Professional Psychology: Research and Practice
2010, Vol. 41, No. 5, 413–419

0735-7028/10/$12.00 DOI: 10.1037/a0020226

Looking to Science Rather Than Convention in Adjusting IQ Scores When Death Is at Issue

Mark D. Cunningham
Independent practice, Dallas, TX

Marc J. Tassé
Ohio State University

The progressive obsolescence of IQ test norms and associated score inflation (i.e., the Flynn effect) may have literal life and death significance in capital mental retardation determinations (i.e., *Atkins* hearings). Hagan, Drogin, and Guilmette (2008) asserted that IQ score corrections for the Flynn effect were inconsistent with a "standard of practice" they deduced from custom, convention, and authority. More accurately, this reflected a proposed practice guideline or recommendation for practice, rather than a standard of practice. Whether a proposed guideline or recommendation for practice, these are better informed by an analysis of the available science than accepted convention. The authors reviewed research findings regarding the occurrence of the Flynn effect in the "zone of ambiguity" (IQ = 71–80), and proposed a best practice recommendation for discussing and reporting Flynn effect correction of IQ scores in capital mental retardation determinations.

Keywords: Flynn effect, death penalty, IQ, Atkins, mental retardation, practice recommendations

Consider the following scenario, reflecting an amalgam of several actual cases: A claim of mental retardation is brought by a 35-year-old death row inmate pursuant to *Atkins V. Virginia* (2002), the U.S. Supreme Court decision that barred the execution of individuals with mental retardation. There is particular focus in the postconviction *Atkins* hearing on whether the offender was a person with mental retardation at the time of the capital offense in 1995 and at the time of trial in 1997. Consistent with accepted definitions of mental retardation (American Psychiatric Association, 2000; Schalock et al., 2010), the inquiry is concerned with whether there is historical evidence of significantly subaverage intellectual functioning (i.e., IQ $\leq$ 70 ($\pm$5 when considering *SEM*), with concurrent deficits in adaptive behavior, before age 18. Review of the records revealed a WISC-R (Wechsler, 1974) Full Scale IQ score of 74 $\pm$ *SEM* in 1988 and a WAIS-R (Wechsler, 1981) Full Scale IQ score of 73 $\pm$ *SEM* in 1996. Significant deficits in several areas of adaptive functioning were evident before the defendant was imprisoned. Though informed of the imprecision of a specific IQ score, the court may make a "bright line" determination of whether the inmate's historical IQ score was 70 or below in ruling whether he is a person with mental retardation. The psychologist has extensive familiarity with the research findings regarding the progressive obsolescence of IQ test norms (i.e., Flynn effect) and the associated average 0.3 point annual IQ score inflation from the date the norms were collected for the respective scale. When the WISC-R was administered to this individual in 1988, 16 years had elapsed since it was normed in 1972. In 1996, the WAIS-R was 18 years beyond the midpoint of its 1976–1980 standardization. Correction for the associated inflation intervals would produce a corrected WISC-R Full Scale IQ score of 69 $\pm$ *SEM* and a corrected WAIS-R Full Scale IQ score of 68 $\pm$ *SEM*.

What "standard of practice" should guide the response of a psychologist in assisting the court to understand and make informed application of these historical IQ scores when the implications are literally life and death? In "Adjusting IQ scores for the Flynn Effect: Consistent with the standard of practice?" (see Hagan, Drogin, & Guilmette, 2008), Hagan et al. concluded regarding this standard:

> The current accepted convention does not support subtracting IQ points in a way that departs from the requirements of the test manual . . . Psychologists cannot conclude that adjusting scores is the generally accepted practice in evaluations for special education, parental rights termination, disability, or any other purpose. (p. 623)

This article was published Online First September 6, 2010.

MARK D. CUNNINGHAM received his PhD in clinical psychology from Oklahoma State University. He maintains an independent practice in greater Dallas, TX. His areas of research and practice include forensic evaluations, assessment of mental retardation, capital sentencing determinations, characteristics of capital offenders, and rates and correlates of prison violence.

MARC J. TASSÉ received his PhD in clinical psychology from Université du Québec à Montréal. He is a professor of psychology and psychiatry at Ohio State University, as well as director of the Ohio State University Nisonger Center, University Center for Excellence in Developmental Disabilities. His research and clinical interests include intellectual disability and autism spectrum disorders; adaptive behavior, test development, and support needs; and the assessment and treatment of psychiatric problems or problem behaviors co-occurring with developmental disabilities.

THE AUTHORS each derive income from evaluations and testimony at capital sentencing regarding issues of mental retardation. Drs. Cunningham and Tassé have been called by the defense in capital mental retardation determinations and have testified that the obsolescence of test norms is a potential source of error when interpreting historical IQ performances on tests of intelligence. Accordingly, each has reported *both* observed and Flynn effect adjusted IQ scores in respective reports and testimony in capital cases.

CORRESPONDENCE CONCERNING THIS ARTICLE should be addressed to Mark D. Cunningham, 6860 North Dallas Parkway, Suite 200, Plano, TX 75024. E-mail: mdc@markdcunningham.com

Atkins hearings are apparently subsumed under "any other purpose" by Hagan et al. (2008). We disagree with their method of analysis in arriving at the above "standard" and their conclusions regarding it.

The Flynn Effect Briefly Explained

To provide a brief context and overview, IQ scores are standard scores, no more than points of comparison with the ostensible mean and normal distribution of scores in the general population (i.e., $M = 100$, $SD = 15$). Accordingly, incremental inflation of IQ scores in the general population (i.e., $M > 100$) results in any observed IQ score being a progressively less accurate point of comparison as the interval increases between scale standardization and any particular test administration. Had the examinee taken the IQ test the year it was standardized, a more accurate comparison could be made between the examinee and the standardization sample. However, should the examinee take the same instrument 15 years later, the original standardization sample no longer accurately reflects the *contemporaneous* population. Both the Flynn effect and the associated necessity of periodically updating the norms of IQ tests were succinctly summarized by Kanaya, Scullin, and Ceci (2003) in their seminal article. Kanaya et al. described:

> Ever since the introduction of standardized IQ tests in the early 20th century, there has been a systematic and pervasive rise in IQ scores all over the world, including the United States. Known as the *Flynn effect* after James Flynn, the political scientist who has extensively documented this rise, the Flynn effect causes IQ test norms to become obsolete over time (Flynn, 1984, 1987, 1998). In other words, as time passes and IQ test norms get older, people perform better and better on the test, raising the mean IQ by several points within a matter of years. Once a test is renormed, which typically happens every 15–20 years, the mean is reset to 100, making the test harder and "hiding" the previous gains in IQ scores. (p. 778)

Psychological vs. Legal Standards

As a beginning point, there is a terminology problem. Hagan et al. (2008) utilize a definition of "standard" taken from a legal dictionary: "a model accepted as correct by custom, consent or authority" (p. 619, citing Black, 2004, p. 1441). However, in psychological practice, "standards" have a quite different meaning. As defined by the American Psychological Association (APA), "standards" are promulgated by APA as opposed to accepted convention. Further, ". . . standards are mandatory and may be accompanied by an enforcement mechanism" (p. 1048, APA, 2002; see also p. 2, Committee on Professional Practice and Standards, APA, 2005). Even the terminology of aspirational "practice guidelines" is the purview of a vetting process by APA. Thus, Hagan et al. are more properly either proposing guidelines for practice or arguing their view of recommendations for practice or "best practices," rather than "the standard of practice." This is not an inconsequential differential, as the courts and other legal consumers of our literature may not appreciate the role of "standards" as this terminology is applied to psychological practice.

The Unacknowledged Elephant in the Room

Though Hagan et al. (2008) did not overtly grapple with a capital scenario in their article, or even directly reference capital sentencing applications, *Atkins* cases are almost certainly the primary intended audience for their analysis and commentary. Indeed, Drs. Hagan and Drogin are practicing forensic psychologists. As noted above, the operational definition of "a standard" was taken from a *law* dictionary (i.e., Black, 2004). The case law cited by Hagan et al. involved mental retardation determinations in capital cases. Dr. Hagan testified in November 2005 as a prosecution-retained expert in a mental retardation determination for capital sentencing (*Walker v. True*, 2005). Dr. Hagan described in testimony that in the course of his case preparation, he first became aware of the "Flynn effect" by that name, a term he described as "a misnomer" and "a mischaracterization" (p. 460, 524, 525, *Walker v. True*, 2005). Further, Dr. Hagan has subsequently expressed opinions in his court testimonies that mirror the analysis of the article when called as an expert by the prosecution in *Atkins*-related proceedings, as illustrated in the following summary by the federal district court:

> Dr. Hagan testified that there is a lack of consensus as to the cause of the Flynn effect, though the generally accepted practice is to account for the Flynn effect by renorming standardized tests or by "address[ing] it in narrative form, but not to subtract IQ points that the individual has earned." (Resp. Ex. A at 32; *Winston v. Kelly*, 2009)

The backdrop of life or death hinging on a few IQ points must be acknowledged and engaged in any discussion of practice standards, practice guidelines, and/or best practices regarding IQ score adjustments for the Flynn effect.

The Unique Context and Implications of the Flynn Effect for Capital Sentencing

Whether scientifically informed IQ score adjustments should be made in Social Security disability determinations and special education classifications, as well as in capital sentencing, are certainly legitimate questions. However, we would argue that the necessity of precision and reliability in the determination increases with the stakes. Quite simply, death is different (see *Gardner v. Florida*, 1977; *Gregg v. Georgia*, 1977; *Lockett v. Ohio*, 1978; *Woodson v. North Carolina*, 1976). Further, the assessment and classification activities associated with intellectual assessment in general clinical practice or school psychology are distinct from those encountered in capital sentencing. Though not available for the consideration of Hagan et al. (2008), and quoted for its descriptive eloquence rather than authority, we find compelling the analysis of the federal district court in its capital mental retardation findings in *United States v. Davis* (2009) regarding this differential between clinical and forensic assessments in the application of the Flynn effect:

> Next, Dr. [name redacted] states that the Flynn effect is not routinely applied in *clinical* settings as a matter of professional practice . . . . While this may be true, the Court finds this to be completely irrelevant. This *is* a forensic context, and an important one in which a man's life hangs in the balance. The goals of an IQ assessment are dramatically different in the clinical versus the forensic setting. In the clinical context, the purpose of such an assessment is typically to get an accurate picture of the individual's current functioning so that appropriate systems of support may be devised to assist that individual in everyday living. In most cases, a recently normed instrument will be used for the IQ assessment, rendering unnecessary any Flynn adjust-

> ments. In the forensic context, however, where an individual's eligibility for a death sentence depends on a somewhat arbitrary numerical cutoff, precision and accuracy in determining that individual's score, both at present and in the past, become critically important. Eligibility for the death penalty is not a lottery, and a greater effort to achieve accurate results is both necessary and appropriate. (p. 22 of Memorandum Opinion)

It is not that "mental retardation" is defined differently in a capital context (see Macvaugh & Cunningham, 2009). Rather, historical testing is likely to take a greater role in *Atkins* cases, and the importance of "getting it right" is of graver magnitude when death is at issue.

Finding the Best Practice in Capital Applications of the Flynn Effect

The Frye Test or General Acceptance Standard

Hagan et al. (2008) framed their inquiry and discussion of the "standard" regarding adjusting IQ scores for the Flynn effect as "a model accepted as correct by custom, consent or authority" (p. 619, citing Black, 2004, p. 1441). In this construction of a standard of psychological practice, Hagan et al. have effectively adopted a well-known standard for the admissibility of scientific evidence in a legal context known as the *Frye* test or general acceptance standard (*Frye v. United States*, 1923): ". . . the thing from which the deduction is made must be sufficiently established to have gained general acceptance in the particular field in which it belongs" (at 1013).

Consistent with an application of the *Frye* test, the methodology of Hagan et al. (2008) focused on various sources of "general acceptance" as reflected in prevailing "custom, consent, and authority" (p. 620). These included doctoral training programs, practice patterns of ABPP-certified school psychologists, manuals from test publishers, contemporary applied texts, ethical canons and guidelines, and statutes and case law. Hagan et al. did not address practice patterns for *Atkins* evaluations that might reflect whether there is "general acceptance" of adjusting IQ scores for the Flynn effect in a capital context.

General Acceptance Versus Other Metrics for Evaluating Science

There are fundamental problems with framing a discussion of a standard of practice for psychologists (or more properly "best practices") in terms of the general acceptance or *Frye* standard. Of immediate import, if the question is engaged as a legal analysis, the *Frye* test has been superseded in federal court and a majority of states by the *Daubert* standard (*Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 1993). The *Daubert* decision calls upon courts to determine the admissibility of scientific evidence not simply in light of its general acceptance, but also or alternatively (i.e., nonexclusively) in light of a number of science-related factors. These include the relevance and reliability of the theory or technique, as reflected in considerations of whether the theory or technique is derived from the scientific method, has been or can be empirically tested, has a known or potential error rate, has been subjected to peer review, and/or has standards or controls concerning its operation. Though the *Daubert* standard incorporates "general acceptance" as one of the factors to consider, the additional considerations focus on the *quality of the science* supporting the methodology in question. Thus, from the standpoint of a legal admissibility standard, Hagan et al. (2008) framed their analysis in terms of a single-dimensional standard of general acceptance, without reference to the more recent and more prevalent admissibility standard that emphasizes examination of the underlying science.

Prevailing Practice Versus Scientifically Informed Practice

These two standards of admissibility for scientific evidence in the courtroom (i.e., general acceptance vs. quality of science) represent a critically important differential for how the Flynn effect is applied to mental retardation assessments in capital cases. To explain, in IQ testing and interpretation, "prevailing practice" (i.e., general acceptance) and "scientifically informed practice" may not be synonymous. We would assert that the highest levels of professional practice are exemplified by applications of the best available science. Training programs and patterns of practice, however, may lag behind this science by years or even decades. Indeed, Hagan et al. found in their survey that fewer than half of faculty respondents who taught or supervised graduate students in IQ test administration and interpretation self-described being "very familiar" with the Flynn effect. Further, among the responding program directors for APA-approved clinical, counseling, and school psychology doctoral programs who were *not* involved in teaching or supervising IQ testing, 90% self-described slight or no familiarity at all with the Flynn effect. Similarly, among board-certified (ABPP) school psychologists surveyed by Hagan et al. (2008), a third reported slight or no familiarity at all with the Flynn effect.

These findings are not disparate from those of Young, Boccaccini, Conroy, and Lawson (2007), who surveyed 20 mental health professionals (13 psychologists and 7 psychiatrists) who had conducted at least one evaluation of mental retardation in a capital case. Thirty percent of the psychologists and *all* of the psychiatrists acknowledged that they were *unfamiliar* with the Flynn effect by name, even though their orientation to this issue had been assisted by providing them with a description before questioning. A quarter of the psychologists and three-fourths of the psychiatrists reported that they were unaware of the name and of the effect of rising IQ scores and norm obsolescence. Young et al. further detailed:

> Several evaluators who had not heard of the effect made comments such as "what you described doesn't make very much sense to me" (psychiatrist) and "I've seen the opposite occur; they tend to rise a little bit" (psychologist). (p. 175)

Because scientific advances may neither be quickly nor ubiquitously reflected in instruction or practice, discussions of "standards of practice" that are anchored to "prevailing convention" may do little more than describe professional performance that is not overtly negligent. A clinician can hardly be faulted for a practice pattern that is common among professional peers, however tenuous the empirical underpinnings of that practice may be. A case in point is the centuries-long reliance of the medical profession on blood-letting as a therapeutic technique. Blood-letting was the

prevailing convention and by this rubric was inarguably the "standard of practice."

Taken to its logical conclusion, tying the standard of practice (or even "best practice") to prevailing convention may impose a veritable straightjacket of circularity on the ability of professional psychology to remain scientifically abreast. To illustrate the circularity problem of anchoring "standards of practice" to prevailing convention:

1. Prevailing convention defines standards of practice.
2. Practice outside of prevailing convention is pejoratively inconsistent with the standard.
3. Scientific advancements cannot be legitimately incorporated into professional practice until they become the prevailing convention.
4. The standard of practice does not allow the adoption of scientific advancements until they are the prevailing convention.

An alternative to the general acceptance or prevailing convention approach to professional standards is to employ a best science or *Daubert*-like analysis. Such a best science emphasis and the continuing progression in scientifically informed practice this emphasis allows are among the elements that inform "practice guidelines" as these are promulgated by APA (2002a):

> 2.8 Basis. Practice guidelines take into account the best available sources on *current theory*, *research* [emphasis added], ethical and legal codes of conduct, and/or practice within existing standards of care so as to provide a defensible basis for recommended conduct. (p. 1049)

Examining the Flynn Effect in Light of Science Rather Than Convention

Scientific Support and Practical Implications

Empirical and peer-reviewed support. The Flynn effect is the long-recognized and empirically demonstrated phenomenon of improving performances on IQ tests over the past half-century. An APA PsycINFO search utilizing key words "Flynn effect," "IQ gains," and "IQ inflation" yielded 112 peer-reviewed articles, books, book chapters, and dissertations addressing this phenomenon. An unabridged discussion of the Flynn effect and its impact on the mean IQ scores that serve as the basis for comparison of any particular observed IQ score is beyond the scope of this article (for an orientation see Flynn, 1984a, 1984b, 1987, 1998, 2000, 2006, 2007, 2009; Flynn & Weis, 2007; Kanaya et al., 2003; Neisser, 1998; Psychological Corporation, 1997).

Practical implications of progressively obsolete norms. The twin problems of IQ score inflation and associated progressively obsolete norms have been acknowledged by the publishers of the Wechsler scales. Indeed, the *WAIS-III Technical Manual* (Psychological Corporation, 1997) explained that IQ-score gains were a fundamental rationale for the periodic re-standardization of IQ tests, including their own scale.

> 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 score of the U.S. population on the WAIS-R was 100 in 1981, the inflation might cause it be about 105 in 1997. (p. 8)

Weiss (2008), in a Pearson technical report, described a 0.17 point annual IQ score inflation on the WAIS-III. Though lower than the 0.3 annual rate of IQ score inflation for the WAIS-III asserted by Flynn (2006), who also recommended an additional 2.34 correction for what he termed "the tree stump effect," some progressive score inflation is not disputed. Other evidence of norm obsolescence was provided with the technical information accompanying the WAIS-IV. Counterbalanced administrations of the WAIS-III and WAIS-IV accomplished as part of the WAIS-IV standardization yielded mean WAIS-III scores that were 2.9 points higher for general examinees ($n = 238$, 12-year annual inflation rate = 0.26 points; Pearson, 2008).

In light of the above findings by the test publisher, the scientific foundation for *not* authorizing corrections in historically obtained scores is elusive. Admittedly, debate and varied perspectives continue on precisely what score correction should be made to the WAIS-III in light of norms that were contemporaneous at the time of any particular administration. This variation in correction makes a strange argument, however, for making *no* correction at all to WAIS-III scores, or other tests in the Wechsler series for that matter (see Flynn, 2009). In agreement with Flynn, we would argue that the approximately true is preferable to the certainly false.

Though not addressing the inflation associated with scores obtained late in the standardization life of a particular IQ test, score inflation can be reset to zero by re-norming. Of course, remaining absolutely current with IQ score inflation would require test publishers to conduct virtually continuous re-standardization of their intelligence tests. This would be cost-prohibitive for test developers, not to mention the marketing challenge in recurrently persuading practitioners to update their testing materials and scoring procedures. Instead, IQ tests are re-normed at intervals dictated by practical economics rather than optimal accuracy. For example, the Wechsler series of intelligence tests reflect the following intervals in revisions, re-standardizations, and republishing (see Flynn, 2006): WISC (normed 1947-48, Wechsler, 1949) to WISC-R (normed 1972; Wechsler, 1974) = 25 years; WISC-R to WISC-III (normed 1989; Wechsler, 1991) = 17 years; WISC-III to WISC-IV (normed 2001, Wechsler, 2003) = 12 years; WAIS (normed 1953-54; Wechsler, 1954) to WAIS-R (normed 1978; Wechsler, 1981) = 25 years; WAIS-R to WAIS-III (normed 1995; Wechsler, 1997) = 17 years; WAIS-III to WAIS-IV (normed 2007-08; Wechsler, 2008) = 12 years. If a 0.3 point annual inflation rate of Full Scale IQ score is accepted, the group mean may have moved as much as seven points between standardization evolutions (25 years $\times$ 0.3 per year = 7.5).

Individual applications of group data. Hagan et al. (2008) frame the consideration of correcting individual IQ scores for the Flynn effect in terms of whether data regarding the group mean can be reliably applied to a specific individual. To illustrate, Hagan et al. stated:

> Of particular importance to the evaluating psychologist is whether the observed changes in group mean scores over time apply reliably to a specific individual. The question here is whether the FE's broad construct applies to a specific evaluee's IQ test scores, particularly when the individual's obtained score is offered as evidence in support of a theory to prove a legal fact. (p. 620)

This is a curious point of contention, at best. The interpretation of any IQ score involves utilizing information from the standardization group (which almost never contained the individual being assessed) to interpret the performance of a specific individual. Indeed, this application of group data to the individual constitutes virtually the entirety of the field of psychometrics, as well as being the scientific foundation for the practice of medicine and mental health sciences. The issue is not that group data will form the basis for deriving, understanding, and interpreting the individual IQ score. Rather, the issue is whether the group data are sufficiently representative and contemporary to form a sound basis for this individualization.

The Flynn Effect at the Mental Retardation Threshold

Though not raised by Hagan et al. (2008), a relevant consideration in whether to correct IQ scores for the Flynn effect in capital or other mental retardation assessment contexts involves whether progressive score inflation occurs at the lower portion of the bell curve. In other words, it is conceivable that the Flynn effect may occur toward the central area, but not at the tails of the IQ distribution. As applied to mental retardation determinations, this hypothesis is informed by group data regarding score inflation (i.e., the Flynn effect) in the "zone of ambiguity" (i.e., Full Scale IQ = 71–80). To explain, persons with Full Scale IQ $\leq$ 70 will meet the first diagnostic prong for mental retardation whether or not the Flynn effect is considered. Those with Full Scale IQ $>$ 80 will likely not meet the first diagnostic prong for mental retardation, regardless of any correction for the Flynn effect. Several studies demonstrate that the Flynn effect does occur between Full Scale IQ = 71–80, in the zone of ambiguity.

Spitz (1989) examined 15 studies comparing WAIS and WAIS-R Full Scale IQ scores, which in the aggregate, reflected a large portion of the intelligence curve. These studies utilized various combinations of counterbalanced, partially counterbalanced, and concurrent administrations of these scales. Lines of best fit demonstrated score inflation (Flynn effect) between Full Scale IQ scores 70–80. Spruill and Beck (1988) reported on WAIS vs. WAIS-R IQ scores for examinees with WAIS Full Scale IQ scores 70–84 (N = 35). Consistent with the expected score inflation associated with obsolete norms, these examinees exhibited Full Scale IQ scores that were 4.75 points higher on the WAIS. Fitzgerald, Gray, and Snowden (2007) compared WAIS-R vs. WAIS-III IQ scores for examinees in the mild mental retardation and borderline categories (N = 32). Again consistent with the expected score inflation, examinees averaged Full Scale IQ scores that were 4.1 points higher on the WAIS-R than they demonstrated on the WAIS-III.

The score inflating impact of obsolete norms has also been demonstrated in the lower IQ ranges in comparisons of the WAIS-III with the WAIS-IV. The WAIS-IV Technical Manual (Pearson, 2008) reported that examinees classified as "intellectual disability–mild" (n = 24) exhibited Full Scale IQ scores that were 4.1 points higher on the WAIS-III as compared to the WAIS-IV (12-year annual inflation rate = 0.34 points). Pearson (2008) additionally reported that examinees classified as "borderline intellectual functioning" obtained Full Scale IQ scores that were 2.2 points higher on the WAIS-III than WAIS-IV (12-year annual inflation rate = 0.18).

It could be argued that the sample sizes associated with the above studies are too small to provide reliable information. This assertion is substantially undercut by the small sample sizes of persons with mental retardation in the standardization samples of the WAIS series, particularly in the mild mental retardation classification that constitutes 85% of persons with mental retardation:

> WAIS IQ $\leq$ 70 (n not reported); WAIS-R IQ $\leq$ 69 (n = 43); WAIS-III IQ = 55–69 (n = 46); WAIS-IV IQ = 55–70 (n = 73).

It seems disingenuous or uninformed to complain of small samples in studies demonstrating the Flynn effect in the zone of ambiguity, while simultaneously asserting the reliability of scores from a Wechsler scale derived from small numbers of mildly mentally retarded persons in the standardization sample.

As part of a large-scale (N = 8,944) study of special education assessments of children (ages 6–17) reported by Kanaya et al. (2003), a subsample were examined regarding whether score inflation was demonstrated among those who had initial WISC-series Full Scale IQ scores that were 71 to 85 (n = 526). Consistent with the expectations of the Flynn effect, Kanaya et al. found a median IQ score inflation of 5.0 points for the WISC-R Full Scale IQ scores in comparison to WISC-III Full Scale IQ scores (n = 157), but no or negligible differences for comparisons of WISC-R to WISC-R (n = 192) or WISC-III to WISC-III (n = 177). Kanaya et al. concluded:

> Our results also show that the Flynn effect has an impact on which individuals are diagnosed MR and which are not, regardless of their actual cognitive ability. (p. 787)

The aggregate of these studies support a conclusion that the Flynn effect applies to Wechsler series scores in the IQ = 71–80 "zone of ambiguity."

Peer-Reviewed Support for Correcting Individual Scores for the Flynn effect

In light of the strong scientific evidence for the Flynn effect, and evidence that this progressive score inflation extends to the zone of ambiguity, a number of scholars have recommended correcting individual IQ scores for the Flynn effect in mental retardation assessments. Such peer review is a factor in the previously described *Daubert* standards for admissibility of scientific evidence in legal proceedings.

More specifically, professional guidelines propagated by the American Association on Intellectual and Developmental Disabilities (AAIDD), formerly the American Association on Mental Retardation (AAMR), an organization whose primary focus is on research, practice, and public policy regarding persons with mental retardation, recommended that professionals should consider the obsolescence of test norms when interpreting historical IQ scores (see Schalock et al., 2007; Schalock et al., 2010). Schalock et al. (2007) recommended making adjustments based on the Flynn

effect to the referent group's mean when interpreting an obtained IQ score from a test with old norms for the purpose of ruling-in or -out a diagnosis of mental retardation. More specifically, the *User's Guide: Mental Retardation* (Schalock et al., 2007), promulgated by AAIDD, prescribed: "Recognize the 'Flynn effect.'. . . In cases where a test with aging norms is used, a correction for the age of the norms is warranted" (pp. 20–21).

Other scholars have also advocated adjustment of individual test scores to account for the Flynn effect in *Atkins* cases (see Flynn, 2006, 2009; Greenspan, 2006, 2007; Macvaugh & Cunningham, 2009; Scullin, 2006). Young et al. (2007) left open the option of Flynn effect correction of IQ scores in capital mental retardation evaluations. Finally, though not overtly prescribing IQ score corrections for the Flynn effect, other scholars have come near that recommendation (Kanaya et al., 2003; Neisser, 1998; Reschly & Grimes, 2002; Tulsky, Saklofske, & Ricker, 2003).

Pandora's Box

Some might assert that corrections for progressive norm obsolescence in IQ scores in *Atkins* evaluations would open the door to all manner of score adjustments for gender, culture, or race (e.g., Moore, 2006). Regarding the latter, considerations of race in the application of the death penalty are particularly troubling. It bears noting that a number of Texas capital cases were remanded for new sentencing trials because racial factors had been incorporated into expert testimony regarding the violence risk assessments of these offenders (see *Saldano v. Texas*, 2000). Otherwise, when score adjustment considerations are accompanied by the depth of scientific findings that accompany the Flynn effect, and are not otherwise discriminatory in their impact, they may indeed warrant consideration of score correction.

Others may caution that correction of scores participates in the reification of IQ scores as having a precision that is unjustified. We do not advocate the use of a "bright line" when determining whether or not a person's intellectual functioning is significantly subaverage. However, rigidly adhering to the sole report of the obtained score, even when that score is derived from demonstrably obsolete norms, seems an even greater reification of what are simply norm-referenced performances. Further, courts in *Atkins* hearings inquire regarding IQ *scores* and may regard that it is the province of the court to evaluate the ecological validity of those scores.

Recommendations for "Best Practice"

This response began with a sobering and practical scenario, a scenario that must be engaged in any discussion of best practices in intellectual assessments made when life and death hang in the balance. In place of convention, prevailing practice, and authority, we assert that *best science* illuminates *best practice* and is fundamental to ethical conduct and professional standards. We find that a sufficient body of science supports interpreting obtained IQ scores in capital mental retardation hearings in reference to best estimates of norms that were contemporaneous to date of test administration, rather than historical standardization means. More specifically, we propose that best practice at capital sentencing is characterized by the following:

1. Report the obtained IQ scores from the historical testing.

2. Describe the Flynn effect and associated studies demonstrating the progressive inflation in the group mean and the effect of this on observed IQ scores, including in the zone of ambiguity (IQ = 71–80).

3. Report the corrected IQ scores calculated from the interval between the year the test was normed and the year the test was administered, multiplied by the associated annual inflation rate from the best synthesis of available normative data. The comparative norm group at the time the test was administered is specified as this is the most meaningful interpretation of a norm-referenced performance, i.e., what did the obtained score mean in relation to the contemporaneous norm group at the time that it was obtained?

We assert that this procedure constitutes a scientifically informed, ethically sound, and clinically transparent practice at capital sentencing (see APA, 2002a, 2.04 Bases for Scientific and Professional Judgments, 3.04 Avoiding Harm, 9.02 Use of Assessments; Committee on Ethical Guidelines for Forensic Psychologists, 1991: VI. Methods and Procedures, Section A). The death implications of *Atkins* evaluations and the application of best science call for supplementary reporting of IQ scores that are adjusted in light of progressively inflating norms when describing intellectual assessments in a capital context.

References

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

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

American Psychological Association. (2002a). Criteria for practice guideline development and evaluation. *American Psychologist, 57,* 1048–1051.

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

Black, H. C. (2004). *Black's law dictionary* (8th ed.). St. Paul, MN: Thomson-West.

Committee on Ethical Guidelines for Forensic Psychologists. (1991). Specialty guidelines for forensic psychologists. *Law and Human Behavior, 15,* 655–665.

Committee on Professional Practice and Standards, Board of Professional Affairs. (February, 2005). *Determination and documentation of the need for practice guidelines.* Washington, DC: Author.

Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993).

Fitzgerald, S., Gray, N. S., & Snowden, R. J. (2007). A comparison of WAIS-R and WAIS-III in the lower IQ range: Implications for learning disability diagnosis. *Journal of Applied Research in Intellectual Disabilities, 20,* 323–330.

Flynn, J. R. (1984a). IQ gains and the Binet decrements. *Journal of Educational Measurement, 21,* 283–290.

Flynn, J. R. (1984b). 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. (1998). WAIS-III and WISC-III: IQ gains in the United States from 1972 to 1995; how to compensate for obsolete norms. *Perceptual and Motor Skills, 86,* 1231–1239.

Flynn, J. R. (2000). The hidden history of IQ and special education: Can the problem be solved? *Psychology, Public Policy, and Law, 6,* 191–198.

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. (2007). Capital offenders and the death sentence: A scandal that must be addressed. *Psychology in Mental Retardation and Developmental Disabilities, 32,* 3–7.

Flynn, J. R. (2009). The WAIS-III and WAIS-IV: Daubert motions favor the certainly false over the approximately true. *Applied Neuropsychology, 16,* 1–7.

Flynn, J. R., & Weis, L. G. (2007). American IQ gains from 1932 to 2002: The WISC subtests and educational progress. *International Journal of Testing, 7,* 209–224.

Frye v. United States, 293 F. 1013 (D. C. Cir. 1923).

Gardner v. Florida, 430 U.S. 349 (1977).

Greenspan, S. (2006). 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. (2007). Flynn-adjustment is a matter of basic fairness: Response to Roger B. Moore, Jr. *Psychology in Mental Retardation and Developmental Disabilities, 32,* 7–8.

Gregg v. Georgia, 428 U.S. 153, 231 (1976).

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

Kanaya, T., Scullin, M. H., & Ceci, S. J. (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.

Locket v. Ohio, 438 U.S. 586, 604 (1978).

Macvaugh, G., & Cunningham, M. D. (2009). *Atkins v. Virginia*: Implications and recommendations for forensic practice. *Journal of Psychiatry and Law, 37,* 131–187.

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

Moore, R. B. (2006). Modification of individual's IQ scores is not accepted professional practice. *Psychology in Mental Retardation and Developmental Disabilities, 32,* 11–12.

Neisser, U. (Ed.). (1998). *The rising curve: Long-term gains in IQ and related measures.* Washington, DC: American Psychological Association.

Pearson (2008). *WAIS-IV technical and interpretive manual.* San Antonio, TX: Author.

Psychological Corporation. (1997). *WAIS-III, WMS-III technical manual.* San Antonio, TX: Author.

Reschly, D. J., & Grimes, J. P. (2002). Best practices in intellectual assessment. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology IV* (pp. 1337–1350). Bethesda, MD: The National Association of School Psychologists.

Saldano v. Texas, 530 U.S. 1212 (2000).

Schalock, R. L., Buntinx, W. H. E., Borthwick-Duffy, S., Bradley, V., Craig, E. M., Coulter, D. L., . . . Yeager, M. H. (2010). *Mental retardation: Definition, classification, and system of supports.* Washington, DC: American Association on Intellectual and Developmental Disabilities.

Schalock, R. L., Buntinx, W. H. E., Borthwick-Duffy, S., Luckasson, R., Snell, M. E., Tassé, M. J., & Wehmeyer, M. L. (2007). *User's guide: Mental retardation: Definition, 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.

Scullin, M. H. (2006). Large state-level fluctuations in mental retardation classifications related to introduction of renormed intelligence test. *American Journal of Mental Retardation, 111,* 322–335.

Spitz, H. H. (1989). Variations in Wechsler interscale IQ disparities at different levels of IQ. *Intelligence, 13,* 157–167.

Spruill, J., & Beck, B. L. (1988). Comparison of the WAIS and WAIS-R: Different results for different IQ groups. *Professional Psychology: Research and Practice, 19,* 31–34.

Tulsky, D. S., Saklofske, D. H., & Ricker, J. H. (2003). *Clinical interpretation of the WAIS-III and WMS-III: Practical resources for the mental health professional.* Boston: Elsevier.

United States v. Davis, 611 F. Supp. 2d 472, 488 (D. Md. 2009).

Walker vs. True, Case No. 1:03-cv-764, in the U.S. District Court for the Eastern District of Virginia, Alexandria Division. Trial transcript. Volume 3, November 1, 2, and 8, 2005.

Wechsler, D. (1949). *Wechsler Intelligence Scale for Children (WISC).* New York: The Psychological Association.

Wechsler, D. (1954). *Wechsler Adult Intelligence Scale (WAIS).* New York: The Psychological Corporation.

Wechsler, D. (1974). *Wechsler Intelligence Scale for Children–Revised (WISC-R).* New York: The Psychological Corporation.

Wechsler, D. (1981). *Wechsler Adult Intelligence Scale–Revised (WAIS-R).* New York: The Psychological Corporation.

Wechsler, D. (1991). *Wechsler Intelligence Scale for Children–Third edition. (WISC-III).* San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale–Third edition. (WAIS-III).* San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children–Fourth edition. (WISC-IV).* San Antonio, TX: Pearson.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale–Fourth edition. (WAIS-IV).* San Antonio, TX: Pearson.

Weiss, L. G. (2008). *WAIS-III Technical report: Response to Flynn.* Retrieved from Pearson Website: http://www.pearsonassessments.com/NR/rdonlyres/98BBF5D2–F0E8–4DF6–87E2–51D0CD6EE98C/0/WAISIII_TR.pdf

Winston v. Kelly, Civil Action No. 7:07cv00364, in the U.S. District Court for the Western District of Virginia, Roanoke Division, Memorandum opinion by Samuel G. Wilson, United States District Judge, 3–6-09.

Woodson v. North Carolina, 438 U.S. 304 (1976).

Young, B., Boccacini, M. T., Conroy, M. A., & Lawson, K. (2007). Four practical and conceptual assessment issues that evaluators should address in capital case mental retardation evaluations. *Professional Psychology: Research and Practice, 38,* 169–178.

Received December 23, 2009
Revision received May 7, 2010
Accepted May 10, 2010 ■