NO. C-712270 RFP (N. C. Cal.).

Maker, J., & Schiever, S. W. (Eds.). (1989). *Critical issues in gifted education: Defensible programs for cultural and ethnic minorities* (Vol. II). Austin, TX: Pro-Ed.

Mensh, E., & Mensh, H. (1991). *The IQ mythology: Class, race, gender and inequality*. Carbondale, IL: Southern Illinois University Press.

Messick, S. (1989). Validity. In R. L. Lynn (Ed.), *Educational measurement*.(3rd ed., pp. 13-103). New York: Macmillan.

Miller, J. G. (1996). A cultural-psychological perspective on intelligence. In R. J. Sternberg & E. L. Grigorenko (Eds.), *Intelligence, heredity, and environment* (pp. 269-302). New York: Cambridge University Press.

Mills, C. J., & Tissot, S. L. (1995). Identifying academic potential in students from under-represented populations: Is using the Raven's Progressive Matrices a good idea? *Gifted Child Quarterly, 39,* 209-217.

Montagu, A. (1999). *Race and IQ* (expanded ed.). New York: Oxford University Press.

Naglieri, J. A. (1985a). *Matrix Analogies Test—Expanded Form*. San Antonio, TX: The Psychological Corporation.

Naglieri, J. A. (1985b). *Matrix Analogies Test—Short Form*. San Antonio, TX: The Psychological Corporation.

Naglieri, J. A. (1997a). *NNAT multilevel technical manual*. San Antonio, TX: The Psychological Corporation.

Naglieri, J. A. (1997b). *Naglieri Nonverbal Ability Test*. San Antonio, TX: The Psychological Corporation.

Naglieri, J. A., & Ford, D. Y. (2003). Addressing the underrepresentation of gifted minority children using the Naglieri Nonverbal Ability Test (NNAT). *Gifted Child Quarterly, 47,* 155-160.

Naglieri, J. A., & Prewett, P. N. (1990). Nonverbal intelligence: A selected review of instruments and their use. In R. W. Kamphaus & C. R. Reynolds (Eds.), *Handbook of psychological and educational assessment: Volume I, intelligence and achievement* (pp. 348-370). New York: Guilford Press.

Naglieri, J. A., & Ronning, M. E. (2000). Comparison of White, African-American, Hispanic, and Asian Children on the Naglieri Nonverbal Ability Test. *Psychological Assessment, 12,* 328-334.

National Association for Gifted Children. (1997). *Position paper on testing*. Washington, DC: Author.

Nicholson, C. L. (1989). Matrix Analogies Test (MAT). *Diagnostique, 15,* 115-123.

Nieto, S. (Ed.). (1999). *The light in their eyes: Creating multicultural learning communities*. New York: Teachers College Press.

Office for Civil Rights. (1998). *OCR elementary and secondary civil rights survey: 1998*. Retrieved September 1, 2004, http://205.207.175.80/ocrpublic/wds_list98P.asp .

Office for Civil Rights. (2000). *The use of tests as part of high-stakes decision-making for students: A resource guide for educators and policy-makers*. Washington, DC: Author.

Olmedo, E. L. (1981). Testing linguistic minorities. *American Psychologist, 36,* 1078-1085.

Padilla, A. M., & Medina, A. (1996). Cross-cultural sensitivity in assessment: Using tests in culturally appropriate ways. In L. A. Suzuki, J. P. Meller, & J. G. Ponterotto (Eds.), *Handbook of multicultural assessment: Clinical, psychological, and educational applications* (pp. 3-28). San Francisco: Jossey-Bass.

Raven, J. C. (1947). *Coloured Progressive Matrices.* London: H. K. Lewis.

Reynolds, C. R. (1998). Cultural bias in testing of intelligence and personality. In A. Bellack & M. Hersen (Series Eds.), & C. Belar (Vol. Ed.), *Comprehensive clinical psychology: Vol. 10—Sociocultural and individual differences* (pp. 52-92). New York: Elsevier Science.

Rodriquez, E. R., & Bellanca, J. (1996). *What is it about me you can't teach: An instructional guide for the urban educator.* Arlington Heights, IL: SkyLight.

Rushston, J. P. (2003). Brain size, IQ and racial-group differences: Evidence from musculoskeletal traits. *Intelligence, 31*(2), 139-155.

Saccuzzo, D. P., Johnson, N. E., & Guertin, T. L. (1994). *Identifying underrepresented disadvantaged gifted and talented children: A multifaceted approach* (Vols. 1 and 2). San Diego, CA: San Diego State University.

Samuda, R. J. (1998). *Psychological testing of American minorities: Issues and consequences* (2nd ed.). Thousand Oaks, CA: Sage.

Samuda, R. J., Feuerstein, R., Kaufman, A. S., Lewis, J. E., & Sternberg, R. J. (1998). *Advances in cross-cultural assessment.* Thousand Oaks, CA: Sage.

Sandoval, J., Frisby, C. L., Geisinger, K. F., Scheuneman, J. D., & Grenier, J. R. (Eds.). (1998). *Test interpretation and diversity: Achieving equity in assessment.* Washington, DC: American Psychological Association.

Sattler, J. M. (1992). *Assessment of children* (Rev. ed.). San Diego: Jerome M. Sattler Publisher.

Shade, B., Kelly, C., & Oberg, M. (1997). *Creating culturally responsive classrooms.* Washington, DC: American Psychological Association.

Smith, C., Constantino, R., & Krashen, S. (1997). Differences in print environment for children in Beverly Hills, Compton, and Watts. *Emergency Librarian, 24*(4), 8-9.

Smitherman, G. (1977). *Talking and testifying: The language of Black America*. Boston: Houghton Mifflin.

Smitherman, G. (1999). *Talkin that talk: Language, culture and education in African America*. New York: Routledge.

Steele, C. M. (1999, August). Thin ice: "Stereotype threat" and Black college students. *The Atlantic Online,* 1-6.

Steele, C. M., & Aronson, J. (1998). In C. Jencks & M. Phillips (Eds.), *The Black-White test score gap* (pp. 401-427). Washington, DC: The Brookings Institute.

Sternberg, R. J. (1982). *Handbook of human intelligence*. New York: Cambridge University Press.

Sternberg, R. J., & Grigorenko, E. L. (Eds.). (1997). *Intelligence, heredity, and environment*. New York: Cambridge University Press.

Storti, C. (1998). *The art of crossing cultures*. Yarmouth, MN: Intercultural Press.

Suzuki, L. A., Vraniak, D. A., & Kugler, J. F. (1996). Intellectual assessment across cultures. In L. A. Suzuki, P. J. Meller, & J. G. Ponterotto (Eds.), *Handbook of multicultural assessment: Clinical, psychological, and educational applications* (pp. 141-177). San Francisco: Jossey-Bass.

Suzuki, L. A., Meller, P. J., & Ponterotto, J. G. (Eds.). (1996). *Handbook of multicultural assessment: Clinical, psychological, and educational applications*. San Francisco: Jossey-Bass.

Taylor, O. L. (1990). *Cross-cultural communication: As essential dimension of effective education*. Washington, DC: The Mid-Atlantic Equity Center.

Thorndike, R. I., Hagen, E. P., & Sattler, J. M. (1986). *Stanford-Binet Intelligence Scale—Fourth Edition*. Itasca, IL: Riverside Publishing.

Tomlinson, C. A., Ford, D.Y., Reis, S. M., Briggs, C. J., & Strickland, C. A. (2004). *In search of the dream: Designing schools and classrooms that work for high potential students from diverse cultural backgrounds*. Washington, DC: National Association for Gifted Children and The National Research Center on the Gifted and Talented.

U.S. Department of Education. (1993). *National excellence: A case for developing America's talent*. Washington, DC: Author.

U.S. Department of Education. (1998). *Talent and diversity: The emerging world of limited English proficient students in gifted education*. Washington, DC: Author.

U.S. Department of Education, National Center for Education Statistics. (2003). *Status and trends in the education of Blacks*. Washington, DC: Author.

Valencia, R. R., & Suzuki, L. A. (2001). *Intelligence testing and minority students: Foundations, performance factors, and assessment issues*. Thousand Oaks, CA: Sage.

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

Walsh, W. B., & Betz, N. E. (1995). *Tests and assessment* (4th ed.). Indianapolis, IN: Prentice Hall.

Wasserman, J. D., & Becker, K. A. (2000). *Racial and ethnic group mean score differences on intelligence tests*. Paper presented at the American Psychological Association convention, Washington, DC.

Wechsler, D. (1974). *Wechsler Intelligence Scale for Children—Revised*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1991). *Wechsler Intelligence Scale for Children—Third Edition*. San Antonio, TX: The Psychological Corporation.

Whitmore, J. R. (1980). *Giftedness, conflict, and underachievement*. Boston: Allyn and Bacon.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 70,* 703-713.

Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock-Johnson Revised tests of achievement: Standard and supplemental batteries*. Itasca, IL: Riverside Publishing.

Zigler, E. F., & Butterfield, E. C. (1968). Motivational aspects of changes in IQ test performance of culturally deprived nursery school children. *Child Development, 39,* 1-14.

Research Monograph

The National Research Center on the Gifted and Talented
University of Connecticut
2131 Hillside Road Unit 3007
Storrs, CT 06269-3007
www.gifted.uconn.edu

Also of Interest

State Policies Regarding Education of the Gifted as Reflected in Legislation and Regulation
A. Harry Passow and Rose A. Rudnitski

Residential Schools of Mathematics and Science for Academically Talented Youth: An Analysis of Admission Programs
Fathi A. Jarwan and John F. Feldhusen

The Status of Programs for High Ability Students
Jeanne H. Purcell

Recognizing Talent: Cross-Case Study of Two High Potential Students With Cerebral Palsy
Colleen Willard-Holt

Also of interest from the

Research Monograph Series

The Prism Metaphor: A New Paradigm for Reversing Underachievement
Susan M. Baum, Joseph S. Renzulli, and Thomas P. Hébert

Attention Deficit Disorders and Gifted Students: What Do We Really Know?
Felice Kaufmann, M. Layne Kalbfleisch, and F. Xavier Castellanos

Gifted African American Male College Students: A Phenomenological Study
Fred A. Bonner, II

Counseling Gifted and Talented Students
Nicholas Colangelo

E. Paul Torrance: His Life, Accomplishments, and Legacy
Thomas P. Hébert, Bonnie Cramond, Kristie L. Speirs Neumeister, Garnet Millar, and Alice F. Silvian

The Effects of Grouping and Curricular Practices on Intermediate Students' Math Achievement
Carol L. Tieso

Developing the Talents and Abilities of Linguistically Gifted Bilingual Students: Guidelines for Developing Curriculum at the High School Level
Claudia Angelelli, Kerry Enright, and Guadalupe Valdés

Development of Differentiated Performance Assessment Tasks for Middle School Classrooms
Tonya R. Moon, Carolyn M. Callahan, Catherine M. Brighton, and Carol A. Tomlinson

Society's Role in Educating Gifted Students: The Role of Public Policy
James J. Gallagher

Middle School Classrooms: Teachers' Reported Practices and Student Perceptions
Tonya R. Moon, Carolyn M. Callahan, Carol A. Tomlinson, and Erin M. Miller

Assessing and Advocating for Gifted Students: Perspectives for School and Clinical Psychologists
Nancy M. Robinson

Giftedness and High School Dropouts: Personal, Family, and School Related Factors
Joseph S. Renzulli and Sunghee Park

Also of interest from the

Research Monograph Series

Assessing Creativity: A Guide for Educators
Donald J. Treffinger, Grover C. Young, Edwin C. Selby, and Cindy Shepardson

Implementing a Professional Development Model Using Gifted Education Strategies With All Students
E. Jean Gubbins, Karen L. Westberg, Sally M. Reis, Susan T. Dinnocenti, Carol L. Tieso, Lisa M. Muller, Sunghee Park, Linda J. Emerick, Lori R. Maxfield, and Deborah E. Burns

Teaching Thinking to Culturally Diverse, High Ability, High School Students: A Triarchic Approach
Deborah L. Coates, Tiffany Perkins, Peter Vietze, Mariolga Reyes Cruz, and Sin-Jae Park

Advanced Placement and International Baccalaureate Programs for Talented Students in American High Schools: A Focus on Science and Mathematics
Carolyn M. Callahan

The Law on Gifted Education
Perry A. Zirkel

School Characteristics Inventory: Investigation of a Quantitative Instrument for Measuring the Modifiability of School Contexts for Implementation of Educational Innovations
Tonya R. Moon, Catherine M. Brighton, Holly L. Hertberg, Carolyn M. Callahan, Carol A. Tomlinson, Andrea M. Esperat, and Erin M. Miller

Content-based Curriculum for Low Income and Minority Gifted Learners
Joyce VanTassel-Baska

Reading Instruction for Talented Readers: Case Studies Documenting Few Opportunities for Continuous Progress
Sally M. Reis, E. Jean Gubbins, Christine Briggs, Fredric J. Schreiber, Susannah Richards, Joan Jacobs, Rebecca D. Eckert, Joseph S. Renzulli, and Margaret Alexander

Issues and Practices in the Identification and Education of Gifted Students From Under-represented Groups
James H. Borland

Also of interest from the

Research Monograph Series

The Social and Emotional Development of Gifted Students
Carolyn M. Callahan, Claudia J. Sowa, Kathleen M. May, Ellen Menaker Tomchin, Jonathan A. Plucker, Caroline M. Cunningham, and Wesley Taylor

Promoting Sustained Growth in the Representation of African Americans, Latinos, and Native Americans Among Top Students in the United States at All Levels of the Education System
L. Scott Miller

Evaluation, Placement, and Progression: Three Sites of Concern for Student Achievement
Samuel R. Lucas

Latino Achievement: Identifying Models That Foster Success
Patricia Gándara

Modern Theories of Intelligence Applied to Assessment of Abilities, Instructional Design, and Knowledge-based Assessment
Robert J. Sternberg, Elena L. Grigorenko, Bruce Torff, and Linda Jarvin

Giftedness and Expertise
Robert J. Sternberg, Elena L. Grigorenko, and Michel Ferrari

Academic and Practical Intelligence
Robert J. Sternberg, Elena L. Grigorenko, Jerry Lipka, Elisa Meier, Gerald Mohatt, Evelyn Yanez, Tina Newman, and Sandra Wildfeuer

Developing Creativity in Gifted Children: The Central Importance of Motivation and Classroom Climate
Beth A. Hennessey

NRC G/T

The National Research Center on the Gifted and Talented Research Teams

University of Connecticut

Dr. Joseph S. Renzulli, Director
Dr. E. Jean Gubbins, Associate Director
Dr. Sally M. Reis, Associate Director
University of Connecticut
2131 Hillside Road Unit 3007
Storrs, CT 06269-3007
860-486-4676

Dr. Del Siegle

University of Virginia

Dr. Carolyn M. Callahan, Associate Director
Curry School of Education
University of Virginia
P.O. Box 400277
Charlottesville, VA 22904-4277
804-982-2849

Dr. Mary Landrum
Dr. Tonya Moon
Dr. Carol A. Tomlinson
Dr. Catherine M. Brighton
Dr. Holly L. Hertberg

Yale University

Dr. Robert J. Sternberg, Associate Director
Yale University
Center for the Psychology of Abilities, Competencies, and Expertise
340 Edwards Street, P.O. Box 208358
New Haven, CT 06520-8358

Dr. Elena L. Grigorenko