Standard of Practice and Flynn Effect Testimony in Death Penalty Cases

Frank M. Gresham and Daniel J. Reschly

Abstract

The Flynn Effect is a well-established psychometric fact documenting substantial increases in measured intelligence test performance over time. Flynn's (1984) review of the literature established that Americans gain approximately 0.3 points per year or 3 points per decade in measured intelligence. The accurate assessment and interpretation of intellectual functioning becomes critical in death penalty cases that seek to determine whether an individual meets the criteria for intellectual disability and thereby is ineligible for execution under *Atkins v. Virginia* (2002). We reviewed the literature on the Flynn Effect and demonstrated how failure to adjust intelligence test scores based on this phenomenon invalidates test scores and may be in violation of the *Standards for Educational and Psychological Testing* as well as the "Ethical Principles for Psychologists and Code of Conduct." Application of the Flynn Effect and score adjustments for obsolete norms clearly is supported by science and should be implemented by practicing psychologists.

DOI: 10.1352/1934-9556-49.3.131

The Flynn Effect is a well-established psychometric fact documenting substantial increases in measured intelligence test performance over time. These increases are not generally believed to reflect actual gains in the construct of intelligence but, rather, the creeping obsolesce of test norms (see Flynn, 1984, 1987). Flynn's (1984) seminal review of the literature established that Americans gain an average of approximately 0.3 IQ points per year or 3 points per decade in measured intelligence. His subsequent paper published in 1987 showed a similar increase in measured intelligence worldwide (Flynn, 1987). An intelligence test normed in 1977 and used today has a population mean of approximately 110 (0.3×33 years $= 9.9$). A score of 75 today using the obsolete norms from 1977 is 2.33 *SD* below the population mean and is comparable to a score of 65 if the actual population mean was 100 with an *SD* of 15. The critical issue for psychologists is which score reflects most accurately the individual's current status compared to the overall population.

Our purpose in this article is to provide a discussion of the Flynn Effect and describe how failure to consider it in death penalty cases can have life or death consequences for individuals with intellectual disability. First, we provide an overview of intellectual disability and discuss how so-called *Atkins* cases have exclusively involved individuals having mild intellectual disability rather than more severe forms. We provide a brief overview of relevant aspects of measurement theory and tie this to the legal implications of the Flynn Effect in death penalty cases. We present three actual *Atkins* cases and show how the failure to consider the Flynn Effect, in part, lead to executions in two of the three cases. We conclude the article with a discussion of standards of practice and validity considerations in employing the Flynn Effect in capital cases involving individuals with intellectual disability.

Although widely accepted by scholars, measurement experts, and researchers in the area of intellectual measurement, why, then, is the Flynn Effect important for the everyday practice of clinical assessment? In other words, what practical difference would it make to clinical practitioners

that the population mean changes systematically with the degree of obsolescence of test norms? Moreover, because the scores on tests of intellectual functioning only become meaningful through comparisons to population means, how can clinicians ensure that these comparisons are statistically accurate? Failure to consider changes in measured phenomena or construct over time often can have dire consequences for individuals, and to not account for these changes is to deny this reality.

The accurate assessment of intellectual functioning becomes critical in death penalty cases when determining whether an individual meets the criteria for intellectual disability, in Social Security Administration disability determinations (Reschly, Meyers, & Hartel, 2002), and in eligibility for special education placement and services (MacMillan, Gresham, Siperstein, & Bocian, 1996). In these cases, the use of obsolete norms without appropriate corrections or considerations has enormous consequences for the individual (Flynn, 2010; Flynn & Widaman, 2008). As pointed out by Hagan, Drogin, and Guilmette (2008), psychologists assist in thousands of legal determinations in which the accurate assessment of intellectual functioning is a central issue.

In 2002, the Supreme Court in *Atkins v. Virginia* ruled that it was a violation of the U.S. Constitution Eighth Amendment's prohibition against cruel and unusual punishment to execute individuals with mental retardation. During the Atkins trial, two board certified forensic psychologists came to diametrically opposed opinions concerning whether or not the defendant Daryl Atkins had intellectual disability. One psychologist who evaluated Atkins concluded that he had intellectual disability, with a tested Full-Scale IQ (FSIQ) of 59 on the Wechsler Adult Intelligence Scale-III (WAIS-III). Another forensic psychologist testified that Atkins was functioning in the range of average intelligence. How is it possible that two board certified forensic psychologists can come to vastly different opinions concerning the presence or absence of intellectual disability? As will be illustrated throughout this article, this is neither unexpected nor unusual.

Intellectual Disability

Three prongs have guided the diagnoses of intellectual disability for 70 years (Doll, 1934, 1941): intellectual functioning, adaptive behavior (social competence), and developmental origin. Although classification criteria and terminology differ slightly, *intellectual disability* has been defined by virtually all organizations and states as significantly subaverage intellectual functioning that exists concurrently with deficits in adaptive behavior and which has an onset prior to age 18 years. Most states adopt diagnostic criteria that follow the definition contained in either the *Diagnostic and Statistical Manual* (*DSM*)*-TR* (American Psychiatric Association, 2000) or the definition specified by the American Association on Intellectual and Developmental Disabilities—AAIDD (Schalock et al., 2010). Greenspan (2009) has noted that the three criteria specified in the *DSM* and AAIDD manuals have remained conceptually unchanged over nearly 5 decades.

Classification Criteria

What has changed, however, are the operational standards for diagnosing an individual as having intellectual disability based on the criteria of intellectual functioning and adaptive behavior. For example, in the 1961 definition of intellectual disability specified by the American Association on Mental Deficiency—AAMD, Heber (1961) used an intellectual functioning criterion of 85 and below as being indicative of intellectual disability. Twelve years later, the AAMD lowered the intellectual functioning criterion to 70 and below, effectively eliminating 14% of all cases of intellectual disability based on the intellectual functioning criterion (Grossman, 1973).

It is important that both AAIDD and the American Psychiatric Association recognize that measurement error of approximately 5 points is contained in all standardized tests of intelligence and should be taken into account in diagnosing intellectual disability. As such, it is possible to diagnose an individual with intellectual disability who has an IQ up to 75 if they also have significant limitations in adaptive behavior and an onset prior to age 18. One should also realize that there are over twice as many potential cases of intellectual disability with IQs between 70–75 (.0475) than with IQs below 70 (.0222) (Reschly et al., 2002).

The debate in *Atkins* cases has never been about individuals with more severe levels of intellectual disability. It has always been about persons who may be considered to have mild intellectual disability. In the AAIDD *Manual*,

Schalock et al. (2010) defined intellectual disability in much the same way as it was defined in the *DSM-TR* with two exceptions: (a) AAIDD does not specify levels of severity and (b) AAIDD specifies a numerical cutoff score for limitations in adaptive behavior (i.e., greater than 2 *SD*s below the mean) in conceptual, practical, or social adaptive skills.

Types of Intellectual Disability

A crucial issue in *Atkins* cases that is often either misunderstood by the courts or at least is not made clear by defense attorneys is the nature of mild intellectual disability as being distinct from more severe forms. First, mild intellectual disability has no identified or specified biological etiology, whereas more severe forms of intellectual disability often have an identified biological etiology (e.g., Down syndrome, fragile X syndrome, Tay Sachs). Second, mild intellectual disability is most often diagnosed only at school entry or shortly thereafter, whereas severe forms of intellectual disability are often diagnosed at birth or shortly thereafter. Third, some genuine cases of mild intellectual disability are not diagnosed by schools or are misdiagnosed as learning disability (MacMillan et al., 1996). Fourth, adaptive behavior functioning of persons with mild intellectual disability may be adequate in some areas (e.g., practical skills) and severely deficient in others (e.g., conceptual skills). Individuals with severe mental retardation almost always have pervasive deficits in adaptive behavioral functioning. Finally, persons with mild intellectual disability may "blend" into society after school exit (Edgerton, 1993) in that many are not officially diagnosed with intellectual disability in the adult years because they appear to function typically in community settings, whereas persons with severe forms of mental retardation will always "stand out" because of their physical anomalies and severe pervasive intellectual and adaptive behavior deficits. Persons with mild intellectual disability continue, however, to exhibit significant limitations in reasoning and judgment, and the seemingly "normal" performance usually depends on significant assistance from a benefactor (Edgerton, Ballinger, & Herr, 1984).

Many courts may have a preconceived notion of what intellectual disability looks like that is inconsistent with what mild intellectual disability looks like to professionals with training and experience in the field of intellectual disability. Unfortunately, these preconceived notions are often perpetuated by forensic experts who testify for the prosecution and who, more often than not, have little or no training in the field of intellectual disability (Olley, 2009).

Measurement Theory and Intellectual Assessment

A major challenge for any expert witness in *Atkins* cases is to explain to courts the nuances of intellectual assessment and interpretation in understandable terms. Many times, judges, opposing attorneys, and juries have a difficult time understanding how intelligence tests are constructed, what they measure, and how they should be interpreted (Flynn, 2009). For example, in *Atkins* cases, it is important for the court to understand that in a psychometric world, an individual can have more than one true score for his or her level of intellectual functioning. This is particularly true in *Atkins* cases, where defendants often have taken different versions of the same test over time (e.g., the Wechsler scales) and/or different intelligence tests (e.g., Stanford Binet, Woodcock-Johnson, Differential Ability Scales). In many of these cases, an *Atkins* defendant may show higher scores on some intelligence tests and lower scores on others. This is not unusual and can be due to a host of factors, such as different norming periods, different test content, presence or absence of practice effects, and the degree to which the test measures different facets of intelligence (Gresham, 2009).

In classical test theory, an individual's true score on any attribute is entirely dependent on the measurement process that is used (Crocker & Algina, 1986). This is not the case in the biological and physical sciences, in which an individual can have only one true score and that score is independent of the measurement process used. This is known as the *absolute true score*. A relevant example in forensics science is the analysis of a defendant's DNA. Individuals can have only one true score for their DNA, and the courts have come to understand this phenomenon. It is true that different labs may sometimes obtain different results and errors of measurement can occur. This does not alter the fact that only one true score exists for an individual's DNA, and different labs would never average the results of various DNA lab tests to derive a "true DNA score." Yet, this is precisely how we

interpret true scores on psychological measures of intelligence and other attributes.

In classical test theory, an individual can have many true scores for his or her intelligence depending on the number of different intelligence tests administered over his or her lifetime. This logic has been well accepted in the psychometric literature for over 100 years (Spearman, 1904). An *Atkins* case in which we testified brings this interpretative difficulty to light (see *Walker v. True*, 2006). Darick DeMorris Walker was convicted of two capital murders and sentenced to death in Virginia. Walker claimed that the death penalty violated his Eighth Amendment rights to protect him from cruel and unusual punishment because he is mentally retarded. Walker had a history of below-average intellectual functioning and a school history of special education placement. Eventually, Walker dropped out of school in the eighth grade; he had substantial deficits in reading and math skills and a long school history of disruptive and noncompliant behavior.

Seven intelligence tests had been administered to Walker throughout his lifetime, with each test producing somewhat different results. On the various Wechsler tests, Walker's Verbal IQ (VIQ) ranged from 70 to 87, with a median of 78. On the Performance IQ (PIQ) measures, Walker's scores ranged from 61 to 68, with a median of 63. The question before the court in this case was whether or not these scores were indicative of mental retardation. If one takes the VIQ measures at face value, then it is clear that Walker did not meet the Virginia standard for mental retardation. On the other hand, if one takes the various PIQ measures at face value, then it is clear that Walker did meet the Virginia standard for mental retardation. Dilemmas such as these are not uncommon in *Atkins* cases across the country (Greenspan & Switzky, 2006).

In any event, the U.S. District Court (Eastern District of Virginia) ruled against Walker, stating that he failed to show by a preponderance of the evidence that he had intellectual disability. 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 had intellectual disability under Virginia law. It further ordered that the District Court should consider all relevant evidence pertaining to Walker's developmental origin, intellectual functioning, and adaptive behavior. The District Court conducted this evidentiary hearing and again reached the conclusion that Walker did not have intellectual disability. Darick Walker was executed by lethal injection at Greensville Correctional Center in Virginia on May 20, 2010.

Legal Implications of the Flynn Effect

There is no doubt that the Flynn Effect can have substantial legal implications in *Atkins* cases in which the presence of intellectual disability for an individual is being contested. As mentioned earlier, in all of these cases, the issue focuses on the category of mild intellectual disability, not more severe cases. Flynn (2006) used the example of a boy who was tested twice during his school years. In 1973, he scored 75 on the WISC that was normed in 1947–1948; thus, the norms were 25.5 years out of date. In 1975, the boy was tested at age 8 with the WISC-R, which was normed in 1972, and, therefore, with norms only 3 years out of date. He obtained an IQ of 68. The score at age 6 of 75 and at age 8 of 68 are, in fact, statistically the same score based on the Flynn Effect because the 1973 score was inflated by 7 points and the 1975 score was not influenced by the Flynn Effect because of the recency of the WISC-R norms.

How is this example relevant to present day *Atkins* cases? Suppose two defendants were tested in 2004 to provide evidence that would be presented in *Atkins* cases. The first defendant was tested with the WAIS-III that was normed in 1989 and obtained an IQ of 73. The second defendant was tested with the WAIS-IV that was normed in 2002 and obtained a score of 69. The first defendant was convicted and sentenced to death because his score did not meet the "bright line" of IQ 70 or below, whereas the second defendant was not sentenced to death because his IQ of 69 met the state's bright line of IQ less than 70. The fact is that both of these scores for the two defendants are statistically identical when viewed in light of the Flynn Effect.

This is precisely what happened in a recent Florida *Atkins* case (Cherry v. State, 2007). Roger Cherry was convicted of capital murder and sentenced to death. On a postconviction appeal, Cherry claimed he had intellectual disability and, therefore, was ineligible for the death penalty. His tested WAIS-III score of 72 did not meet the Florida bright line criterion of IQ 70 and below, and the court denied Cherry's appeal. In fact, when

Cherry took the WAIS-III, the norms were 13 years out of date, thereby producing a Flynn Effect of approximately 4 points. Based on the Flynn Effect, Cherry's IQ of 72 is actually 68, thereby meeting the Florida bright line standard. As Flynn (2006) indicated: "Failure to adjust IQ scores in light of IQ gains over time turns eligibility for execution into a lottery" (pp. 174–175).

Some of the illustrations above might be criticized because they are hypothetical; however, we next present three actual *Atkins* cases that show the real legal ramifications of the Flynn Effect in death penalty cases. The first case presented in Table 1 is Darick Walker (previously mentioned), who was convicted of two capital murders (*Walker v. True, 2006*) and executed on May 20, 2010. Recall that the U.S. District Court ruled twice that Walker did not have intellectual disability and upheld his death penalty sentence. Table 1 shows that Walker's Wechsler IQs for VIQ, PIQ, and FSIQ were 70, 85, and 76, respectively. When Flynn corrections were applied, these scores more accurately were 66, 81, and 72, respectively, and clearly placed Walker in the range of mild intellectual disability based on *DSM-TR* and AAIDD intellectual criteria.

The second case presented in Table 1 is Kevin Green, who was convicted of capital murder, denied a status of mental retardation in an appeal of the death penalty (*Green v. Johnson, 2006*), sentenced to death, and executed on May 27, 2008. Green's IQs were 67, 80, and 71 for VIQ, PIQ, and FSIQ, respectively. In 1991, while a 14-year-old student in fourth grade (having failed three school grades previously and described by his teacher as fitting in well socially with children 4 to 5 years younger), Green was referred for a psychological evaluation as part of the consideration of special education eligibility. The 1974 version of the Wechsler Scale (WISC-R) was used, despite the publication of the updated WISC-III in 1991. The FSIQ of 71 was derived from a test with norms that were 19 years obsolete. The WISC-R population mean in 1991 was approximately 106. The score of 71 on the WISC-R in 1991 was 2.33 *SD*s below the population mean, clearly exceeding the traditional standard of intellectual functioning approximately 2 *SD* below the population mean. However, the Flynn corrections show that Green's scores in comparison to the existing population mean were 61, 74, and 65, respectively, clearly placing him in the range of mild intellectual disability based on the intellectual criterion. Nevertheless, a board certified forensic psychologist urged the court to ignore the Flynn Effect because it did not represent the current standard of practice in psychology (see later discussion).

Table 1 Uncorrected and Flynn Corrected Wechsler Scores for Three *Atkins* Cases

Score[a]	Walker[b]	Green[c]	Johnston[d]
VIQ	70	67	69
FVIQ	66	61	63
PIQ	85	80	89
FPIQ	81	74	83
FSIQ	76	71	76
FFSIQ	72	65	71

[a]VIQ = Verbal IQ, FVIQ = Flynn Corrected VIQ, PIQ = Performance IQ, FPIQ–Flynn Corrected PIQ, FSIQ=Full Scale IQ, FFSIQ–Flynn Corrected FSIQ. [b]Based on WAIS-III normed in 1989 and administered in 2004. [c]Based on WISC-R normed in 1972 and administered in 1991. [d]Based on WAIS-III normed in 1989 and administered in 2005.

Finally, Table 1 shows the Wechsler IQs for David Johnston, who was convicted of capital murder in Florida (see *Johnston v. State*, 1986) and sentenced to death. Table 1 shows that Johnston's IQs were 69, 89, and 76 for VIQ, PIQ, and FSIQ, respectively. Flynn corrections lower these scores to 63, 83, and 70, respectively, again placing Johnston in the range of mild intellectual disability based on the intellectual criterion.

All three of the above cases consistently show how failure to account for the Flynn Effect can produce IQs that move defendants out of the range of intellectual disability on the Wechsler scales. In 2 of the 3 cases (Walker and Green), this failure contributed to their execution in the state of Virginia. The third case (Johnston) was before the Florida Supreme Court; however, Johnston died of natural causes on Death Row before the Supreme Court could rule on his case.

Some have questioned whether or not the Flynn Effect applies reliably to specific individuals, particularly those who find themselves in *Atkins* cases and death penalty appeals (Hagan et al., 2008). This is, frankly, a specious argument simply because any individual's IQ is entirely dependent upon group mean scores of the standardization sample. If the group mean has shifted upward, then the score that meets the intellectual disability

standard has likewise increased by the same amount (Flynn, 1985). If this standardization sample is obsolete, then any individual score calculated in reference to the obsolete norms will be inflated by a factor of 0.3 points per year, or 3 points per decade from when the test was standardized.

The Flynn Effect has a substantial influence on the number of persons who might be classified as having intellectual disability using a specified cutoff score based on a large scale of the proportions of persons identified as having intellectual disability and placed in special education programs. For example, Kanaya, Ceci, and Scullin (2003) found that the number of children who were diagnosed with intellectual disability nearly tripled with the introduction of the WISC-III (from the WISC-R) because more and more children obtained an IQ of 70 and below with the comparison to the more difficult norm. The Flynn Effect produces situations in which a given individual's IQ can fluctuate above and below a specified IQ cutoff that most states use to determine eligibility for the death penalty (Flynn, 2009; Kanaya et al., 2003). In effect, this is like playing dice with IQ scores, except the stakes in *Atkins* cases are most certainly higher.

Two recent court cases in capital trials applied the Flynn Effect as well as acknowledging the standard error of measurement and an intellectual disability cutoff score at 75 to evidence similar to that in the *Walker* and *Green* cases, leading to decisions forbidding the death penalty (*U.S. v. Hardy*, 2010; *U.S. v. Lewis*, 2010). It is significant that these cases were trials in federal district courts, where the judges are appointed for life, rather than in state courts, where judges often are elected and more responsive to public opinion, which frequently favors strong retribution against capital defendants. In both of the recent cases, the Flynn Effect was accepted as a scientific fact, and testimony that the Flynn Effect is not currently taught in graduate programs preparing psychologists was essentially discounted. We can only speculate on whether state courts will increasingly adopt what we see as clear scientific evidence cases confirming the Flynn Effect.

We acknowledge that acceptance of the Flynn Effect will not always yield decisions forbidding the death penalty. In fact, in both *Green* and *Walker*, the appellants were also found ineligible for the intellectual disability classification on the adaptive behavior criterion. It is our impression, however, that courts, much like practitioners making diagnoses of intellectual disability in school settings, are strongly influenced by the individual's status on the general intellectual functioning prong, with decisions about adaptive behavior following rather than being equally weighted with intelligence in intellectual disability decisions (Reschly & Ward, 1991). Greater weighting of the intellectual prong also occurs because of less well-developed measures of adaptive behavior and difficulties with gathering adaptive behavior information for adults prior to age 18 (Reschly, 2009).

Standard of Practice and the Flynn Effect

What, then, are practicing psychologists to do when presented with an *Atkins* case, and they find themselves as expert witnesses in courts or in SSI disability evaluations involving intellectual disability? In other words, what is the appropriate standard of practice for interpreting IQs in light of the Flynn Effect? Opinions regarding this issue understandably vary depending on who is asked that question. Greenspan (2006) suggested that adjusting an individual's IQ in light of the Flynn Effect is essential. Others have made similar suggestions based on their analysis of the Flynn Effect in various reviews of the literature (Ceci & Kanaya, 2010; Fletcher, Stuebing, & Hughes, 2010; Kanaya et al., 2003; McGrew, 2010).

Hagan et al. (2008) addressed this issue by conducting a survey of 358 APA-approved clinical, counseling, and school psychology program directors. One surprising result was the fact that over one third (36%) of program directors had either not heard of the Flynn Effect or were slightly familiar with the concept. Of the remaining 64% of the respondents, almost 92% of them indicated they would never teach students to recalculate IQs based on the Flynn Effect. Similarly, a survey of 28 Diplomates in School Psychology revealed that 94% of them had never adjusted IQs based on the Flynn Effect.

Survey results depend heavily on how questions are worded and the use of context descriptions. Apparently, Hagan et al. (2008) simply inquired about subtracting points based on the Flynn Effect without any description of context or implications. Under these circumstances the clear majority of the small proportions of each sample who responded rejected score adjustments. These results likely would have been different if the respondents were given SSI or death penalty contexts, such as those described above in the Walker, Green, and Johnston cases.

Hagan et al. (2008) also reported that primary source assessment texts and test manuals did not recommend changing scores. Again, however, context and vested interests likely make a difference. Moreover, test publishers have a vested interest in ignoring the Flynn Effect in test manuals because of the tacit admission attendant to discussing this phenomenon that tests have a limited shelf life and need to be updated frequently (Kaufman, 2010; Weiss, 2007, 2010). One exception is the following content from the WAIS-III *Manual* (Wechsler, 1997).

> **Updating of Norms**. Because there is a real phenomenon of IQ-score inflation over time, norms for a test of intellectual functioning should be updated regularly (Flynn 1984, 1987; Matarazzo, 1972). Data suggest that an examinee's IQ score will generally be higher when outdated rather than current norms are used. The inflation rate of IQ scores is about 0.3 points each year. Therefore, if the mean IQ of the U.S. population on the WAIS-R was 100 in 1981, the inflation might cause it to be about 105 in 1997. (pp. 8–9)

Not surprisingly, the most recent WAIS version does not discuss the Flynn Effect (Wechsler, 2008), perhaps reflecting the rather defensive denial of Flynn's criticism of the WAIS-III standardization sample by a test company official involved with the development of the Wechsler scales (Weiss, 2007). To set the record straight, the Flynn Effect continues to be prominent and well supported statistically through the most recent revisions of the Wechsler scales (Flynn, 2009).

Hagan et al. (2008) concluded that adjusting IQ scores and recalculating scores based on the Flynn Effect do not represent custom or standard of practice in professional psychology based on a survey with a participation rate among those surveyed. This so-called standard of practice, however, was based on a survey in which over one third of the sample responding was fundamentally unfamiliar with the concept at issue—namely, the Flynn Effect. The majority of the remaining respondents said they would never teach students to adjust scores based on the Flynn Effect. This finding is not scientifically convincing and should not be taken at face value. The Flynn Effect is a well-established measurement phenomenon based on years of replicated research findings across the world. The fact that most program directors would never teach students to interpret scores in light of the Flynn Effect is to ignore scientific reality and potentially could be in violation of the *Standards for Educational and Psychological Testing* (American Educational Research Association, 1999).

Perhaps the most well-known and qualified group of professionals who deal with the diagnosis and treatment of persons with intellectual disability are members of the AAIDD. Founded in 1876, this organization has, through 11 editions of its diagnostic manual, provided guidance for professionals working in the field of intellectual disability. Reschly (1992) established that the AAIDD leads the world, including the *DSM*, in the development and refinement of the intellectual disability diagnostic construct. In the User's Guide of the 10th edition of the AAIDD Manual, Schalock et al. (2006) stated that best practices require recognition of the Flynn Effect when older editions of an intelligence test are used in assessment or interpretation of an IQ score. The authors go further:

> The main recommendation resulting from this work [regarding the Flynn Effect] is that all intellectual assessment must use a reliable and appropriate individually administered intelligence test. In cases with multiple versions, the most recent version with the most current norms should be used at all times. *In cases where a test with aging norms is used, a correction for the age of the norms is warranted* [italics added]. (pp. 20, 21)

Validity Considerations

Validity is the centerpiece concept in every aspect of psychological assessment. Validity is an evaluative judgment of the extent to which empirical evidence and theoretical explanations support the adequacy and appropriateness of test score interpretations and actions (Messick, 1995). We emphasize that validity is not a characteristic of a given test, but rather is a property of the meaning of test scores. Cronbach (1971) argued that what is validated in psychological testing is the meaning and interpretation of the test score and the implications for actions that the meaning entails.

Based on this conceptualization of validity, what impact does the Flynn Effect have on the meaning and interpretation of intelligence test scores? The most obvious implication is that failure to account for the Flynn Effect in the interpretation of such scores renders that interpretation inaccurate. For example, interpretation of a WAIS-III score of 72 administered in 2006 and deciding that the individual does not meet the criterion of IQ 70 or less would be erroneous. A Flynn correction of this score, in fact, would yield a more accurate score of 69, thereby meeting the IQ criterion. It is unknown how prevalent these validity violations are in *Atkins* cases, but we believe this to be a

common phenomenon, particularly based on the Hagan et al. (2008) survey of clinical, counseling, and school psychology program directors.

The *Standards for Educational and Psychological Testing* (American Educational Research Association, 1999) indicate that proper interpretations of test scores may be compromised by *construct-irrelevant variance*, which is defined as the degree to which test scores are affected by processes that are extraneous to the construct being measured. We argue that the failure to adjust IQ scores based on the Flynn Effect introduces construct-irrelevant variance into the proper interpretation of intelligence test scores. Failure to make this adjustment diminishes the quality and accuracy of test score interpretation and invalidates the inferences that can be made from those test scores.

Messick (1995) discussed the issue of consequential validity in his seminal paper on validity of psychological assessment. Using the language of Cronbach and Meehl (1955), Messick suggested that unintended consequences occurring in psychological testing are strands in the nomological network that should be taken into account in test score interpretation and use. We maintain that failure to account for the Flynn Effect in death penalty cases can produce adverse social consequences for individuals and, thus, invalidate their test scores. Messick (1995) suggested that:

> The primary measurement concern with respect to adverse consequences is that any negative impact on individuals or groups should not derive from any source of test invalidity, such as construct underrepresentation or construct-irrelevant variance. Moreover, low scores should not occur because the measurement contains something irrelevant that interferes with the affected persons' demonstration of competence. (p. 746)

We argue that this same logic also works in the opposite direction. That is, higher scores should not occur because the measurement contains something irrelevant that interferes with an affected person's demonstration of lowered intellectual functioning. The Flynn Effect injects such construct irrelevant variance into the interpretation of test scores when professional psychologists do not account for it.

The Flynn Effect and its proper use in professional psychological practice might be cast in terms of the value implications to proper test score interpretation. Value implications are an integral aspect of proper test score interpretation and often link the construct being assessed to questions of applied practice and social policy (Messick, 1995). The proper use of the Flynn Effect in *Atkins* cases, we think, captures the essence of what Messick meant by value implications and proper test score interpretation. To this we would add that Principle 9.08 (Obsolete Tests and Outdated Test Results) of the "Ethical Principles of Psychologists and Code of Conduct" (American Psychological Association, 2002) states in part: "*(B) Psychologists do not base such decisions or recommendations on tests and measures that are obsolete and not useful for the current purpose* [italics added]." Failure to account for the Flynn Effect in test score interpretation in *Atkins* or any other cases is a violation of this ethical principle. In addition, failure to ensure the accurate interpretation of test scores in *Atkins* cases may possibly be a violation of the ethical Principle A: Beneficence and Nonmaleficence of the APA Code of Ethics. The principle states, in part, "Psychologists strive to benefit those with whom they work and take care to *do no harm* [italics added]." In their professional actions, psychologists seek to safeguard the welfare and rights of those with whom they interact professionally and other affected persons.

Given that *Atkins* held that it is a violation of the Eighth Amendment to the Constitution to execute persons who suffer from intellectual disability, it would seem that concluding individuals do not have intellectual disability without considering the Flynn Effect most certainly would cause undue harm and would violate the Constitutional rights of these individuals.

Conclusion

Standard of practice in the use of the Flynn Effect in the context of high stakes decisions must be guided by scientific evidence, not by opinion of psychologists. As Hagen et al. (2008) found in their survey, many psychologists are not aware of the underlying science and likely not cognizant of the high stakes contexts. Practicing psychologists claim to use an underlying psychological science as the foundation for clinical work. Application of the Flynn Effect and score adjustments for obsolete norms clearly is supported by science and should be implemented by professional psychologists.

References

American Educational Research Association. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., Text. rev.). Washington, DC: Author.

American Psychological Association. (2002). Ethical principles of psychologists and code of conduct. *American Psychologist, 57*, 1060–1073.

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

Ceci, S. J., & Kanaya, T. (2010). "Apples and oranges are both round": Furthering the discussion of the Flynn Effect. *Journal of Psychological Assessment, 28*, 441–447.

Cherry v. State, 959 So. 2d 702, 712-13 (Fla. 2007)

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

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302.

Doll, E. A. (1934). Social adjustment of the mental subnormal. *Journal of Educational Research*, *28*, 36–43.

Doll, E. A. (1941). The essentials of an inclusive concept of mental deficiency. *American Journal of Mental Deficiency*, *46*, 214–219.

Edgerton, R. B. (1993). *The clock of competence: Revised and updated*. Berkeley: University of California Press.

Edgerton, R. B., Ballinger, M., & Herr, B. (1984). The cloak of competence: After two decades. *American Journal of Mental Deficiency*, *88*, 345–351.

Fletcher, J. M., Stuebing, K. K., & Hughes, L. C. (2010). IQ scores should be corrected for the Flynn Effect in high-stakes decisions. *Journal of Psychoeducational Assessment, 28*, 469–473.

Flynn, J. R. (1985). Wechsler intelligence tests: Do we really have a criterion of mental retardation? *American Journal on Mental Deficiency, 90*, 236–244.

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

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101*, 171–191.

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

Flynn, J. R. (2009). The WAIS-III and WAIS-IV: *Daubert* motions favor the certainly false over the approximately true. *Applied Neuropsychology, 16*, 98–104.

Flynn, J. R. (2010). Problems with IQ gains: The huge vocabulary gap. *Journal of Psychoeducational Assessment, 28*, 412–433.

Flynn, J. R., & Widaman, K. F. (2008). The Flynn Effect and the shadow of the past: Mental retardation and the indefensible and indispensable role of IQ. *International Review of Research in Mental Retardation, 35*, 121–149.

Green v. Johnson, 2006 U.S. Dist. LEXIS 90644 (E.D. Va.), adopted by, 2007 U.S. Dist. LEXIS 21711 (E.D. Va.), aff'd., 2008 U.S. App. LEXIS 2967 (4th Cir.), cert. denied, 128 S. Ct. 2527 (2008).

Greenspan, S. (2006, spring). Issues in the use of the "Flynn Effect" to adjust IQ scores when diagnosing MR. *Psychology in Mental Retardation and Developmental Disabilities, 31*, 3–7.

Greenspan, S. (2009). Assessment and diagnosis of mental retardation in death penalty cases: Introduction and overview of the special "Atkins" issue. *Journal of Psychoeducational Assessment, 16*, 89–90.

Greenspan, S., & Switzky, H. (2006). Lessons from the *Atkins* decision for the next AAMR manual. In H. Switzky & S. Greenspan (Eds.), *What is mental retardation? Ideas for an evolving disability in the 21st century* (pp. 281–300). Washington, DC: American Association on Mental Retardation.

Gresham, F. M. (2009). Interpretation of intelligence test scores in *Atkins* cases: Conceptual and psychometric issues. *Applied Neuropsychology, 16*, 91–97.

Grossman, H. J. (Ed.). (1973). *Manual on terminology and classification in mental retardation*. Washington, DC: American Association on Mental Deficiency.

Hagan, L. D., Drogin, E. Y., & Guilmette, T. J. (2008). Adjusting IQ scores for the Flynn Effect: Consistent with standard of practice? *Professional Psychology: Research and Practice, 39*, 619–625.

Heber, R. (1961). *A manual on terminology and classification in mental retardation* (Rev. ed.). Washington, DC: American Association on Mental Deficiency.

Johnston v. State. 497 So. 2d 863 (Fla.1986).

Kanaya, T., Ceci, S., & Scullin, M. (2003). The Flynn Effect and U.S. policies: The impact of rising IQ scores on American society via mental retardation diagnoses. *American Psychologist, 58*, 778–790.

Kaufman, A. S. (2010). "In what way are apples and oranges alike?" A critique of Flynn's interpretation of the Flynn Effect. *Journal of Psychoeducational Assessment, 28*, 382–398.

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.

Matarazzo, D. (1972). *Wechsler's measurement and appraisal of adult intelligence* (5th enlarged ed.). Baltimore: Williams & Wilkins.

McGrew, K. S. (2010). The Flynn Effect and its critics: Rusty linchpins and "lookin' for g and Gf in some of the wrong places." *Journal of Psychoeducational Assessment, 28*, 448–468.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performance as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749.

Olley, J. G. (2009). Knowledge and experience required for experts in *Atkins* cases. *Applied Neuropsychology, 16*, 135–140.

Reschly, D. J. (1992). Mental retardation: Conceptual foundations, definitional criteria, and diagnostic operations. In S. R. Hooper, G. W. Hynd, & R. E. Mattison (Eds.), *Developmental disorders: Diagnostic criteria and clinical assessment* (pp. 23–67). Hillsdale, NJ: Erlbaum.

Reschly, D. J. (2009). Documenting the developmental origins of mild mental retardation. *Applied Neuropsychology, 16*, 124–134.

Reschly, D. J., Myers, T. G., & Hartel, C. R. (Eds.). (2002). *Mental retardation: Determining eligibility for Social Security benefits*. Washington, DC: National Academy Press.

Reschly, D. J., & Ward, S. M. (1991). Use of adaptive measures and overrepresentation of black students in programs for students with mild mental retardation. *American Journal of Mental Retardation*, 96, 257–268.

Schalock, R. L., Borthwick-Duffy, S. A., Bradley, V. J., Buntinx, W. H. E., Coulter, D. L., Craig, E. M., et al. (2010). *Intellectual disability: Definition, classification, and systems of supports* (11th ed.). Washington, DC: American Association on Intellectual and Developmental Disabilities.

Schalock, R., Buntinx, W., Borthwick-Duffy, Luckasson, R., Snell, M., Tassé, M., & Wehmeyer, M. (2006). *User's guide: Mental retardation, classification, and systems of supports,10th Edition: Applications for clinicians, educators, disability program managers, and policy makers*. Washington, DC: American Association on Intellectual and Developmental Disabilities.

Spearman, C. (1904). General intelligence: Objectively determined and measured. *American Journal of Psychology, 15*, 201–293.

United States v. Hardy (2010, November 24). U.S. District Court Eastern District of Louisiana, CA No. 94–381.

United States v. Lewis (2010, December 23). Case No.: 1:08 CR 404. U.S. District Court Northern District of Ohio Eastern Division. (Trial Opinion)

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale* (3rd ed.). San Antonio, TX: Psychological Corp.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale* (4th ed.). San Antonio, TX: Psychological Corp.

Weiss, L. G. (2007). *WAIS-III: Technical report, Response to Flynn*. San Antonio, TX: Psychological Corp.

Weiss, L. G. (2010). Considerations on the Flynn Effect. *Journal of Psychoeducational Assessment, 28*, 482–493.

Walker v. True (2006, August 30). U.S. District Court for the Eastern District of Virginia, Alexandria Division, Case No. 1:03–cv–00764.

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

Received 10/26/10, first decision 1/19/11, accepted 1/31/11.

Editor-in-Charge: Steven J. Taylor

Authors:
Frank M. Gresham, PhD (e-mail: frankgresham@yahoo.com), Professor, Department of Psychology, Louisiana State University, Baton Rouge, LA 70803.
Daniel J. Reschly, PhD, Professor, Department of Special Education, Vanderbilt University, Nashville, TN 37203.

