Interpretation of Intelligence Test Scores in Atkins Cases: Conceptual and
Psychometric Issues

Frank M. Gresham [a]

[a] Department of Psychology, Louisiana State University, Baton Rouge, Louisiana

PLEASE SCROLL DOWN FOR ARTICLE

Psychology Press
Taylor & Francis Group

Interpretation of Intelligence Test Scores in *Atkins* Cases: Conceptual and Psychometric Issues

Frank M. Gresham

Department of Psychology, Louisiana State University, Baton Rouge, Louisiana

So-called *Atkins* cases refer to individuals who have been sentenced to death for capital crimes who claim that the death penalty constitutes "cruel and unusual punishment" under the Eighth Amendment. Psychological testimony is influential because this testimony strikes at the very core issue in these cases; namely, whether or not the individual is mentally retarded. Despite the importance of psychological testimony, courts have not been made to understand the subtleties and complexities of the issues in diagnosing mental retardation. Five such issues are discussed in this article: (a) the nature of intellectual functioning, (b) the Flynn Effect, (c) measurement error, (d) practice effects, and (e) the nature of school "diagnoses." Examples of each of these issues are illustrated with an actual Atkins case (*Walker v. True*, 2006).

Key words: *Flynn Effect, intelligence, measurement error, psychometric*

Around midnight on August 16, 1996, Daryl Renard Atkins and an accomplice (William Jones) abducted Erich Nesbitt with a semiautomatic handgun and robbed him of his money. Subsequently, they drove Nesbitt to an ATM and forced him to withdraw cash. He was then taken to an isolated spot where he was shot eight times and killed. During trial, both Atkins and Jones testified and confirmed each other's account of the incident, except that Jones' testimony was considered more credible than Atkins'. In fact, Atkins' court testimony was substantially inconsistent with the testimony he gave police upon his arrest, whereas Jones declined to make a statement to authorities upon his arrest (Miranda Rights). During the penalty phase of the trial, the defense relied on Dr. Evan Nelson, a forensic psychologist, who had evaluated Atkins prior to trial and concluded that he was mildly mentally retarded based on a review of school and court records and a tested full scale IQ of 59 on the Wechsler Adult Intelligence Scale-III (WAIS-III). Atkins, however, was sentenced to death, and Jones plea bargained with

the prosecution in return for testimony against Atkins and was spared the death penalty.

At a second sentencing hearing, another forensic psychologist, Dr. Stanton Samenow, expressed the opinion that Atkins was not mentally retarded and was functioning in the range of "average" intelligence. This opinion was based on two interviews with Atkins, a review of school records, the Wechsler Memory Scale (Wechsler, 1972), and interviews with correctional officers. Dr. Samenow did not administer an intelligence test but opined that Atkins' poor academic performance while in school was due to his frequent inattention and his overall tendency toward noncompliance in school.

How can two board-certified, licensed, forensic psychologists come to two diametrically opposed opinions regarding the presence or absence of mental retardation? Atkins' measured intelligence was over 2.7 standard deviations below the mean, which almost pushed him into the moderate range of mental retardation (American Psychiatric Association, 2000). Despite this fact, the prosecution's psychologist considered Atkins to be of average intelligence. This finding, as is demonstrated throughout this special issue, is neither unusual nor unexpected for a variety of reasons that will be discussed

Address correspondence to Frank M. Gresham, Department of Psychology, Louisiana State University, Baton Rouge, LA 70803. E-mail: gresham@lsu.edu

in this article. I start with a very brief overview of mental retardation, particularly mild mental retardation, and continue with a discussion of interpretative and psychometric issues in the assessment of intelligence. The article concludes with recommendations for psychologists who may one day find themselves as experts in *Atkins* cases.

MENTAL RETARDATION

Mental retardation is defined by most organizations and states as significantly subaverage intellectual functioning that concurrently exists with deficits in adaptive behavior and which has an onset prior to age 18. The *Diagnostic and Statistical Manual of Mental Disorders* (DSM-IV; American Psychiatric Association, 2000) specifies that significantly subaverage intellectual functioning should be two standard deviations below the mean, however, it acknowledges that the existence of five points in measurement error should be considered in making a diagnosis of mental retardation. As such, it is possible to diagnose an individual as having mental retardation with an IQ up to 75 if they also have substantial deficits in adaptive behavior. Adaptive behavior refers to how well an individual copes with life demands and how well they meet the standards of personal independence expected of someone in their age group, sociocultural background, and community setting (APA, 2000). DSM-IV specifies four degrees of severity for mental retardation: mild mental retardation (IQ 50–55 to 70–75), moderate mental retardation (IQ 35–40 to 50–55), severe mental retardation (IQ 20–25 to 35–40) and profound mental retardation (IQ below 20 or 25). As will be described later, the debate in the *Atkins* cases has never been about individuals with moderate, severe, or profound mental retardation. It has always been about persons who might be considered to have *mild mental retardation*.

The American Association on Intellectual and Developmental Disabilities (AAIDD, 2002) defines an intellectual disability as being characterized by significant limitations both in intellectual functioning and in adaptive behavior as expressed in conceptual, social, and practical adaptive skills and originates before the age of 18. Similar to DSM-IV, significant limitations in intellectual functioning is defined as performance that is two standard deviations below the mean (70–75 and below); imitations in adaptive behavior is defined as performance that is at least two standard deviations below the mean in one of the three adaptive behavior domains (conceptual, social, or practical) or a total adaptive behavior (composite) score on a standardized adaptive behavior measure (see Greenspan and Reschly's discussion of adaptive behavior, this issue). Unlike DSM-IV, however,

AAIDD does not classify mental retardation by severity (mild, moderate, severe, or profound), but rather uses the concept of levels of supports needed to promote the development, education, interests, and personal well-being of an individual with intellectual disability.

An extremely important issue in *Atkins* cases that is often misunderstood by the courts is the nature of mild mental retardation (MMR) as being distinct from more severe forms. First, MMR has no identified or specified biological etiology, whereas more severe forms of mental retardation often have an identified biological etiology (e.g., Down syndrome, Fragile X syndrome, and microcephaly). Second, MMR is most often diagnosed only at school entry or shortly thereafter, whereas severe forms of mental retardation are often diagnosed at birth or shortly thereafter. Third, adaptive behavior functions of persons with MMR may be adequate in some areas (e.g., practical skills), but severely deficient in others (e.g., conceptual). Individuals with severe forms of mental retardation almost always have pervasive adaptive behavior deficits. Finally, persons with MMR may "blend" into society after school exit (Edgerton, 1993) and appear to function normally in community settings, whereas persons with severe forms of mental retardation will always "stand out" because of their physical anomalies and severely pervasive intellectual and adaptive behavior deficits. It is apparent that the courts have a preconceived notion of what mental retardation looks like that is inconsistent with what MMR looks like to professionals in the field who have training and experience in the field of mental retardation. Unfortunately, this bias is often perpetuated by forensic experts who testify for the prosecution, who, more often than not, have little or no training in the field of mental retardation.

INTERPRETIVE ISSUES IN INTELLECTUAL ASSESSMENT

The remainder of this article will discuss various interpretive issues in intellectual assessment that courts have failed to understand or consider in deciding *Atkins* cases. These interpretive issues are: (a) the nature of intellectual functioning, (b) the Flynn effect, (c) the concept of measurement error, (d) practice effects, and (e) the effect of school diagnoses. Each of these issues will be illustrated with actual *Atkins* cases and court decisions.

Nature of Intellectual Functioning

A major issue confronting the courts in *Atkins* cases resides in their understanding (or misunderstanding) of what intelligence tests measure and how well they

measure the construct of intelligence. The courts have a difficult time comprehending that in a psychometric world; an individual can have more than one true score. For example, suppose an individual is administered a WAIS-III, a Stanford Binet Intelligence Scale-IV (SB-V), and a Woodcock-Johnson Test of Cognitive Abilities-III (WJ Cognitive-III). All three tests yield an overall or composite intelligence score, and an individual taking all three tests will have three true scores, one for each test.

In classical test theory, an individual's true score on any attribute is entirely dependent on the measurement process that is used. In the biological and physical sciences, an individual can have only one true score and that score is independent of the measurement process that is used. This is known as the absolute true score (Crocker & Algina, 1986). For example, a laboratory may analyze an individual's DNA as part of evidence presented in court in a capital case. Individuals have only one true score for their DNA, and the courts have come to understand this phenomenon. However, different labs may obtain different results in their DNA analyses and thus errors of measurement occur. This does not alter the fact that only one true score exists, and different labs would never average the results of various lab tests to derive a true score. Yet, this is precisely how we interpret true scores on psychological measures of intelligence and other attributes.

An *Atkins* case in which I testified brings this interpretive difficulty to light. Darick DeMorris Walker was convicted of two capital murders and sentenced to death in Virginia. Walker claimed that the death penalty violated his Eight Amendment rights that protect him from "cruel and unusual punishment" because he is mentally retarded. Walker had a history of below-average intelligence and a school history of being placed into special education classrooms. Eventually, Walker dropped out of school in the eighth grade with substantial deficits in reading and math skills and a long school history of disruptive/noncompliant behavior.

Throughout his life, Walker has been administered no less than seven intelligence tests, each producing different results. What is particularly notable in these results is the disparity between Walker's crystallized and fluid intelligence. On the various Wechsler tests, Walker's Verbal IQ ranged from 70 to 87 with a median of 78. On various measures of fluid intelligence, his scores ranged from 61 to 68 with a median IQ score of 63. The question before the court was whether or not these scores were indicative of mental retardation. There are two answers to this question which, as expected, confused rather than enlightened the court. If one takes the crystallized measures as being indicative of mental retardation, it is clear that Walker is not mentally retarded. If one takes the fluid measures as indicators

of mental retardation, Walker is, clearly, mildly mentally retarded.

One approach that could be taken would be to argue that different measures of intelligence have different *g loadings*, or that they vary in how well they measure a general intelligence factor. It is well established that measures of crystallized intelligence (vocabulary, verbal abstract reasoning, and general information) have much higher *g loadings* than most measures of fluid intelligence. As such, it could be argued that measures of crystallized intelligence in most circumstances provide better estimates of *g* than most measures of fluid intelligence. This, however, could be disputed on the basis that some measures of fluid intelligence have *g loadings* approaching loadings that are produced by measures of crystallized intelligence (Keith, 2005).

Apart from this argument, the U.S. District Court (Eastern District) ruled against Walker, stating that he failed to show by a preponderance of the evidence that he is mentally retarded. His case was appealed to the U.S. Fourth Circuit Court of Appeals which vacated and remanded the District Court's judgment and granted Walker an evidentiary hearing to determine whether he is mentally retarded under Virginia law. It further ordered that the district court should consider all relevant evidence pertaining to the developmental origin, intellectual functioning, and adaptive behavior aspects of Walker's claim.

Flynn Effect

It is well established that there has been a substantial increase in measured intelligence test performance over time because IQ test norms become obsolete. As such, intelligence test norms have to periodically be recalibrated to maintain their accuracy in reflecting an individual's level of intelligence. The general upward trend in IQ scores has become known as the *Flynn Effect*, named after James Flynn who first documented this phenomenon (Flynn, 1984). Based on his extensive review of the literature, Flynn established that Americans gain approximately 0.3 IQ points per year or 3 points per decade in measured intelligence. Thus, an IQ test normed in 1972 would reflect a 10.8 point gain in IQ today ($36 \times 0.3 = 10.8$ points).

The Flynn Effect has a substantial influence on the number of persons who might be classified as mentally retarded using a specified cutoff score (Ceci, Scullin, & Kanaya, 2003). For example, if you used the WISC-R that was normed in 1972 and specified a cutoff score of 70 and below, you would identify 2.27% of the population as being mentally retarded using the intellectual criterion. However, if you used the WISC-III that was normed in 1989, you would identify 5.48% of the population as being mentally retarded—more than double the

prevalence rate based on a normal distribution. Based on the Flynn Effect it is not unusual for an individual's IQ score to fluctuate above and below a specified IQ cutoff that most states used to determine eligibility for the death penalty (Kanaya, Ceci, & Scullin, 2003).

Flynn (2006) has argued that an individual's true IQ score does not change over time, only the norms change. For instance, suppose you test a girl at age eight with the WISC-R and she obtains an IQ score of 74. You retest that same girl at age 12 with the WISC-III and she obtains an IQ score of 69. There is a five-point difference between these two IQ scores, with one score being above the level for mental retardation and the other score being below that level. The girl's intelligence, however, did not change, only the norms changed, separated by 17 years.

The Flynn Effect differentially affects certain Wechsler scores. For instance, the effect is rather large for Similarities and Block Design and nonexistent for Vocabulary and Information (Flynn, 2006). One could argue that Similarities and Block Design have rather high g loadings (.81 and .70, respectively), therefore this must reflect "real" changes in general intelligence. However, the two subtests that are considered the best single measures of g (Vocabulary and Information) remain unchanged by the Flynn Effect.

In summary, Flynn argues that intelligence has not changed over time, and that changes in measured IQ reflect the fact that norms start becoming obsolete the day they are collected. If this is true, then it could be argued that the Flynn Effect is irrelevant in determining an individual's eligibility for the death penalty because it does not address the level of intelligence, but rather the accuracy of norms that are not a part of any definition of mental retardation. However, states use IQ *scores* which are inextricably and directly dependant on norms for their meaning. These scores often are rigidly adhered to by many states (e.g., Virginia) to determine a person's eligibility for the death penalty. The view that the Flynn Effect does not reflect real changes in intelligence is moot because the courts often use an absolute level of intelligence ($IQ < 70$) to determine whether an individual is eligible for capital punishment.

Measurement Error

It is obvious to any well-trained psychologist that all measurement contains error, but this is far from obvious to the courts in deciding *Atkins* cases. For example, in *Walker v. True* (2006) the United States District Court stated that use of the standard error of measurement (SEM) to lower an IQ score could just as likely be used to raise an IQ score, and that the use of such as statistic is inherently "speculative." Clearly, there is nothing speculative in the standard error of measurement given

that it is entirely dependent on the reliability of the test that is used to obtain a score. The concept of measurement error goes back to the notion of a psychometric true score versus an absolute true score described earlier—a concept that courts have a difficult time understanding. Experts for the defense in *Atkins* cases have been unsuccessful in making courts understand the band of error concept (plus or minus the SEM) and the notion of a psychometric true score that falls within this band of error. Experts for the prosecution have often downplayed the importance of measurement error in these cases because it diminishes the credibility of their testimony (*Walker v. True,* 2006).

An issue relating to measurement error in these cases is the selection of the most appropriate estimate of measurement error: should it be based on internal consistency estimates, stability estimates, or both? Internal consistency estimates will almost always yield higher reliability estimates and thus will produce lower SEMs than stability estimates because stability coefficients are almost always lower.

These two estimates of measurement error reflect two different interpretations of test scores. An internal consistency estimate is based on the average interitem correlation in a test and reflects the ratio of true score variance to total variance (i.e., the reliability index), and the square root of this index is the reliability coefficient (Suen, 1990). As such, this statistic reflects how much error is contained in the obtained score and how well that score estimates the true score. This is known as the *coefficient of internal consistency*. Errors of measurement based on stability estimates reflect the fluctuations in test scores obtained at two points in time.

The problem in classical test theory is that one can have more than one reliability coefficient and thus have more than one standard error of measurement. This is inherently self-contradictory (Suen, 1990) and therefore is more likely to confuse than inform the courts. Conceptually, what is needed is a *coefficient of precision* (Coombs, 1950), which is defined as the correlation between test scores when examinees respond to the same test items (internal consistency) over time (stability) and there are no changes in examinees over time. Unfortunately, this coefficient is a theoretical entity in classical test theory and no completely defensible way of calculating it is possible. Perhaps the best that can be done at this time is to indicate that SEMs based on internal consistency estimates contain an individual's true score at one point in time, whereas SEMs based on stability estimates contain an individual's true score over repeated testings.

Practice Effects

In *Atkins* cases it is likely that defendants have been administered intelligence tests repeatedly; often

beginning in their school years. This was true in *Atkins, Walker v. True*, and *Green v. Johnson*. School records in all of these cases show that these defendants began taking intelligence tests relatively early in their school careers because they were referred to special education. *Walker* had taken seven intelligence tests by the time his case came before the United States Eighth District Court. One argument in *Walker v. True* made by the defense was that his IQ scores should be adjusted downward, in part, because of well-known *practice effects* due to repeated administrations of the same test. The Court ruled, however, in *Walker v. True* that "Petitioner has failed to present evidence that such an adjustment would be anything other than speculation" (p. 8).

Practice effects refer to gains in test scores on intelligence tests that occur when an individual is retested on the same or similar instruments. This is not a speculation but rather a well-established empirical fact. These gains are due to having been exposed to the same or very similar test items and not due to any specific performance feedback given by examiners. Practice effects for the various Wechsler scales from ages 5 to 50 years show median gains in Verbal IQ of 3 points, Performance IQ of 9 points, and Full Scale IQ of 7 points (Kaufman, 2003). Walker had taken the WISC-R three times before the age of 18 and the WAIS-III twice after the age of 18. Thus, the practice effects on Wechsler scales beginning at age 9 to 20 (his last WAIS-III) must have been quite substantial, thereby producing inflated IQ scores.

In *Atkins* cases the courts must be made to understand the average practice effect gains in IQ scores and how these artificially inflated test scores produce an overestimate of an individual's true score. This is particularly true when experts from either side administer the same test within relatively short periods of time, because the shorter the retest interval, the larger the practice effect. If we apply the median practice effect to Walker's median Full Scale IQ, his IQ goes from 76 to 69; not considering measurement error. Quite clearly, this is extremely important in *Atkins* cases, particularly in states that inflexibly adhere to IQ < 70 standard for mental retardation.

SCHOOL DIAGNOSES

A major source of evidence used by courts in *Atkins* cases is the documentation of whether or not the defendant had ever been identified by a school as mentally retarded. This is considered an essential piece of evidence, given that one of the eligibility prongs for a diagnosis of mental retardation is onset prior to age 18. In *Walker v. Virginia*, the defendant received special education services during elementary school, first under the

label of "learning disability" and later under the label of "emotionally disturbed." Walker was never labeled as being mentally retarded by the Richmond, Virginia public schools, despite evidencing significantly subaverage intellectual functioning and deficits in conceptual, social, and practical adaptive skills. A similar educational history was evidenced in the *Atkins* and *Green v. Virginia* cases.

The fact that none of these individuals had received the label of mental retardation by the public schools in not unusual, particularly for African Americans for whom the issue of overrepresentation in special education programs for the mentally retarded has been an issue since the late 1970s. A study by MacMillan and colleagues showed how this mislabeling" occured in a series of studies conducted in California (Gresham, MacMillan, & Bocian, 1998; MacMillan, Gresham, Bocian, & Siperstein, 1997; MacMillan, Gresham, Siperstein, & Bocian, 1996).

In one study, MacMillan, Gresham, Siperstein, and Bocian (1996) selected a sample of 43 students from grades two, three, and four who had WISC-III IQ scores of 75 and below. The schools that these students attended classified 44% of these students as "learning disabled" (19 students) despite the group having a mean IQ of 68. Only 14% (six students) were classified as mentally retarded with a mean IQ of 63. The remaining 18 students received no formal classification by schools and remained in general education. Similar results were reported by Kanaya, Ceci, and Scullin (2003), who showed that 48.1% of children with IQs below 70 were classified as learning disabled ($M = 66$) and 48.5% were classified as mentally retarded ($M = 64$).

Clearly, relying on a school history of being classified as mentally retarded and receiving special education services under that label is not very reliable in establishing the onset of mental retardation prior to age 18. Courts should be presented with evidence such as that cited by MacMillan, Gresham, Siperstein, and Bocian (1996), MacMillan, Gresham, Bocian, and Siperstein (1997), and Kanaya, Ceci, and Scullin (2003) to demonstrate that the use of the mentally retarded label, especially for individuals with mild mental retardation, is uncommon and is often replaced with a label of learning disabled. Unfortunately, courts often take the failure of schools to diagnose defendants as mentally retarded to be proof that they are not mentally retarded.

CONCLUSIONS

It is clear that experts in *Atkins* cases have provided the court with varying opinions regarding the presence or absence of mental retardation. This was made particularly clear in the *Atkins* trial when one expert diagnosed

Atkins as mentally retarded with an IQ of 59 and the other expert indicated that he had "normal intelligence." An issue that continues to confuse the courts is the nature of *mild mental retardation* (MMR) as distinguished from more severe forms of mental retardation. It is likely that the courts have a preconceived notion of mental retardation that frequently does not include the construct of MMR. Courts are often not convinced that mental retardation, particularly MMR, is a *relative* concept and that an individual's limitations have meaning only in terms of social conditions (Edgerton, 1993). Limitations in intellectual and adaptive behavior functioning must be interpreted within the context of a person's age, culture, and peers and are not absolute concepts. Courts, on the other hand, seek to discover "absolute truths" and are often confounded by arguments that introduce relative concepts into a legal defense.

Differences in expert opinion may stem from a lack of understanding by experts of the concept of mental retardation. Most experts for the prosecution in *Atkins* cases have little or no training or experience in the field of mental retardation (Greenspan, 2006). This was the case in *Walker v. True* in which the expert had a long history of testifying in forensic cases, but no formal training whatsoever in the field of mental retardation.

Another issue that often confuses the courts is the nature of measurement error and how it can affect the interpretation of test scores. This is a nonissue with more severe forms of mental retardation, but a key issue with MMR. If an individual obtains an IQ of 75 and a state uses an IQ below 70 for its intellectual criterion for mental retardation, the prosecution almost always argues that the person cannot be mentally retarded. This argument, however, ignores the fact that there is measurement error in all test scores. For most IQ test scores, the accepted degree of measurement error is five points meaning that an IQ of 75 could be between 70 and 80. More confusing is the fact that measurement error can come from different sources such as internal consistency and stability reliability estimates. In *Walker v. True*, the court considered the concept of measurement error to be "speculative," and defense experts were unsuccessful in arguing against this inaccurate notion.

A controversial issue in *Atkins* cases is the Flynn effect that shows the mean IQ of Americans increases over time by about 0.3 points per year and 3 points per decade (Flynn, 1984). The Flynn Effect can produce a substantial increase in the number of persons diagnosed with MMR, depending on the date a test was normed. For instance, if one used the WISC-III that was normed in 1989 and specified a cutoff of 70 and below, about 2.27% of the population would be identified as mentally retarded. On the other hand, if one used the WISC-IV that was normed in 2001, one would identify approximately 4% of the population as being mentally retarded. The concepts to be understood in interpreting the Flynn Effect are twofold: (1) the *mean* IQ increases over time (the mean shifts upward), and (2) intelligence does not change, only the norms change (i.e., they get "tougher").

Finally, it has been difficult for defendants in *Atkins* cases to meet the developmental criterion in a diagnosis of mental retardation. It must be shown in these cases that an individual's mental retardation had an onset prior to age 18. In many, if not most, *Atkins* cases, this has proven difficult because all defendants have been adults with no prior diagnosis of mental retardation. Consulting defendants' school records frequently show that many of these individuals have a long history of poor academic performance, retention in grade, and a history of special education. In *Atkins, Walker v. True,* and *Green v. Virginia*, all defendants had a history of school difficulties and/or special education, but none were ever diagnosed as being mentally retarded by schools. Instead, Walker was diagnosed as "learning disabled" and "emotionally disturbed" by Richmond, Virginia schools, and Green was diagnosed as "speech and language impaired" and "learning disabled" by the Washington, DC public schools.

It is well-established that schools were and are reluctant to classify children as mentally retarded, particularly African-American students since the 1970s (MacMillan & Siperstein, 2002). Schools frequently assign a more "palatable" label to students who would otherwise be classified as mentally retarded, using labels such as "specific learning disability" or "speech and language impairment." In *Atkins* cases, this frequently works against the defense's efforts because there is no developmental history of an individual ever being diagnosed as mentally retarded, thereby making it difficult to prove the developmental criterion of mental retardation.

Experts testifying for the defense in *Atkins* cases should be well prepared to testify about the nature of mild mental retardation, to be extremely knowledgeable of psychometric theory and measurement error, to understand and be able to articulate the Flynn effect, and to testify about the failure of schools to diagnose mental retardation and their tendency to use "softer" labels for students who may have been mentally retarded. Ultimately, the courts will be the final arbiter of the convincingness of this testimony.

REFERENCES

American Association on Intellectual and Developmental Disabilities. (2002). *Mental retardation: Definition, classification, and systems of supports* (10th ed.). Washington, DC: Author (2002).

American Psychiatric Association. (2000). *Diagnostic and statistical manual for mental disorders* (Text revision). Washington, DC: Author.

Atkins v. Virginia. (2002). 536, U.S. 304, 122, S. CT 2242.

Coombs, C. H. (1950). The concepts of reliability and homogeneity. *Educational and Psychological Measurement*, *10*, 43–56.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart, & Winston.

Edgerton, R. B. (2003). *The cloak of competence*. Los Angeles: University of California Press.

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin*, *95*, 29–51.

Flynn, J. R. (2006). Tethering the elephant: Capital cases, IQ, and the Flynn effect. *Psychology, Public Policy, and Law*, *12*, 170–189.

Green v. Johnson. United States District Court, Eastern District of Virginia, No. 2:05 cv 340. (Testimony in October 2006) Death penalty appeal.

Greenspan, S. (2006). Adaptive behavior in *Atkins* proceedings. Paper presented at a symposium at the Annual Meeting of the American Psychological Association. San Francisco, August 2006.

Gresham, F. M., MacMillan, D., & Bocian, K. (1998). Agreement between school study team decisions and authoritative definitions in classification of students at-risk for mild disabilities. *School Psychology Quarterly*, *13*, 181–191.

Kanaya, T., Ceci, S., & Scullin, M. (2003). The difficulty of basing death penalty eligibility on IQ cutoff scores for mental retardation. *Ethics & Behavior*, *13*, 11–17.

Kaufman, A. (2003). *Practice effects: The clinical café*. Bloomington, MN: Pearson Assessments.

Keith, T. (2005). Using confirmatory factor analysis to aid in understanding the constructs measured by intelligence tests. In D. Flanagan & P. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 581–614). New York: Guilford Press.

MacMillan, D., & Siperstein, G. (2002). Learning disabilities as operationally defined by schools. In R. Bradley, L. Danielson, & D. Hallahan (Eds.), *Learning disabilities: Research to practice* (pp. 287–333). Mahwah, NJ: Erlbuam.

MacMillan, D., Gresham, F. M., Bocian, K., & Siperstein, G. (1997). The role of assessment in qualifying students as eligible for special education: What is and what's supposed to be. *Focus on Exceptional Children*, *30*, 1–18.

MacMillan, D., Gresham, F. M., Siperstein, G., & Bocian, K. (1996). The labyrinth of IDEA: School decisions on referred students with subaverage general intelligence. *American Journal on Mental Retardation*, *101*, 161–174.

Suen, H. K. (1990). *Principles of test theories*. Hillsdale, NJ: Erlbaum.

Walker v. True (2006). 399, F. 3d, 327 (4th Cir. 2005).

Wechsler, D. (1972). *Wechsler Memory Scale Manual*. San Antonio: Psychological Corporation.