No. _____

IN THE

Supreme Court of the United States

TAVARES J. WRIGHT,

Petitioner,

v.

SECRETARY, DEPARTMENT OF CORRECTIONS, AND ATTORNEY GENERAL, STATE OF FLORIDA,

Respondents.

ON PETITION FOR A WRIT OF CERTIORARI TO THE UNITED STATES COURT OF APPEALS FOR THE ELEVENTH CIRCUIT

APPENDIX TO THE PETITION FOR A WRIT OF CERTIORARI

VOLUME IV

DEATH PENALTY CASE

ADRIENNE JOY SHEPHERD
FLORIDA BAR NUMBER 1000532
SHEPHERD@CCMR.STATE.FL.US

ALI A. SHAKOOR
FLORIDA BAR NUMBER 0669830
SHAKOOR@CCMR.STATE.FL.US

LAW OFFICE OF THE CAPITAL COLLATERAL
REGIONAL COUNSEL - MIDDLE REGION
12973 NORTH TELECOM PARKWAY
TEMPLE TERRACE, FLORIDA 33637
(813) 558-1600

Counsel for Petitioner

APPENDIX TABLE OF CONTENTS

Contents **Page**

Volume I:

Volume II:

Volume III:

Appendix Table of Contents

Contents **Page**

Volume IV:

No. _____

IN THE

Supreme Court of the United States

TAVARES J. WRIGHT,

Petitioner,

v.

SECRETARY, DEPARTMENT OF CORRECTIONS, AND ATTORNEY GENERAL, STATE OF FLORIDA,

Respondents.

ON PETITION FOR A WRIT OF CERTIORARI TO THE UNITED STATES COURT OF APPEALS FOR THE ELEVENTH CIRCUIT

APPENDIX TO THE PETITION FOR A WRIT OF CERTIORARI

DEATH PENALTY CASE

APPENDIX N

Chart of IQ Scores for Tavares Wright Entered as Defense Exhibit One at January 5, 2015 Hearing

IN THE Circuit COURT, CRIMINAL DIVISION

POLK COUNTY, FLORIDA

Defendant Tavares Wright : Case No. 00CF-2727

Attorney Ahmad : Judge Jacobsen

Clerk K. Blackshear : State Attorney Pattey

Court Reporter D. Lamb : Date 1.5.15 + T.6.15

Law Enforcement Case No. _____ : Verdict Evid NRG

EVIDENCE REPORT

Def EXHIBIT NO. 1 ✓

Filed for I.D. this date 1.5.15 : Filed in Evidence this date 1.5.15

Testing Chart

States EXHIBIT NO. 1 ✓

Filed for I.D. this date 1.6.15 : Filed in Evidence this date 1.6.15

Chart of IQ Testing

State EXHIBIT NO. 2 ✓

Filed for I.D. this date 1.6.15 : Filed in Evidence this date _____

CD

State EXHIBIT NO. 3 ✓

Filed for I.D. this date 1.6.15 : Filed in Evidence this date 1.6.15

Event # 00-04-05801

Trans. of TJ's Taped Statement

_____ EXHIBIT NO. _____

Filed for I.D. this date _____ : Filed in Evidence this date _____

Defense EXHIBIT NO. 2 ✓ 1-23-15

Filed for I.D. this date 1.6.15 : Filed in Evidence this date 1.6.15

Full Resume of Dr. Freid eval 1997

_____ EXHIBIT NO. _____

Filed for I.D. this date _____ : Filed in Evidence this date _____

_____ EXHIBIT NO. _____

Filed for I.D. this date _____ : Filed in Evidence this date _____

_____ EXHIBIT NO. _____

Filed for I.D. this date _____ : Filed in Evidence this date _____

RECEIVED AND FILED

JAN 07 2015

STACY M. BUTTERFIELD, CLERK

_____ EXHIBIT NO. _____

Filed for I.D. this date _____ : Filed in Evidence this date _____

1. Original - White 2. Clerk's Evidence - Canary 3. Property Evidence - Pink 4. State Attorney - Goldenrod

Wright Age	Test	Date Admin	Date Norm	Yrs old	FSIQ	Flynn Cor FSIQ	Cutoff Scores for MR/ID AAMR Retarded Benchmark
10-0	WISC-R	2/1991 (?)	1972	19	76	70	76.27+ 2 SEM(3.14)= 82.55
10-2	WISC-R	4/9/91	1972	19	80	74	76.27+ 2 SEM(3.14)= 82.55
10-7	WISC-R	9/11/91	1972	19	81	75	76.27+ 2 SEM(3.14)= 82.55
16-6	WAIS-R	8/25/97	1978	19	75	69	76.27 +2SEM(2.96) = 82.19
			SCORES OBTAINED AFTER AGE 18				
20	WASI	Around 10/01	1997/98	3	None	None	
21	WASI	2/4/03	1997/98	5	None	None	
24	WAIS-III	7/15/05	1995	[redacted] 10	82	~~80~~ 77	72.34 with 2 SEM CI = ~~90~~ 87
24	WAIS-III	7/25/05	1995	[redacted] 10	75	[redacted] 70	72.34 with 2 SEM CI = ~~90~~ 87

1461

2534

Case No. 00CF-2727
Def. Exhibit No. 1
Filed for I.D. 1.5.15
Filed in Evidence
Stacy M Butterfield, Clerk

No. _____

IN THE

Supreme Court of the United States

TAVARES J. WRIGHT,

Petitioner,

v.

SECRETARY, DEPARTMENT OF CORRECTIONS, AND ATTORNEY GENERAL, STATE OF FLORIDA,

Respondents.

ON PETITION FOR A WRIT OF CERTIORARI TO THE UNITED STATES COURT OF APPEALS FOR THE ELEVENTH CIRCUIT

APPENDIX TO THE PETITION FOR A WRIT OF CERTIORARI

DEATH PENALTY CASE

APPENDIX O

Academic Articles Concerning the Flynn Effect

Leigh D. Hagan et al., *Adjusting IQ Scores for the Flynn Effect: Consistent With the Standard of Practice?*, 39 PROFESSIONAL PSYCHOLOGY: RESEARCH AND PRACTICE 619 (2008)

Leigh D. Hagan, et al., *IQ Scores Should Not Be Adjusted for the Flynn Effect in Capital Punishment Cases*, JOURNAL OF PSYCHOEDUCATIONAL ASSESSMENT 474 (2010)

Jack M. Fletcher, et al., *IQ Scores Should Be Corrected for the Flynn Effect in High-Stakes Decisions,* JOURNAL OF PSYCHOEDUCATIONAL ASSESSMENT 469 (2010)

Mark D. Cunningham & Marc J. Tasse, *Looking to Science Rather Than Convention in Adjusting IQ Scores When Death is at Issue*, 41 PROFESSIONAL PSYCHOLOGY: RESEARCH AND PRACTICE 413 (2010)

Cecil R. Reynolds, et al., *Failure to Apply the Flynn Correction in Death Penalty Litigation: Standard Practice of Today Maybe, but Certainly Malpractice of Tomorrow*, JOURNAL OF PSYCHOEDUCATIONAL ASSESSMENT 477 (2010)

Frank M. Gresham and & Daniel J. Reschly, *Standard of Practice and Flynn Effect Testimony in Death Penalty Cases*, 49 INTELLECTUAL AND DEVELOPMENTAL DISABILITIES 131 (June 2011)

Lisa Trahan, et al., *The Flynn Effect: A Meta-analysis*, PSYCHOLOGICAL BULLETIN, Sept. 2014

Marc J. Tasse, *Adaptive Behavior Assessment and the Diagnosis of Mental Retardations in Capital Cases*, 16 APPLIED NEUROPSYCHOLOGY 114 (2009)

Professional Psychology: Research and Practice
2008, Vol. 39, No. 6, 619–625

0735-7028/08/$12.00 DOI: 10.1037/a0012693

Adjusting IQ Scores for the Flynn Effect: Consistent With the Standard of Practice?

Leigh D. Hagan
Virginia Commonwealth University

Eric Y. Drogin
Harvard Medical School

Thomas J. Guilmette
Providence College

Should psychologists adjust obtained IQ scores to accommodate the *Flynn effect* (J. R. Flynn, 1985)? The authors surveyed directors of doctoral training programs approved by the American Psychological Association and board-certified school psychologists and completed a systematic review of IQ test manuals, contemporary textbooks on IQ testing, federally regulated IQ testing protocols, and various sources of legal and ethical guidance. They confirmed in each instance that such adjustments to IQ scores do not comport with prevailing standards of psychological practice. Results of IQ testing may be applied to a broad range of psycholegal issues, many of which cannot be anticipated. Psychologists assist examinees, courts, and other 3rd parties most effectively by administering and interpreting IQ tests in their intended fashion.

Keywords: practice standards, Flynn effect, IQ, intelligence testing

Each year psychologists assist in hundreds of thousands of legal determinations through evaluation reports and expert testimony based on scientific knowledge of measurement procedures, including intelligence testing. Psychologists' reports of IQ test data can have a major impact on access to services and even life-and-death decisions (*Atkins v. Virginia,* 2002). In addition to specific medicolegal evaluations, psychologists administering an IQ test for one purpose, such as treatment planning or special education, might find their work product used for a different purpose years later in a criminal proceeding, disability evaluation, or claim of damages in a lawsuit. Given the possible intended and unintended consequences of intelligence test records, understanding and comporting with practice standards is essential. A debatable but potentially emerging standard is whether psychologists should subtract points from an individual's obtained IQ score on the basis of the Flynn effect (FE; Flynn, 1985), a phenomenon in which IQ means have been shown to increase in the general population across time.

Why Standards Make a Difference

A standard is "a model accepted as correct by custom, consent, or authority" (Black, 2004, p. 1441). Standards establish parameters of practice and communicate the prevailing views of psychology to those outside of behavioral science. Psychological practice standards do not exist in a vacuum. Law, science, and ethical principles impact each other; none stand in isolation. In the psycholegal context, each guides the psychologist who, in turn, advises the court about prevailing standards.

The FE and Adjusting IQ Scores

The FE refers to the finding that the general population's average IQ test scores have increased over the past several decades (Flynn, 1985). Although some studies have reported an increase of about 0.30 IQ points per year (Flynn, 1999), the issues underpinning the changes in average scores over time are complex and exceed the scope of this article. The research-informed practitioner should note the differential impact of a host of variables, including gender and ethnicity ("Latest Thinking," 2007), age and culture (Flynn, 1987), level of industrial and technological development (Daley, Whaley, Sigman, Espinosa, & Neumann, 2003; Flynn, 1987), the type of cognitive task being measured (fluid or crystallized), and where the score falls along the distribution curve (Zhou & Zhu, 2007).

LEIGH D. HAGAN received his PhD in counseling psychology from the University of Missouri—Columbia. He maintains an independent clinical and forensic practice in Chesterfield, Virginia, and is an affiliate assistant clinical professor of psychology at Virginia Commonwealth University. His areas of professional interest include capital defendant sentencing, custody evaluation, and practice standards for mental health professionals in the forensic arena.

ERIC Y. DROGIN received his PhD in clinical psychology from Hahnemann University and received his JD from the Villanova University School of Law. He maintains an independent forensic consulting practice and serves on the faculty of the Harvard Medical School, Beth Israel Deaconess Medical Center, as a clinical instructor and member of the Program in Psychiatry and the Law. His areas of professional interest include trial consultation and expert testimony.

THOMAS J. GUILMETTE received his PhD in counseling psychology from the University of Missouri—Columbia. He is a professor of psychology at Providence College and an adjunct associate professor of psychiatry and human behavior at the Warren Alpert Medical School of Brown University. His professional and research interests include neuropsychology and psychology and law.

CORRESPONDENCE CONCERNING THIS ARTICLE should be addressed to Leigh D. Hagan, P.O. Box 350, Chesterfield, VA 23832. E-mail: lhagan@leighhagan.com

Flynn (2007) documented a wide range of score fluctuations, including a slight reverse of the FE, depending on which Wechsler scale was used. Some countries have actually shown a reverse FE in more recent years (Shayer, Ginsburg, & Coe, 2007).

Our research focuses on the straightforward question: Is it the standard of practice to adjust obtained IQ scores in light of the FE? To the extent that the empirical impact of the FE is blind to the purpose for which a test is administered, then practicing psychologists need to be cognizant of this issue, not just for criminal evaluations, but for special education, disability, employment, and any other purpose. Although mainstream recognition of the FE as an authentic psychometric consideration has increased, the question of how to accurately represent its impact for a particular individual's earned scores on IQ tests is a different question altogether.

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. Specifically, is it the generally accepted practice in the field of psychological testing to adjust a particular person's earned IQ scores or to recalculate norm means on the basis of the FE?

Flynn has advanced several different positions on this point. In 1987, he cautioned against placing unwarranted emphasis on individual IQ scores, asserting that "IQ tests do not measure intelligence but rather a correlate with a weak causal link to intelligence" (Flynn, 1987, p. 171). Later, he took the position that the Wechsler Adult Intelligence Scale (3rd ed.; Wechsler, 2002) might be reliable for scores below 70 and concluded that the FE was a factor of 0.25 rather than 0.30 (Flynn, 1998). Shortly thereafter, in 2000, he proposed abandoning the use of IQ scores for mental retardation determination rather than adjusting obtained scores, arguing that "the fact that people will get quite different scores on different IQ tests can be manipulated by psychologists to suit their clients' needs" (Flynn, 2000, p. 191).

In 2006, Flynn advocated adjusting individual IQ scores on the premise that doing so creates no greater error than failing to do so. He argued that resistance to the practice of subtracting points from an individual's obtained score was not particularly defensible. Yet, within the same article, he pointed out that the FE is not generally accepted in the clinical field. Most recently, with respect to deducting 0.30 IQ points per year, Flynn (2007) acknowledged that "recommending such a simple cure for obsolete norms assumes too much" (Flynn, 2007, p. 134).

Although Flynn's position about IQ scores varies in his scholarly articles, he steadfastly advocates subtracting obtained IQ points in criminal sentencings (e.g., *Berry v. Mississippi*, 2005; *Walker v. True,* 2005). To the extent that the FE is a function of IQ tests generally, and if adjusting an individual's obtained IQ scores is the accepted convention in clinical practice, then one would expect to find empirically based support for individual score adjustments across all IQ test purposes. One would not expect to find the discussion limited to a narrow range of purposes, such as capital case advocacy. Yet, the professional literature is almost silent on individual score adjustments outside of the criminal forensic arena.

Although the FE appears in hundreds of articles, most are of a technical nature or focus on social policy implications. Very few psychologists forward the position that an individual's obtained IQ scores should be reduced by a numerical factor based on the FE. Kanaya, Scullin, and Ceci (2003) argued for score adjustments on the basis of a large scale empirical study. Greenspan (2006), in a discussion article absent new empirical data, asserted that subtracting IQ points from an individual's obtained score is not only appropriate, but essential. Other psychologists have argued through their reports and testimony in the capital-sentencing context that adjusting scores is the normative practice (*Bowling v. Kentucky,* 2005; *Green v. Johnson,* 2008; *Howell v. Tennessee,* 2004; *People v. Superior Court [Vidal],* 2005; *Walker v. True,* 2005; *Walton v. Johnson,* 2006;), but they drew from work previously cited without adding to the empirical research base of knowledge.

Division 33 of the American Psychological Association (APA) called for an ad hoc committee to further study this issue and to find those areas in consensus on standards for psychologists (Olley, Greenspan, & Switzky, 2006). Beyond the works previously cited, we found no empirical studies advocating for FE-based score adjustments in special education, disability, parental rights termination, or any other purpose for which psychologists ordinarily administer IQ tests.

A dichotomy sometimes emerges between scholarly empirical research and expert testimony in the courtroom. Cases abound in which expert witnesses have testified that adjusting an individual's obtained IQ score is the standard (*Commonwealth v. Prieto,* 2007; *Green v. Johnson,* 2008; *People v. Superior Court [Vidal],* 2005; *State v. Keel,* 2003; *Walker v. True,* 2005). In these same cases, however, other qualified experts have testified that adjusting IQ scores is not the accepted practice.

Other scholars and expert witnesses oppose adjusting IQ scores for several reasons. Moore (2006) challenged the proposition that adjusting individual IQ scores is the standard of practice. Lacritz and Cullum (2003) advised that "caution should be used in applying Flynn's philosophy to actual patients, as there are many sources of variance unaccounted for by his formulas that could impact an individual's score" (p. 529).

Young, Boccaccini, Conroy, and Lawson (2007) provided the closest analysis to date with respect to the standard of practice and IQ score adjustment in death penalty evaluations. They found that among experienced death penalty evaluators, most psychologists reported being aware of the FE either by name or the underlying construct, yet most (71%) of the psychiatrists surveyed had never heard of the concept underpinning the FE. Olley et al. (2006) also pointed out the lack of consensus about how to present IQ data for *Atkins* hearings (see *Atkins v. Virginia,* 2002) for the court to determine if the capital defendant meets the statutory criteria for mental retardation. We investigate whether there presently exists a standard for adjusting individually obtained IQ scores in a way that is accepted as correct in light of custom, consent, and authority.

Search for a Standard

Survey 1: Doctoral Program Directors

Participants were program directors of APA-approved clinical, counseling, and school psychology doctoral programs as identified by their respective APA Web sites. Of the surveys sent to each of 358 program directors, all respondents were program directors, IQ/intelligence instructors or supervisors, or a combination of both cat-

egories. The largest portion (43%) received their doctoral degree more than 20 years ago. Most (69%) taught or supervised doctoral students' IQ testing in the previous 3 years. We did not solicit information about the respondents' forensic experience specifically but did inquire about their knowledge of the FE in any arena.

The survey questions were not limited to any specific IQ testing purpose. Respondents were instructed to stop filling out the survey and return it if they were not at all familiar with the FE. The remaining items sought to determine whether graduate school faculty members were teaching their students to calculate, adjust, and list scores on the basis of the FE in ways that have previously been described in some cases as the accepted professional standard (*Commonwealth v. Prieto,* 2007) or as near universal (*Green v. Johnson,* 2008).

We found that of the 89 respondents, 36% indicated that their familiarity with the FE was slight or that they had no familiarity at all; 37% were moderately familiar, whereas 27% were very familiar.

Because our focal interest was in contemporary teaching practices, the balance of the data analysis was derived from the responses of those faculty who indicated that they had taught or supervised graduate student IQ testing and interpretation within the previous 3 years. Excluding those who had not taught or supervised students also eliminated respondents who were not at all familiar with the FE. Of the remaining 57 respondents, 93% reported that they had taught or supervised IQ testing in the past 3 years.

Table 1 reveals that, of this group, 82% indicated that it was only slightly important or not at all important for students to learn to calculate the FE when listing actual scores in the written report. In addition, although 61% believed that it was moderately or very important for students to learn to consider the FE when interpreting scores, only 18% indicated that it was very important, which is the same as the percentage who believed that it was not at all important.

Simply considering the FE is not the same enterprise as memorializing that thought process in the narrative of a written report. Thus, Table 2 reveals the frequency with which the participants taught their graduate students to comment on the FE in reports or to actually recalculate or adjust IQ scores based on the FE. As can be seen in Table 2, two thirds of the respondents never taught students to comment on the FE, and 9 out of 10 never taught their students to adjust or recalculate IQ scores.

The survey inquired about teaching students to adjust IQ scores depending on where in the distribution the score might fall. The vast majority (94%) reported that they never taught students to adjust obtained IQ scores, irrespective of their position in the distribution. Only 2% advocated adjusting IQ scores across the entire range.

Rather than adjusting obtained IQ scores, some psychologists have proposed compensating for the FE by adjusting the mean score from the published norms and then reporting the obtained score relative to the newly adjusted mean (*Green v. Johnson,* 2008). Teaching students to adjust obtained scores after recalculating the published means was even less likely, with 95% never instructing in this practice. None of the respondents indicated that they promoted this practice for all IQ testing referral questions.

Some researchers and testifying experts (Flynn, 2006; *Green v. Johnson,* 2008; Kanaya et al., 2003; *People v. Superior Court [Vidal],* 2005; *Walker v. True,* 2005;) have advocated adjusting the obtained IQ score, not just for each year after the publication of the test, but also for each year after the normative data were collected. This procedure accounts for the postulated lag between data collection and publication of the test manual. Flynn (2006) referred to this process as "the general rule" (p. 179).

The survey polled for this practice. Of the participants, 79% (45 out of 57) did not teach their students to make numerical adjustments to the obtained IQ, but of those who did, the majority (75% or 9 out of 12) relied on the year the norm group was collected when adjusting the IQ.

No consensus emerged about a scientific authority for adjusting scores. The much larger majority (86%) declined to identify any scientific, legal, regulatory, or ethical authority for adjusting obtained scores or recalculating means because they did not train students to use this practice.

Survey 2: Diplomates in School Psychology

The second survey queried clinicians who had achieved the advanced credential of board certification in school psychology from the American Board of Professional Psychology. We chose these psychologists because they frequently engage in intelligence testing and have considerable experience and expertise in interpreting archival test data.

Participants in Survey 2 were all of the 141 American Board of Professional Psychology school psychologists identified by the board's Web site. We received 28 usable returns, or 23% of the viable pool. The majority had over 20 years of experience. Most (93%) of the viable respondents had personally administered, scored, and interpreted more than 200 individual IQ tests.

The majority (68%) were moderately or very familiar with the FE. A large majority (94%) of the viable participants reported that they had never adjusted obtained IQ scores on the basis of the FE when reporting numerical IQ scores. Only one participant reported adjusting obtained scores in some cases (few but less than most). None reported doing so in most or all cases. Only one reported

Table 1
Percentage of Participants Who Considered It Important for Students to Learn to Calculate or Consider the FE in Written Reports

Item	Not important	Slightly important	Moderately important	Very important
Learning to calculate the FE when listing scores in written reports	46.4	35.7	14.3	3.6
Learning to consider the FE when interpreting scores in written reports	17.9	21.4	42.9	17.9

Note. Participants were program directors or instructors of IQ testing courses in clinical, counseling, or school psychology programs approved by the American Psychological Association ($n = 56$). FE = Flynn effect.

Table 2
Percentage of Participants Who Taught Students to Comment on the FE or Recalculate IQ Scores on the Basis of the FE in Written Reports

Item	Never	Yes, in all cases	In MR cases only	In certain other cases	In MR and certain other cases
Teach students to comment on the FE in reports	68.5	3.7	7.4	18.5	1.9
Teach students to recalculate IQ scores on the basis of the FE	91.9	0.0	3.6	3.6	0.0

Note. Participants were program directors or instructors of IQ testing courses in clinical, counseling, or school psychology programs approved by the American Psychological Association (n = 54 and 56, respectively, for Item 1 and Item 2). FE = Flynn effect; MR = mental retardation.

commenting on the FE in the written narrative. None of the respondents reported having adjusted archival scores retrospectively when reviewing previous IQ scores. These findings are consistent with the testimony in *Green v. Johnson* (2008) in which, out of 5,000 school-based IQ test reports between 1999–2001, only 6 mentioned the FE. None adjusted the obtained IQ scores.

Other Standards Authorities

The search for other IQ testing standards authorities led to the test manuals themselves because multiple authorities substantiate that the manual is the sine non qua for test administration and scoring.

We included current adult IQ tests fully meeting the criteria of the National Research Council (2002), instruments authorized by the Social Security Administration (SSA, 2006), measures identified from peer-reviewed published surveys of clinical practice patterns (Rabin, Barr, & Burton, 2005; Watkins, Campbell, Nieberding, & Hallmark, 1995), and those approved by the only two states that maintain lists of measures for capital mental retardation evaluations (Fla. Stat. § 921.137 [1], 2005; Virginia Department of Mental Health, Mental Retardation and Substance Abuse, 2005). Excluded were earlier versions of tests that psychologists might encounter in the evaluee's archives (e.g., the Wechsler Intelligence Scale for Children [3rd ed.] or the Stanford–Binet Intelligence Scales [4th ed.]) or tests constructed primarily for minors.

Six IQ tests met the inclusion criteria: the Wechsler Adult Intelligence Scale (3rd ed.; WAIS-III; Wechsler, 2002), the Stanford–Binet Intelligence Scales (5th ed.; Roid, 2003), the Kaufman Adolescent and Adult Intelligence Test (Kaufman & Kaufman, 1993), the Reynolds Intellectual Assessment Scales (Reynolds & Kamphaus, 2003), the Multidimensional Aptitude Battery (2nd ed.; Jackson, 2003), and the Woodcock–Johnson Test (3rd ed.; Mather & Woodcock, 2001). We examined each test manual for citations of Flynn's publications, references to the FE, and any specific recommendation for dealing with the increase in scores over time.

The *WAIS-III Technical Manual–Revised* (Wechsler, 2002) acknowledges "IQ-score inflation over time" and thus recommends that "norms for a test of intellectual functioning should be updated regularly" (Wechsler, 2002, p. 9). The WAIS-III publisher specifically rejects the practice of adjusting obtained scores: "Still, there is no scientific justification for adjusting data to fit theory. As the publisher of the Wechsler series of tests, Harcourt Assessment does not endorse the recommendation made by Flynn to adjust WAIS-III scores" (Weiss, 2007, p. 1).

The Stanford–Binet Intelligence Scales and the Kaufman Adolescent and Adult Intelligence Test manuals cite Flynn (1987) but make no specific recommendation for dealing with this statistical observation beyond the general admonition to follow the scoring rules strictly. The Reynolds Intellectual Assessment Scales, the Multidimensional Aptitude Battery, and the Woodcock–Johnson Test do not reference the FE, either conceptually or by name.

Several other sources of authority illuminate whether adjusting individual obtained IQ scores is the model accepted as correct by custom or consent. The SSA eligibility determination process is one of the largest testing programs in the United States. More than 1 million individuals currently receive SSA benefits under the mental retardation criteria.

In an effort to assess the adequacy of disability determinations, the SSA engaged the National Research Council to "evaluate the existing determination process in the context of state-of-the-art scientific knowledge and clinical practice" (National Research Council, 2002, p. 1). The large-scale effort by the study group produced numerous recommendations but did not include a specific proposal to adjust individual obtained IQ scores either in current testing or for archival assessments. Instead, the study group recommended that "tests should undergo normative update, restandardization, or revision at intervals corresponding to the time expected to produce one *SEM* of change" (National Research Council, 2002, p. 125).

The SSA Program Operations Manual System articulates the disability evaluation protocol for mental retardation (SSA, 2006). The agency's policy specifically bars its reviewing staff psychologists from adjusting current and archival IQ tests scores provided by the examining psychologist (SSA, 2006). To date, no appellate court has reversed or remanded a denial of an SSA entitlement claim because of a failure to adjust IQ scores on the basis of the FE.

The use of IQ testing for special education is another substantial public policy issue impacting a large population. As many as 5 million children receive special education services under the Individuals With Disabilities Education Improvement Act of 2004. This regulation does not reference the FE and does not set a standard for adjusting an individual's obtained scores or recalculating the mean score against which the obtained score should be assessed (Individuals With Disabilities Education Improvement Act, 2004, § 300.532).

Next, our search for a standard of practice examined contemporary textbooks published for practicing clinicians and graduate students. We queried APA Online PsycNET book records for 1984 through 2007, using the keyword *IQ test*. A leading psychology

textbook publisher and current graduate school assessment faculty also contributed to a list of relevant titles. Other titles were found in the IQ testing section of the library of a university with APA-approved training programs in clinical and counseling psychology. Other titles surfaced in research publications cited earlier.

Because our interest focused on practice standards, the search included textbooks only of an applied nature. The search yielded 14 textbooks published between 1999 and 2007 (Flanagan & Harrison, 2005; Gleitman, Fridlund, & Reisberg, 2003; Groth-Marnat, 2003; Kaplan & Saccuzzo, 2005; Kaufman & Lichtenberger, 1999, 2006; Kaufman, Lichtenberger, Fletcher-Janzen, & Kaufman, 2005; Lichtenberger & Kaufman, 2004; Myers, 2007; Prifitera, Saklofske, & Weiss, 2005; Sattler & Hoge, 2006; Tulsky, Saklofske, & Ricker, 2003; Urbina, 2004; Weiss, Saklofske, Prifitera, & Holdnack, 2006).We examined each textbook for the presence of Flynn in the author index, FE in the subject index, and specific recommendations for dealing with the FE when reporting scores.

Most (79%) contemporary applied textbooks cite Flynn's research and mention the FE by name or as a concept. In contrast to the claim in *Walker v. True* (2005), none recommend adjusting scores or recalculating norm means as generally accepted practice. Some specifically recommend following the test manual directions and give detailed instructions toward that end. Others simply advise that the norms should be updated periodically.

Ethical canons and related guidelines serve as a source of authority for practice standards. APA's "Ethical Principles of Psychologists and Code of Conduct" (APA, 2002) do not comment specifically on score adjustment apart from asserting that "psychologists administer, adapt, score, interpret, or use assessment techniques, interviews, tests, or instruments in a manner and for purposes that are appropriate in light of the research on or evidence of the usefulness and proper application of the techniques" (9.02a).

Standards for Educational and Psychological Testing (American Educational Research Association, APA, & National Council on Measurement in Education, 1999) provides criteria for testing practices and the effects of test use. Standards 5.1 and 5.2 require the test administrator to carefully follow the standardized procedures and score the measure according to the test manual without departing from the publisher's instructions. These standards make no reference to the FE, adjusting individual scores, or recalculating norm means separate and apart from the test manual.

Neither the "Specialty Guidelines for Forensic Psychologists" (Committee on Ethical Guidelines for Forensic Psychologists, 1991) nor the latest draft revisions for these guidelines (Committee on Ethical Guidelines for Forensic Psychologists, 2008) advocate diverting from test scoring manual instructions.

The APA has promulgated policy statements regarding psychological testing (APA, 1996; Joint Committee on Testing Practices, 1998), general service guidelines (Committee on Professional Practice and Standards, 2007; Committee on Professional Standards, 1987), practice area guidelines (APA, 1998, 2004; Committee on Professional Practice and Standards, 1998), and related qualification guidelines (APA, 2001). All are pertinent, in part or whole, to professional responsibility when using IQ tests for a wide range of purposes. All are silent with respect to the FE. None establish a standard for adjusting obtained scores or for departing from test manual instructions.

Statutory and Case Law Authority

Duvall and Morris (2006) surveyed the statutes relevant to death penalty evaluations in the United States. Of the 38 death penalty states, none has a statute that mandates adjusting IQ scores on the basis of the FE. Case law in Tennessee (*Howell v. Tennessee,* 2004) and Kentucky (*Bowling v. Kentucky,* 2005) specifies that adjusting obtained scores on the basis of the FE is not sufficiently scientific. The latter court rejected factoring in the impact of the FE, finding that "the scientific community does not agree on the cause of this phenomenon" (*Bowling v. Kentucky,* 2005, p. 37). In *Green v. Johnson* (2008), the Fourth Circuit Court of Appeals observed for both the FE and the standard error of measurement that "neither *Atkins* nor Virginia law appears to require expressly that these theories be accounted for in determining mental retardation status" (p. 8).

Although appellate case law calls for consideration of the FE when not procedurally barred, there is no judicial consensus that adjusting obtained scores or recalculating norm means is generally accepted in the field. Some appellate courts have ruled that a trial court must consider evidence of the FE and determine the persuasiveness of the evidence (*Walker v. True,* 2005). However, this survey found no instance in which an appellate court ruled that the FE is compelling or controlling as a matter of law.

Conclusions and Implications for Practice

Three conclusions emerge. First and foremost, adjusting obtained scores and recalculating norm means on the basis of the FE do not represent the convention and custom in psychology. Adjusting obtained IQ scores for this purpose is not the standard of practice. Second, recalculating an individual's actual data likely violates standardization procedures and departs from training practices, prevailing canons, guidelines, most treatises, and test instructional manuals. In addition, the prevailing consensus calls for publishers to update norms periodically. Third, when choosing IQ tests or reviewing archival test data, psychologists should carefully consider potential compromises to validity and the differential impact of such compromises in light of race, culture, age, gender, and the weighting of cognitive demands of the instrument. Commenting on these issues in the report narrative is appropriate, but adjusting the numerical scores is not. The practitioner should heed the practice standard to use the most current version of a test.

The current accepted convention does not support subtracting IQ points in a way that departs from the requirements of the test manual. "Evaluators must also be aware that there is no agreed-upon method for how diagnostic conclusions should be influenced by the Flynn effect" (Young et al., 2007, p. 176). Psychologists cannot conclude that adjusting scores is the generally accepted practice in evaluations for special education, parental rights termination, disability, or any other purpose.

An accurate score on an IQ test can make a meaningful difference, and the descriptive label the psychologist applies to it can also make a difference (Guilmette, Hagan, & Giuliano, 2008). Highly skilled and conscientiously committed psychologists may find that these critical medico-legal evaluations stir significant personal and ethical dilemmas. Those who thoughtfully reflect on the clinical and forensic issues as well as their qualifications and experience and elect to decline or accept these referrals are to be

commended for their professional posture. Those who decide to undertake these forensic evaluations should proceed cautiously and continuously educate themselves about developments in the law, ethics, practice standards, and science.

References

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

American Psychological Association. (1996). Statement on the disclosure of test data. *American Psychologist, 51,* 644–648.

American Psychological Association. (1998). Guidelines for the evaluation of dementia and age-related cognitive decline. *American Psychologist, 53,* 1298–1303.

American Psychological Association. (2001). APA's guidelines for test user qualifications: An executive summary. *American Psychologist, 56,* 1099–1113.

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

American Psychological Association. (2004). Guidelines for psychological practice with older adults. *American Psychologist, 59,* 236–260.

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

Berry v. Mississippi, 544 U.S. 950 (2005) [Flynn affidavit].

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

Bowling v. Kentucky, 163 S.W.3d. 361 (Kentucky 2005).

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

Committee on Ethical Guidelines for Forensic Psychologists. (2008). *Specialty guidelines for forensic psychology: Third official draft.* Retrieved March 19, 2008, from http://ap-ls.org/links/22808sgfp.pdf

Committee on Professional Practice and Standards. (1998). *Guidelines for psychological evaluations in child protection matters.* Washington, DC: American Psychological Association.

Committee on Professional Practice and Standards. (2007). *Record keeping guidelines.* Washington, DC: American Psychological Association.

Committee on Professional Standards. (1987). *General guidelines for providers of psychological services.* Washington, DC: American Psychological Association.

Commonwealth v. Prieto, Fairfax (VA) Cir. Ct., FE 2005-1764 (March 5, 2007).

Daley, T. C., Whaley, S. E., Sigman, M. D., Espinosa, M. P., & Neumann, C. (2003). IQ on the rise—The Flynn effect in rural Kenyan children. *Psychological Science, 14,* 215–219.

Duvall, J. C., & Morris, R. J. (2006). Assessing mental retardation in death penalty cases: Critical issues for psychology and psychological practice. *Professional Psychology: Research and Practice, 37,* 658–665.

Flanagan, D. P., & Harrison, P. L. (2005). *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed.). New York: Guilford.

Fla. Stat. (2005). § 921.137[1].

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

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 1972–1985: How to compensate for obsolete norms. *Perceptual and Motor Skills, 86,* 1231–1239.

Flynn, J. R. (1999). Searching for justice: The discovery of IQ gains over time. *American Psychologist, 54,* 5–20.

Flynn, J. R. (2000). The hidden history of IQ and special education: Can the problems 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). *What is intelligence? Beyond the Flynn effect.* New York: Cambridge University Press.

Gleitman, H., Fridlund, A. J., & Reisberg, D. (2003). *Psychology* (6th ed). New York: Norton.

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 Newsletter, 31,* 3–7.

Green v. Johnson, WL 352028 (4th Cir. 2008).

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

Guilmette, T., Hagan, L. D., & Giuliano, A. J. (2008). Assigning qualitative descriptions to test scores in neuropsychology: Forensic implications. *Clinical Neuropsychologist, 22,* 122–139.

Howell v. Tennessee, 151 S.W.3d 450 (Tenn. 2004).

Individuals With Disabilities Education Improvement Act of 2004, 20 U.S.C. § 1400 *et seq.*

Jackson, D. N. (2003). *Multidimensional Aptitude Battery: Manual* (2nd ed.). Port Huron, MI: Sigma Assessment Systems.

Joint Committee on Testing Practices. (1998). *Rights and responsibilities of test takers: Guidelines and expectations.* Washington, DC: American Psychological Association.

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.

Kaplan, R., & Saccuzzo, D. (2005). *Psychological testing: Principles, applications, and issues* (6th ed.). Belmont, CA: Thomson-Wadsworth.

Kaufman, A. S., & Kaufman, N. L. (1993). *Kaufman Adolescent and Adult Intelligence Test: Manual.* Circle Pines, MN: American Guidance Services.

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

Kaufman, A. S., & Lichtenberger, E. O. (2006). *Assessing adolescent and adult intelligence* (3rd ed.). New York: Wiley.

Kaufman, A. S., Lichtenberger, E. O., Fletcher-Janzen, E., & Kaufman, N. L. (2005). *Essentials of KABC-II assessment.* New York: Wiley.

Lacritz, L. H., & Cullum, C. M. (2003). The WAIS-III and WMS-III: Practical issues and frequently asked questions. In D. S. Tulsky, D. H. Saklofske, G. J. Chelune, R. K. Heaton, R. J. Ivnik, R. Bornstein, et al. (Eds.), *Clinical interpretation of the WAIS-III and WMS-III* (pp. 491–532). Boston: Elsevier.

The latest thinking on intelligence: Interview with James Flynn. (2007, June). *The Psychologist, 20,* 356–357.

Lichtenberger, E. O., & Kaufman, A. S. (2004). *Essentials of WPPSI-III assessment.* New York: Wiley.

Mather, N., & Woodcock, R. W. (2001). *Woodcock–Johnson 3rd Ed.: Examiner's manual.* Itasca, IL: Riverside.

Moore, R. B. (2006, fall). Modification of individual's IQ scores is not accepted professional practice. *Psychology in Mental Retardation and Developmental Disabilities, 31,* 11–12.

Myers, D. G. (2007). *Psychology* (8th ed.). New York: Worth.

National Research Council. (2002). *Mental retardation: Determining eligibility for social security benefits.* Washington, DC: National Academy Press.

Olley, J. G., Greenspan, S., & Switzky, H. (2006, winter). Division 33 Ad Hoc Committee on Mental Retardation and the Death Penalty. *Psychology in Mental Retardation and Developmental Disabilities, 31,* 11–13.

People v. Superior Court (Vidal), 129 Cal. App. 4th 434, 28 Cal Rptr. 3d 529 (5th Ct. App. 2005), *vacated and later proceedings at* People v. S. C., 2005 Cal. LEXIS 13290 (Cal. Nov. 17, 2005).

Prifitera, A., Saklofske, D. H., & Weiss, L. D. (2005). *WISC-IV clinical use and interpretation.* New York: Academic Press.

Rabin, L. A., Barr, W. B., & Burton, L. A. (2005). Assessment practices of clinical neuropsychologists in the United States and Canada: A survey of INS, NAN, and APA Division 40 members. *Archives of Clinical Neuropsychology, 20,* 33–65.

Reynolds, C. R., & Kamphaus, R. W. (2003). *Reynolds Intellectual Assessment Scales and the Reynolds Intellectual Screening Test: Professional manual.* Lutz, FL: Psychological Assessment Resources.

Roid, G. H. (2003). *Stanford–Binet Intelligence Scales, 5th Edition: Technical manual.* Itasca, IL: Riverside.

Sattler, J. M., & Hoge, R. D. (2006). *Assessment of children: Behavioral, social and clinical foundations* (5th ed.). LaMesa, CA: Sattler.

Shayer, M., Ginsburg, D., & Coe, R. (2007). Thirty years on—A large anti-Flynn effect? The Piagetian test Volume & Heaviness norms 1975–2003. *British Journal of Educational Psychology, 77,* 25–41.

Social Security Administration. (2006). *Disability evaluation under Social Security: DI 22510.021. Consultative examination report content guidelines—Mental disorders.* Washington, DC: Author.

State v. Keel, 90 CRS 8033 (Sup. Ct. Edgecombe Co., N.C.) (unpub.), *cert. denied,* 357 N.C. 465, 586 S.E.2d 462 (N.C. 2003).

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.

Urbina, S. (2004). *Essentials of psychological testing.* New York: Wiley.

Virginia Department of Mental Health, Mental Retardation and Substance Abuse. (2005). *List of standardized measures of intellectual functioning.* Retrieved July 25, 2007, from http://www.dmhmrsas.virginia.gov/documents/ofo-standardizedmeasures.pdf

Walker v. True, 399 F.3d 315, *after remand,* 401 F.3d 574 (4th Cir. 2005).

Walton v. Johnson, 440 F.3d 160 (4th Cir.) (*en banc*), *cert. denied.* 126 S.Ct. 2377 (2006).

Watkins, C. E., Campbell, V. L., Nieberding, R., & Hallmark, R. (1995). Contemporary practice of psychological assessment by clinical psychologists. *Professional Psychology: Research and Practice, 26,* 54–60.

Wechsler, D. (2002). *Technical manual (updated) for the Wechsler Adult Intelligence Scale, 3rd ed. and Wechsler Memory Scale, 3rd ed.* San Antonio: Psychological Corporation.

Weiss, L. G. (2007). *WAIS-III technical report: Response to Flynn.* Retrieved March 14, 2007, from http://harcourtassessment.com/NR/rdonlyres/98BBF5D2–F0E8–4DF6–87E2–51D0CD6EE98C/0/WAISIII_TR_lr.pdf

Weiss, L. G., Saklofske, D. H., Prifitera, A., & Holdnack, J. A. (2006). *WISC-IV: Advanced clinical interpretation.* New York: Academic Press.

Young, B., Boccaccini, 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.

Zhou, X., & Zhu, J. (2007, August). *Peeking inside the "blackbox" of Flynn effect: Evidence from three Wechsler instruments.* Poster session presented at the annual meeting of the American Psychological Association, San Francisco, CA.

Received October 26, 2007
Revision received April 28, 2008
Accepted May 6, 2008 ■

Journal of Psychoeducational Assessment
28(5) 474–476

DOI: 10.1177/0734282913733343
http://jpa.sagepub.com
SAGE

IQ Scores Should Not Be Adjusted for the Flynn Effect in Capital Punishment Cases

Leigh D. Hagan[1], Eric Y. Drogin[2], and Thomas J. Guilmette[3]

Abstract

Atkins v. Virginia (2002) dramatically raised the stakes for mental retardation in capital punishment cases, but neither defined this condition nor imposed uniform standards for its assessment. The basic premise that mean IQ scores shift over time enjoys wide recognition, but its application—including the appropriateness of characterizing it in terms of an allegedly predictable "Flynn effect"—is frequently debated in the course of death penalty litigation. The scientifically and ethically sound approach to this issue is to report IQ scores as obtained and be prepared to address those factors that might affect their reliability. Altering the IQ scores themselves is insufficiently supported by professional literature, legal authority, or prevailing standards of practice.

Keywords

practice standards, Flynn effect, IQ, intelligence testing, death penalty

In *Atkins v. Virginia* (2002), the Supreme Court of the United States banned the execution of persons with mental retardation (MR), but it neither defined MR nor specified how to evaluate it. Some experts maintain that the basic premise of the Flynn Effect (FE)—that mean IQ scores increase over time (Flynn, 1987)—is critical to the accurate identification and depiction of MR in capital murder cases. We do not seek to impugn or debunk the FE or its relevance to these cases; rather, our goal is to insist that those inclined to invoke this theory do so in a valid, responsible, and ethical manner. We conclude that the practice of altering an obtained IQ score based on the FE is insufficiently supported by scholarly literature or legal authority.

The FE is typically conveyed as an annual increase of 0.3 points per year, resulting in an inflation of scores between the time of test development and the test's eventual clinical use with a particular examinee (Flynn, 1987). Decades of FE research and testimony, however, depict the *amount* of this shift as a moving target. For example, Flynn (1998) once identified the annual

[1]Virginia Commonwealth University, Chesterfield, VA, USA
[2]Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA
[3]Providence College, Warren Alpert Medical School of Brown University, Providence, RI, USA

Corresponding Author:
Leigh D. Hagan, P. O. Box 350, Chesterfield, VA 23832, USA
Email: lhagan@leighhagan.com

shift as 0.25 rather than 0.30, but later testified in *Ex Parte Eric Dewayne Cathey* (2010) that 0.29 would be appropriate. Schalock et al. (2010) have called for an annual adjustment of 0.33.

Spitz (1989) found the FE to vary depending on the examinee's obtained range of intellectual functioning. Kanaya, Scullin and Ceci (2003) and the project team at PsychCorp/Pearson (Wechsler, 2008) also identified noteworthy variability across the normal distribution. Zhou, Zhu, and Weiss (2010) analyzed Performance IQs and confirmed that the FE varies by ability level, age group, and specific intelligence test. In fact, whereas most FE studies report gradual IQ score increases over time, some have found stagnation and some noted a reverse **(**Flynn, 2000; Shayer, Ginsburg, & Coe, 2007; Teasdale & Owen, 2000**)**.

Flynn (2006) characterized the notion that the FE cannot be particularized to an individual as a prosecutor's "senseless mantra," asserting that FE gains "render test norms obsolete and inflate the IQ of every individual being scored against obsolete norms" (p. 186). An all-inclusive declaration about "every individual" does not, however, adequately acknowledge the probabilistic nature of group data and potential inconsistency when applied to individuals.

When it comes to analyzing and commenting on the accuracy and applicability of a particular IQ test result, due consideration should be given to other well-documented influences on score variability. The project team at PsychCorp/Pearson (Wechsler, 2008) substantially revised each iteration of its intelligence scales by altering or eliminating subtests, increasing the number of permissible cues, changing the scoring for some subtests, reordering subtest presentation, and other changes. These modifications substantially complicate comparisons across different measures, as do such additional notions as the standard error of measurement (SEM), test–retest phenomena, and variations in examinee effort.

A national survey of American Board of Professional Psychology school psychologists and training directors of American Psychological Association–accredited clinical, counseling, and school psychology doctoral programs showed that most report or teach the practice of reporting obtained scores and—consistent with the dictates of test manuals—do not train future psychologists to alter IQ scores due to the FE (Hagan, Drogin, & Guilmette, 2008). Although several appellate courts have remanded capital murder cases to the trial court for an evidentiary hearing to consider the FE, at this time no appellate court has published a ruling that subtracting IQ points or adjusting the mean based on the FE is a generally accepted practice. None of the 38 states allowing for capital punishment has a statute mandating reduction of a capital defendant's IQ scores based on the FE (Duvall & Morris, 2006).

Altering obtained IQ scores based on the FE does not comport with the standard of forensic psychological practice, and there exists no legal mandate to make such adjustments. Psychologists serve an important function in capital punishment cases when they identify data limitations that may be attributable to the FE or any other error source. If an obtained score is considered to be invalid and if the "true" score is believed to be higher or lower within an estimated range, psychologists are justified in sharing this perspective in narrative form, but the current state of psychological science—particularly in light of the established variability of individual cases—does not support devising some other score based on the FE and then substituting that score for the one obtained.

Declaration of Conflicting Interests

The author(s) declared no conflicts of interest with respect to the authorship and/or publication of this article.

Funding

The author(s) received no financial support for the research and/or authorship of this article.

References

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

Duvall, J. C., & Morris, R. J. (2006). Assessing mental retardation in death penalty cases: Critical issues for psychology and psychological practice. *Professional Psychology: Research and Practice, 37*, 658-665.

Ex Parte Eric Dewayne Cathey. Cause No. 713189, 176th District Court, Harris County, TX. January 25, 2010.

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 & WISC-III IQ gains in the United States from 1972 to 1995: How to compensate for obsolete norms. *Perceptual & Motor Skills, 86*, 1231-1239.

Flynn, J. R. (2000). The hidden history of IQ and special education: Can the problems 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.

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.

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

Shayer, M., Ginsburg, D., & Coe, R. (2007). Thirty years on—A large anti-Flynn Effect? The Piagetian test volume and heaviness norms 1975-2003. *British Journal of Educational Psychology, 77*, 25-41.

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

Teasdale, T., & Owen, D. (2000). Forty-year secular trends in cognitive abilities. *Intelligence, 28*, 115-120.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale, fourth edition: Technical and interpretive manual.* San Antonio, TX: Pearson.

Zhou, X., Zhu, J., & Weiss, L. A. (2010). Peeking inside the "black box" of the Flynn effect: Evidence from three Wechsler instruments. *Journal of Psychoeducational Assessment, 28*, 399-411.

Journal of Psychoeducational Assessment
28(5) 469–473

DOI: 10.1177/0734282910373341
http://jpa.sagepub.com
SAGE

IQ Scores Should Be Corrected for the Flynn Effect in High-Stakes Decisions

Jack M. Fletcher[1], Karla K. Stuebing[1], and Lisa C. Hughes[1]

Abstract

IQ test scores should be corrected for high stakes decisions that employ these assessments, including capital offense cases. If scores are not corrected, then diagnostic standards must change with each generation. Arguments against corrections, based on standards of practice, information present and absent in test manuals, and related issues, ignore expert consensus about the assessment of intellectual disabilities and the acceptance of the Flynn effect in the field. Most psychometric concerns about correction are based on validity studies with small subgroups and do not reflect sufficient effort to estimate the precision of the Flynn estimate. We computed a confidence interval for the Wechsler PIQ across four validity studies that shows a SEM of about 1 around a mean of about 3 points per decade. A meta-analytic weighted mean of the 14 studies in Flynn (2009) is 2.80 (2.50, 3.09), close to Flynn's (2009) unweighted average (2.99). More psychometric research would be helpful, but this level of precision supports the Flynn adjustment of 3 points per decade.

Keywords

IQ, intellectual disability, Flynn effect, Atkins hearings

IQ test scores should be corrected for high-stakes decisions in which a test with older norms is invoked as evidentiary support in the decision-making process. This could include not only Atkins cases involving capital offenses and the death penalty but also intellectual disability (ID) decisions involving social security eligibility or special education where eligibility hinges on a specific score or range of scores. In all these contexts, the person may have previous IQ test scores that are higher than current scores, which may be reconciled by taking into account norms obsolescence.

In Atkins cases as well as other high-stakes assessments, the offender often has multiple IQ scores obtained over a long period of time. Some offenders may have been administered older versions of tests with norms well over 10 years of age, rendering them obsolete and yielding

[1]University of Houston, Houston, TX, USA

Corresponding Author:
Jack M. Fletcher, Department of Psychology, University of Houston Texas Medical Center Annex, 2151 W. Holcombe Boulevard, Suite 222, Houston, TX 77204-5053, USA
Email: jackfletcher@uh.edu

inaccurate estimates of IQ (Flynn, 2009). To illustrate, in one case in which the senior author consulted, the offender had WAIS-III (Wechsler Adult Intelligence Scale–Third Edition) scores of 68 and 71, 3 years apart as an adult. As a child, the offender obtained a WISC (Wechsler Intelligence Scale for Children) score of 79 in 1973, 25 years after the normative sample was collected. A correction for the Flynn effect (FE) of 0.3 per year would be 0.3 × 25 years = 7.5, or 71.5, aligning closely with the WAIS-III assessments. Should an offender be executed because the psychologist who gave the WISC failed to write a note indicating that the IQ score may be an overestimate because of norms obsolescence?

Correcting an IQ score is not a violation of test administration. Rather, it is selecting an appropriate normative comparison (Gresham, 2009). We would not expect pediatricians to use a height/weight chart from another country or century to assess a child's percentile rank in height or weight; if they did, we would expect corrections so that the percentile reflects the current, national distribution. Correcting an IQ score is a simple procedure that avoids having to change standards. Thus, if 15-year-old IQ norms are used, either the score itself must be corrected by about 4.5 points (0.3 × 15 years = 4.5) or the cut-point for ID needs to be corrected to 74.5 because the mean IQ of a contemporary sample using the old norms would be 104.5.

Some argue that correcting for norms obsolescence is not a standard of practice (Hagan, Drogin, & Guilmette, 2008; 2010). However, standards of practice are set by consensus reports written by experts. The most prominent guidelines for the assessment of ID represent the 11 editions of the manual for diagnosis by the American Association of Intellectual and Developmental Disabilities (Schalock et al., 2010), not cited by Hagan et al. (2008). Since 2002, this manual has explicitly recommended correcting IQ scores for norms obsolescence, with other researchers agreeing (e.g., Gresham, 2009; Kanaya & Ceci, 2007; Widaman, 2007).

Other objections to correcting for norms obsolescence confuse issues related to why the FE occurs with whether it occurs; its existence is widely accepted, but the cause is disputed (see Flynn, 2010; Kaufman, 2010). There is also confusion involving Flynn's assertion that the WAIS-III norms are problematic (e.g., Flynn, 2009). The publisher's post on this issue (Weiss, 2008) addressed Flynn's claim that there were problems with the norming of the WAIS-III, but has been misinterpreted as indicating that the correction for norms obsolescence was under dispute (Hagan et al., 2008), which is not the case (Zhou, Zhu, & Weiss, 2010). Some suggest that the standardization and validity samples are different and that group data should not be used to correct individual scores (Zhu & Tulsky, 1999). However, individual scores are not being adjusted; rather, the validity studies are used as a basis for selecting an appropriate normative comparison group.

The major questions should involve the magnitude of the effect and its constancy across age and levels of IQ (Tanaka & Ceci, 2007; Zhou et al., 2010). As Widaman (2007) suggested, much of the variation in estimates of the effect is because of measurement error, especially when small samples across different age and IQ levels are used. This variation is important to understand, and it is surprising that more effort has not been expended toward evaluating the precision of the correction.

We estimated 95% confidence intervals (CIs) for the four comparisons of PIQ (Performance IQ) in Zhou et al. (2010) using the standard deviations for each comparison kindly provided by Dr. Xiaobin Zhou (Table 1). The CIs were computed by estimating the standard error of the mean (SEM) of average change and multiplying by ±1.96 (the critical *z* value). The SEM for matched pairs is the *SD* of the difference divided by the square root of *N*. To create the CI, we used a standard formula [CI_{95} = mean difference ± $z_{.05/2}$ (SEM)]. As Table 1 shows, the confidence intervals do not include 0 and extend approximately 1 point (0.1 per year) on either side of the mean difference of about 3 per decade (0.29-0.31 per year). A simple rubric would be 3 ± 1. An adjustment for Full-Scale IQ (FSIQ) would be similar because it is highly correlated with PIQ. Because the FSIQ is higher in reliability, the CIs may smaller.

Table 1. Confidence Intervals for PIQ Across Four Wechsler Tests

Tests	Mean Change Per Year	*SD* Change Per Year	*N*	Years Between Norming	*SE* or $SD / \sqrt{N}$	*SE* Times 1.96	Lower CI Mean Minus $SE \times 1.96$	Upper CI Mean Plus $SE \times 1.96$	Lower CI in Points Per Decade	Upper CI in Points Per Decade
WPPSI-R/III	0.24	0.86	174	13	0.07	0.13	0.11	0.37	1.12	3.68
WISC-III/IV	0.29	0.96	239	12	0.06	0.12	0.17	0.41	1.68	4.12
WAIS-R/III	0.29	0.61	191	16	0.04	0.09	0.20	0.38	2.03	3.77
WAIS-III/IV	0.31	0.81	240	11	0.05	0.10	0.21	0.41	2.08	4.12

Note. PIQ = Performance IQ; *SD* = standard deviation; *SE* = standard error; CI = confidence interval; WPPSI-R = Wechsler Preschool and Primary Scale of Intelligence–Revised; WISC = Wechsler Intelligence Scale for Children; WAIS-R = Wechsler Adult Intelligence Scale–Revised.

Table 2. Weighted Mean Effects, Confidence Intervals, and Tests of Homogeneity

Newer Tests	Older Tests	Difference Years	*N*	Mean Difference	Difference Per Decade	Deviation Squared Model 1	Deviation Squared Model 2	Deviation Squared Model 3
SB-5	WAIS-III	6	87	5.50	9.17	43.61		
SB-4	WAIS-R	7	47	3.42	4.89	5.07	4.29	4.29
WISC-IV	WAIS-III	6.75	198	3.10	4.59	11.85	9.75	
SB-5	WISC-III	12	66	5.00	4.17	1.72	1.33	1.33
WISC-IV	WISC-III	12.75	244	4.23	3.32	1.14	0.53	0.53
WISC-III	WISC-R	17	206	5.30	3.12	0.43	0.10	0.10
SB-4	WISC-R	13	205	2.95	2.27	0.74	1.29	1.29
SB-5	SB-4	16	104	2.77	1.73	2.62	3.50	3.50
WAIS-III	WAIS-R	17	192	4.20	2.47	0.64	1.47	
SB-4	SB-LM	13	139	2.16	1.66	2.09	2.75	2.75
WAIS-R	WISC-R	6	80	0.90	1.50	2.48	3.17	3.17
WAIS-III	WISC-III	6	184	–0.70	–1.17	53.48		0.00
WAIS-IV	WAIS-III	11	240	3.37	3.06	0.64	0.09	0.09
WAIS-IV	WISC-IV	4.25	157	1.20	2.82	0.00	0.05	0.05
Mean effect						2.80	2.96	2.86
Q						126.52***	28.32**	17.10**

Note. SB = Stanford–Binet Intelligence Scale; WAIS = Wechsler Adult Intelligence Scale; WISC; Wechsler Intelligence Scale for Children.
**$p < .003$. ***$p < .0001$.

Table 2 uses the 14 studies in Flynn (2009) to compute the meta-analytic mean, showing an inverse variance weighted mean effect (Lipsey & Wilson, 2001) per decade of 2.80 (2.50, 3.09), close to Flynn's unweighted average. We tested the distribution of effects for heterogeneity using the *Q* statistic (which is distributed as a chi-square with $k - 1$ degrees of freedom, where k equals the number of studies), and found that the 14 effects were more variable than would be expected because of sampling error alone, $Q_{(13)} = 126.52$, $p < .0001$. Although the CI is small, significant heterogeneity potentially limits the usefulness of the mean effect because of averaging dissimilar effects. Inspection of the contribution of each effect to the *Q* statistic (Deviation Squared Model 1 in Table 1) revealed two outliers, one very large and one very small, both of which involved the WAIS-III. After removing these two outliers, the mean effect per decade was 2.96 (2.65-3.27), with $Q_{(11)} = 28.33$, $p < .003$. Given the questions raised about the normative sample for the WAIS-III (Flynn, 2009), we removed the other two WAIS-III comparisons and found a mean

effect of 2.86 (2.5-3.22) and $Q_{(9)}$=17.1, $p < .047$. Thus, the sources of heterogeneity can be identified. We do not view this finding as supporting Flynn's claim that the WAIS-III norms are problematic. Rather, more research with additional samples and perhaps the inclusion of other tests may enhance understanding of factors responsible for the variability across studies and make possible more precise estimates of the effect of norms obsolescence.

These two approaches to estimating the mean and the precision of the effect support Flynn's aggregated estimate of the magnitude of norms obsolescence and are sufficiently precise to justify corrections for high-stakes decisions. There is variability across studies, and age/ability level, but this is true for any subject matter. The estimate of 3 ± 1 is similar to the estimates for the conversion of WAIS-III and WAIS-IV scores for the middle of the distribution (where the sample size is larger) in table 5.6 of the WAIS-IV technical manual.

The administration/technical manuals' silence over the FE has been interpreted in Atkins cases as evidence that scores should not be corrected. Clearly publishers have acknowledged the FE by renorming tests more frequently and providing validity studies and conversion tables. A publisher should not be expected to address every use of the test. The WAIS-IV manual, for example, provides no guidance on the diagnosis of ID. However, Weiss (2008) is commonly invoked as denying that the FE exists (Hagan et al., 2008) when it actually addresses the adequacy of the WAIS-III norms. In one Atkins hearing, an email from the technical assistance hotline of a publisher was introduced in response to a question about the FE from a testifying psychologist. The email indicated that the publisher did not recommend correcting scores. Telephone calls and emails requesting clarification from the publisher elicited no response and the judge cited the email in ruling against the offender.

Publishers may need to do more by providing data like that in Tables 1 and 2 (and studies like Zhou et al., 2010) and by indicating explicitly that when outdated norms are used, corrections will be necessary to appropriately scale the scores. This would facilitate adoption of practices recommended by the American Association of Intellectual and Developmental Disabilities into the different venues where IQ scores are used for high-stakes decision making. IQ scores based on obsolete norms should be corrected and can be estimated with reasonable precision in high-stakes decisions, including capital offense cases. There is no evidence that Flynn's correction overestimates IQ at the lower end of the distribution (Zhou et al., 2010).

Summary and Conclusions

IQ test scores should be corrected for any high-stakes decision that employ these assessments, including capital offense cases. If scores are not corrected, then diagnostic standards must change with each generation. Arguments against correction ignore expert consensus about the assessment of intellectual disabilities and do not take into account the wide acceptance of the FE. More research on the precision of the estimate would be helpful, but the level of precision we reported of a mean of about 3 and a SEM of about 1 supports the correction and is consistent with the Flynn correction of 3 points per decade.

Declaration of Conflicting Interests

The author(s) declared no conflicts of interest with respect to the authorship and/or publication of this article.

Funding

The author(s) received no financial support for the research and/or authorship of this article.

References

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.

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

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.

Hagan, L. D., Drogin, E. Y., & Guilmette, T. J. (2010). IQ scores should not be adjusted for the Flynn effect in capital punishment cases. *Journal of Psychoeducational Assessment, 28*, 474-476.

Kanaya, T., & Ceci, S. J. (2007). Mental retardation diagnosis and the Flynn effect: General intelligence, adaptive behavior, and context. *Child Development Perspectives, 1*, 62-63.

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.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

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

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

Widaman, K. (2007). Stalking the roving IQ score cutoff: A commentary on Kanaya and Ceci. *Child Development Perspectives, 1*, 57-59.

Zhou, X., Zhu, J., & Weiss, L. G. (2010). Peeking inside the "black box" of the Flynn effect: Evidence from three Wechsler instruments. *Journal of Psychoeducational Assessment, 28*, 399-411.

Zhu, J., & Tulsky, D. S. (1999). Can IQ gain be accurately quantified by a simple difference formula? *Perceptual and Motor Skills, 88*, 1255-1260.

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 ■

Journal of Psychoeducational Assessment
28(5) 477–481

DOI: 10.1177/0734282913733348
http://jpa.sagepub.com
SAGE

Failure to Apply the Flynn Correction in Death Penalty Litigation: Standard Practice of Today Maybe, but Certainly Malpractice of Tomorrow

Cecil R. Reynolds[1], John Niland[2], John E. Wright[3], and Michal Rosenn[4]

Abstract

The Flynn Effect is a well documented phenomenon demonstrating score increases on IQ measures over time that average about 0.3 points per year. Normative adjustments to scores derived from IQ measures normed more than a year or so prior to the time of testing an individual have become controversial in several settings but especially so in matters of death penalty litigation. Here we make the argument that if the Flynn Effect is real, then a Flynn Correction should be applied to obtained IQs in order to obtain the most accurate estimate of IQ possible. To fail to provide the most accurate estimate possible in matters that are truly life and death decisions seems wholly indefensible.

Keywords

Intelligence, Flynn effect, death penalty, forensic psychology

Since the Supreme Court's decision in *Atkins v. Virginia* (536 US 304, 122 S. CT 2242, 2002) that the execution of the mentally retarded violates the Eighth Amendment's prohibition against cruel and unusual punishment, the importance of understanding and assessing mental retardation in criminal defendants has become critical, indeed a true matter of life and death, in capital felony cases. Determining whether a defendant's intellectual functioning is severely limited is essential to a judgment as to whether that individual is able to act with the level of moral culpability that merits particular forms of punishment. As the best measures of intellectual functioning, IQ tests are regarded as one of the primary indicia of mental retardation by both clinicians and courts. The consensus among mental health professionals is that an IQ of 70 to 75 or less on

[1]Texas A&M University, College Station, TX, USA
[2]Texas Defender Service, Austin, TX, USA
[3]Huntsville, TX, USA
[4]Harvard University, Cambridge, MA, USA

Corresponding Author:
Cecil R. Reynolds, 101 Reynolds Court, Bastrop, TX 78602, USA
Email: crrh@earthlink.net

a standardized, individually administered IQ test satisfies the IQ prong of the diagnostic criteria for mental retardation (e.g., see Flynn, 2006, 2007a; Reynolds, Price, & Niland, 2003, as well as various court cases cited in these articles).

IQ tests are periodically revised and renormed to keep the content appropriate to current cultural contexts, ensure the representativeness of the normative or reference group (characteristics of the target population are constantly changing), and to maintain an average score of 100. The findings associated with these periodic revisions led researchers to observe that scores on standardized measures of intelligence have steadily risen over the past century, a phenomenon termed the Flynn Effect (FE; after James Flynn, the man who first documented these changes carefully and comprehensively, e.g., Flynn, 1984). Among the various explanations offered for the effect, the predominant explanation—and the one adopted by Flynn—is that environmental changes relating to modernization have increased people's ability to manipulate abstract concepts, a skill that is heavily emphasized in IQ tests. However, the reason for the FE is controversial, as can easily be seen in other articles in this issue, but the existence of the effect has no significant scholarly challenges of which we are aware. The FE, whatever its cause, is as real as virtually any effect can be in the social sciences. Studies have observed an increase of 0.3 points per year in average IQs; thus, for a test score to reflect accurately the examinee's intelligence, 0.3 points must be subtracted for each year since the test was standardized (Flynn, 2006, 2007a, 2007b). Since the FE's increased scientific acceptance in the 1990s, it has become one of the reasons why IQ tests have been revised and renormed more frequently than in the past, typically occurring on a 10- to 11-year schedule now as opposed to a 20-year or more schedule in the past. Even so, the FE is observable in the years between revisions, and is certainly relevant where outdated test versions are used—especially where even 2 or 3 IQ points may determine whether a defendant is allowed to live or is killed.

Because of the central role IQ tests play in determining an individual's level of mental retardation, and because of the importance of mental retardation in determining a defendant's eligibility to be killed by the State, it is imperative that the FE, if it is real, be taken into account in capital cases. IQ ranges that indicate mental retardation are determined relative to the average score (which has been set by convention, albeit arbitrarily, at 100). The so-called average score is derived from a reference group, which is a snapshot of the population at one particular point in time. The determination of the intelligence component of the diagnosis of mental retardation (we do recognize that the actual diagnosis is far more complex than looking at an IQ—but the IQ is a crucial component, and we deal only with it here) should be based on the person's standing relative to the target population at the time the person was actually tested, not the target population when the test was normed. Because it is at this time a practical impossibility to renorm tests annually to maintain a more appropriate reference group, to the extent corrections are available and valid, they should be applied to obtained scores so the most accurate estimate of standing possible is obtained. To do less is to do wrong—what possible justification could there be for issuing estimates of general intelligence in a death penalty case that are less than the most accurate estimates obtainable?

The way in which the Flynn correction applies to an individual is illustrated by the following scenario: a person taking the same version of the same IQ test 10 years apart will, on average, experience a 3-point increase in his or her score over that time—not because of any actual increase in intellectual functioning, but because of latent social changes that manifest themselves in the test. Therefore, a 3-point correction downward of the obtained IQ is required to provide the most accurate estimate of intellectual functioning relevant to today's population. Without this correction, whether a criminal defendant is deemed mentally retarded and thus eligible for the death penalty can thus turn on when the IQ measure chosen was standardized (e. g., see Ceci, Scullin, & Kanaya, 2003, and Flynn, 2006). If there remains any doubt that we must provide the

most accurate IQ estimates we can, in all cases, but especially matters of death, we can take guidance from the U.S. Supreme Court, the ultimate arbiter of legal issues in the United States, whose members have repeatedly recognized, "the penalty of death is different in kind from any other punishment imposed under our system of criminal justice" (*Gregg v. Georgia*, 428 U.S. 153, 188, 1976). It thus requires "a greater degree of accuracy and factfinding than would be true in a noncapital case" (*Gilmore v. Taylor*, 508 U.S. 333, 342, 1993). This has led to specialized procedures for capital cases: attorney competency requirements; provisions for automatic appeal in many states; special requirements for jury sentencing (e.g., unanimous verdicts, consideration of mitigating evidence); and, in many states, review for proportionality. These measures and others are meant to address the heightened need for accuracy in capital cases to ensure that the exercise of the State's ultimate authority to kill a defendant is meted out only to those deemed legally deserving of such a final and irrevocable punishment.

Though courts usually consider evidence regarding the FE relevant to their interpretation of defendants' IQs, application of the effect is not mandatory. Judges are often particularly hesitant to conclude that, because a general effect exists, an individual's IQ should be adjusted downward accordingly—apparently some psychologists also view generalizing a group effect to an individual as an undue leap of inference and therefore a reason not to make the Flynn correction. This is really a straw man argument.

First, nearly all effects in psychology are based on aggregated data and groups and subsequent probability estimations from groups to individuals. Any prediction formula, and these are used often by most all psychologists involved in forensic cases, in employment decisions, prediction of achievement levels, diagnosis of specific learning disabilities, college admissions, and so on to name a few, is based on groups and then the formulae are applied to individual cases. However, to argue the FE should not be applied to individuals belies the fact that all IQs, obtained or otherwise, are to a significant extent based on a group effect and derived from aggregated data. This is because we have only interval scaling, not ratio scaling, available to us in determining scores such as IQs. The determination of an individual's IQ begins with defining the midpoint of the distribution of performance of a sample (a group of people) of a target population. With interval scaling, the only point we can initially locate accurately is the middle of the score distribution—we then measure outward toward the two ends of the distribution of scores based on the variance of the group we used to derive the scores. We then generalize this set of group effects to the performance of individuals and place them on the group distribution and demarcate their placement with the assignment of an IQ (for a more detailed explanation see Reynolds & Livingston, in press). As the group used to provide these statistics ages and becomes less like the current target population, applying any correction that can improve the accuracy of the placement of the individual on this continuum (e.g., the Flynn correction) improves IQ estimation for the individual. The admonishments of the U.S. Supreme Court, in multiple incarnations, that death penalty cases require special attention to accuracy apply an even more profound legal argument to applying this correction.

Zhou, Zhu, and Weiss (2010) point to another controversy surrounding the FE that may be applicable to the size of the correction needed in an individual case. The FE may not be constant across the full distribution of IQs. Depending on the method and assumptions, they show the FE may vary in nonsignificant magnitudes across the full range of IQ. Using the largest and most stable samples, which were collapsed across scales after a statistical demonstration of constant effects by instrument, and assessed using analysis of variance (ANOVA) and analysis of covariance (ANCOVA), a larger FE was observed at the lower end of the IQ range, that is, obtained IQs less than or equal to 79. However, using an equipercentile approach to equating, disparate results were seen where on some tests the prior pattern was upheld but not on others—a reversal of the pattern occurring in some instances. However, the larger sample size and the use of the verbal composite to block and thus lessen any regression effects in the first analyses, appears more

reliable in its results, and perhaps a stronger fit to theories of cognitive development as well. The changes in magnitude of the FE reported in Zhou et al. may also be related to chronological age at the time of testing as it varied, albeit inconsistently, by age appropriate version of the Wechsler Scales. This could be addressed if tests that examine a large age range with a common set of core tasks could be subjected to similar analyses (e.g., the Reynolds Intellectual Assessment Scales; Reynolds & Kamphaus, 2003); however, the necessary data on these instruments are not available. Taken as a whole, and noting some of the inconsistencies in the results, the Zhou et al. analyses support the idea that an even larger correction may need to be applied to low IQs, especially given the paucity of individuals scoring at the extremes in the samples evaluated, but this remains for future research with much larger samples. For now, best practice is the application of the Flynn correction as a constant by year across the distribution.

Where courts strictly adhere to score cutoffs in determining mental retardation, a single point can mean the difference between a constitutional and unconstitutional execution—even if courts were to drop such a rigid adherence, the most accurate estimation of the defendant's IQ is still required. The FE, though certainly not without its detractors, is nevertheless a generally accepted scientific theory. Flynn has demonstrated its clear applicability to individual cases quite clearly as well.[1] (Flynn, 2006, has dealt eloquently with a number of other objections to applying the Flynn correction to individuals, which space limitations do not allow us to address.) The United States has decided to allow states to determine mental retardation almost exclusively by reference to a ranking system that quantifies an individual's standing relative to a reference sample. In doing so, psychologists who work in this system and the courts themselves must ensure that this system is applied as accurately as possible so that no overinclusion into the category of death-eligible individuals occurs.

Conclusion

In criminal proceedings, the law's primary concern is that justice is meted according to the procedures and guarantees contained in the federal and state constitutions. These constitutional concerns, as well as the need for accuracy, are at their highest when the death penalty is at issue. The highest court in this country has made the determination that executing persons with mental retardation violates the Eighth Amendment's prohibition against cruel and unusual punishment. As a generally accepted scientific theory that could potentially make the difference between a constitutional and unconstitutional execution, the FE must be applied in the legal context. Those who oppose the Flynn correction must either dispute the scientific validity of the FE (and we see no such serious challenges—the remaining issue seems to be over why it occurs, a debate that is irrelevant to whether it should be applied), have a poor understanding of the death penalty and the writings of the Supreme Court on the matter, or perhaps simply do not understand interval scaling and its implications for how test scores are derived, the purpose and application of reference samples (i.e., norm groups), or how predictions are made in psychology (unfortunately, a too common state of affairs in professional psychology, e.g., see Reynolds, 2010). If the FE is real, the failure to apply the Flynn correction as we have described it is tantamount to malpractice. No one's life should depend on when an IQ test was normed.

Declaration of Conflicting Interests

The author(s) declared no conflicts of interest with respect to the authorship and/or publication of this article.

Funding

The author(s) received no financial support for the research and/or authorship of this article.

Note

1. See, for example, affidavit of James R. Flynn in the case of Earl Wesley Berry, August 8, 2004, which noted a 7-point difference in two of the defendant's IQ scores, obtained 1 month apart but from different versions of the Wechsler Adult Intelligence Scale (WAIS) test. Adjusting the scores according to the FE yielded an identical score across the two tests.

References

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

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains from 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.

Flynn, J. R. (2007a). *What is intelligence? Beyond the Flynn effect*. New York, NY: Cambridge University Press.

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

Reynolds, C. R. (2010). Measurement and assessment: An editorial view. *Psychological Assessment, 22*, 1-4. doi:10.1037/a0018811

Reynolds, C. R., & Kamphaus, R. W. (2003). *Reynolds Intellectual Assessment Scales*. Lutz, FL: Psychological Assessment Resources.

Reynolds, C. R., & Livingston, R. A. (in press). *Mastering psychological testing*. Boston, MA: Allyn & Bacon.

Reynolds, C. R., Price, R., & Niland, J. (2003). The role of neuropsychology in capital felony (death penalty) defense. *Journal of Forensic Neuropsychology, 3*, 89-123.

Zhou, X., Zhu, J., & Weiss, L. A. (2010). Peeking inside the "black box" of the Flynn effect: Evidence from three Wechsler instruments. *Journal of Psychoeducational Assessment, 28*, 399-411.

INTELLECTUAL AND DEVELOPMENTAL DISABILITIES VOLUME 49, NUMBER 3: 131–140 | JUNE 2011

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*,

INTELLECTUAL AND DEVELOPMENTAL DISABILITIES VOLUME 49, NUMBER 3: 131–140 | JUNE 2011

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

INTELLECTUAL AND DEVELOPMENTAL DISABILITIES VOLUME 49, NUMBER 3: 131–140 | JUNE 2011

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.

Published in final edited form as:
Psychol Bull. 2014 September ; 140(5): 1332–1360. doi:10.1037/a0037173.

The Flynn Effect: A Meta-analysis

Lisa Trahan, **Karla K. Stuebing**, **Merril K. Hiscock**, and **Jack M. Fletcher**
University of Houston

Abstract

The "Flynn effect" refers to the observed rise in IQ scores over time, resulting in norms obsolescence. Although the Flynn effect is widely accepted, most approaches to estimating it have relied upon "scorecard" approaches that make estimates of its magnitude and error of measurement controversial and prevent determination of factors that moderate the Flynn effect across different IQ tests. We conducted a meta-analysis to determine the magnitude of the Flynn effect with a higher degree of precision, to determine the error of measurement, and to assess the impact of several moderator variables on the mean effect size. Across 285 studies (N = 14,031) since 1951 with administrations of two intelligence tests with different normative bases, the meta-analytic mean was 2.31, 95% CI [1.99, 2.64], standard score points per decade. The mean effect size for 53 comparisons (N = 3,951) (excluding three atypical studies that inflate the estimates) involving modern (since 1972) Stanford-Binet and Wechsler IQ tests (2.93, 95% CI [2.3, 3.5], IQ points per decade) was comparable to previous estimates of about 3 points per decade, but not consistent with the hypothesis that the Flynn effect is diminishing. For modern tests, study sample (larger increases for validation research samples vs. test standardization samples) and order of administration explained unique variance in the Flynn effect, but age and ability level were not significant moderators. These results supported previous estimates of the Flynn effect and its robustness across different age groups, measures, samples, and levels of performance.

Keywords

Flynn effect; IQ test; intellectual disability; capital punishment; special education

Historical Background

The "Flynn effect" refers to the observed rise over time in standardized intelligence test scores, documented by Flynn (1984a) in a study on intelligence quotient (IQ) score gains in the standardization samples of successive versions of Stanford-Binet and Wechsler intelligence tests. Flynn's study revealed a 13.8-point increase in IQ scores between 1932 and 1978, amounting to a 0.3-point increase per year, or approximately 3 points per decade. More recently, the Flynn effect was supported by calculations of IQ score gains between 1972 and 2006 for different normative versions of the Stanford-Binet (SB), Wechsler Adult Intelligence Scale (WAIS), and Wechsler Intelligence Scale for Children (WISC) (Flynn, 2009a). The average increase in IQ scores per year was 0.31, which was consistent with Flynn's (1984a) earlier findings.

The Flynn effect implies that an individual will likely attain a higher IQ score on an earlier version of a test than on the current version. In fact, a test will overestimate an individual's IQ score by an average of about 0.3 points per year between the year in which the test was normed and the year in which the test was administered. The ramifications of this effect are especially pertinent to the diagnosis of intellectual disability in high stakes decisions when an IQ cut point is used as a necessary part of the decision-making process. The most dramatic example in the United States is the determination of intellectual disability in capital punishment cases. These determinations in so-called Atkins hearings represent life and death decisions for death row inmates scheduled for execution. Because an inmate may have received several IQ scores with different normative samples over time, whether to acknowledge the Flynn effect is a major bone of contention in the legal system. In addition, the Flynn effect figures in access to services and accommodations, such as determining eligibility for special education and American Disability Act services and Social Security Disability Insurance (SSDI) in the United States.

More generally, conceptions about IQ as a predictor of success in various domains is pervasive in many domains of the behavioral sciences and in Western societies. Many studies use IQ scores as an outcome variable or to characterize the sample. In clinical practice, most assessments routinely administer an IQ test and most applied training programs teach administration and interpretation of IQ test scores. Organizations like MENSA set IQ levels associated with "genius" and people commonly refer to others as "bright" or use more pejorative terms as an indicator of their level of ability. Although the meaningfulness of these uses of IQ scores is beyond the scope of this investigation, they illustrate the pervasiveness of concepts about IQ scores as indicators of individual differences and level of performance.

The Flynn effect is less well known and often not taught in behavioral science training programs (Hagen, Drogin, & Guilmette, 2008). It is important because the normative base of the test directly influences the interpretation of the level of IQ. MENSA, the "high IQ society," requires an IQ score in the top 2% of the population (www.us.mensa.org/join/testscores/qualifyingscores). The organization accepts scores from a variety of tests, often with no specification of which version of the test. The Stanford-Binet IV and Stanford-Binet 5 are both permitted. If a person applied and took an IQ test in 2014, the required score of 132 on the Stanford-Binet 4 would be equivalent to a score of 126 on the recently normed Stanford-Binet 5 because the normative sample was formed 20 years ago. Although the Flynn effect is not necessarily of general interest to psychology, the pervasive use of IQ test scores in clinical practice and research, in high stakes decisions, and in Western society suggests that it should be. It is not surprising that a PsycINFO® search shows that the number of articles on the Flynn effect rose from 6 in 2001–2002 to 54 in 2010–2011. Most significant is the use of IQ scores in identifying intellectual disabilities and the death penalty, where there are literally hundreds of active cases in the judicial system, and in determining eligibility for social services and special education.

Definition of Intellectual Disability

The identification of an intellectual disability in the United States requires the presence of significant limitations in intellectual functioning and adaptive behavior prior to age 18 (American Association on Intellectual and Developmental Disabilities [AAIDD], 2010). An IQ score at least two standard deviations below the mean (i.e., ≤ 70) is a common indicator of a significant limitation in intellectual functioning, and captures approximately 2.2% of the population. Although the gold standard AAIDD criteria stress the importance of exercising clinical judgment in the interpretation of IQ scores (e.g., accounting for measurement error), a cut-off score of 70 commonly is used to indicate a significant limitation in intellectual functioning (Greenspan & Switzky, 2006). Thus, were an adult to have attained an IQ score of 73 on the Wechsler Intelligence Scale for Children--Revised (WISC-R) as a child, s/he might not be identified as having a significant limitation in intellectual functioning. However, suppose the WISC-R had been administered in 1992, 20 years after the test was normed. The Flynn effect would have inflated test norms by 0.3 points per year between the year in which the test was normed (1972) and the year in which the test was administered (1992). Correction for that inflation would reduce the person's IQ score by six points, to 67, thereby indicating a significant limitation in intellectual functioning and highlighting the problems with obsolete norms. Further, the WISC-III, published in 1989, would have been the current edition of the test when the child was tested. This underscores the importance of testing practices (e.g., acquiring and administering the current version of a test) in formal education settings.

High Stakes Decisions

Capital punishment

The Eighth Amendment of the U.S. Constitution prohibits cruel and unusual punishment, and that prohibition informed the Court's decision in Atkins v. Virginia (2002) to abstain from imposing the death penalty on a defendant with an intellectual disability. In this case, Daryl Atkins, a man determined to have a mild intellectual disability, was convicted of capital murder. The Supreme Court of Virginia initially imposed the death penalty on Atkins; however, the United States Supreme Court reversed the decision due to the presumed difficulty people with intellectual disabilities have in understanding the ramifications of criminal behavior and the emergence of statutes in a growing number of states barring the death penalty for defendants with an intellectual disability.

In 2008, a report indicated that since the reversal of the death penalty in Atkins' case, 80+ death penalty pronouncements have been converted to life in prison (Blume, 2008). This number has increased significantly since 2008. Importantly, Walker v. True (2005) set a precedent for the consideration of the Flynn effect in capital murder cases. The defendant argued in an appeal that his sentence violated the Eighth Amendment; when corrected for the Flynn effect, his IQ score of 76 on the WISC, administered to the defendant in 1984 when he was 11 years old, would be reduced by four points to 72. He alleged that a score of 72 fell within the range of measurement error recognized by the AAIDD (2010) and the American Psychiatric Association (APA, 2000) for a true score of 70. The judges agreed that the Flynn effect and measurement error should be considered in this case. There are

hundreds of Atkins hearings involving the Flynn effect in some manner and other issues related to the use of IQ tests (see AtkinsMR/IDdeathpenalty.com)

Special education

Demonstration of an intellectual disability or a learning disability is an eligibility criterion for receipt of special education services in schools. Kanaya, Ceci, and Scullin (2003a) and Kanaya, Scullin, and Ceci (2003b) documented a pattern of "rising and falling" IQ scores in children diagnosed with an intellectual disability or learning disability as a function of the release date of the new version of an intelligence test. One study (Kanaya et al., 2003a) mapped IQ scores obtained from children's initial special education assessments between 1972 and 1977, during the transition from the WISC to the WISC-R, and between 1990 and 1995, during the transition from the WISC-R to the WISC-III. The authors reported a reduction in IQ scores during the fourth year of each interval (one year after the release of the new test version) followed by an increase in IQ scores during subsequent years. In a second study (Kanaya et al., 2003b), the authors reported a 5.6-point reduction in IQ score for children initially tested with the WISC-R and subsequently tested with the WISC-III, with a significantly greater proportion of these children being diagnosed with an intellectual disability during the second assessment than children who completed the same version of the WISC during both assessments. More recent studies have supported these patterns in children assessed for learning disabilities with the WISC-III (Kanaya & Ceci, 2012).

Taken together, these studies suggest that the use of obsolete norms leads to inflation of the IQ scores of children referred for a special education assessment as a function of the time between the year in which the test was normed and the year in which the test was administered. The use of a test with obsolete norms reduces the likelihood of a child being identified with an intellectual disability and receiving appropriate services, and may increase the prevalence of learning disabilities; the inflated IQ score helps produce a discrepancy between intellectual functioning and achievement, which in education settings has often been interpreted as indicating a learning disability (Fletcher et al., 2007). These studies also highlight the importance of using the current version of a test in education settings, a practice which may be thwarted by a school district's budgetary constraints and challenges associated with learning the administration and scoring procedures for the new test (Kanaya & Ceci, 2007).

Social security disability

As with determination of the death penalty and eligibility for special education, IQ testing remains an important component of the decision-making process for determining eligibility for SSDI as a person with an intellectual disability. Like the AAIDD, the Social Security Administration (2008) requires significant limitations in intellectual functioning and adaptive behavior for a diagnosis of intellectual disability; however, these limitations must be present prior to age 22. Moreover, individuals with an IQ at or below 59 are eligible de facto for SSDI, whereas those with an IQ between 60 and 70 must demonstrate work-related functional limitations resulting from a physical or other mental impairment, or two other specified functional limitations (e.g., social functioning deficits). The manual, like the

AAIDD manual, explicitly discusses the importance of correcting for the Flynn effect, but acknowledges that precise estimates are not available.

Flynn's Work

Flynn's (1984a) landmark study, which revealed increasing IQ at a median rate of 0.31 points per year between 1932 and 1978 across 18 comparisons of the SB, WAIS, WISC, and Wechsler Preschool and Primary Scale of Intelligence (WPPSI), was the first analysis of its kind. Seventy-three studies totaling 7,431 participants provided support for this effect. Whereas Flynn's (1984a) study focused on comparisons documented in publication manuals of primarily the first editions of the Stanford-Binet and Wechsler tests, a second study investigated IQ gains in 14 developed countries using a variety of instruments, including Ravens Progressive Matrices, Wechsler, and Otis-Lennon tests (Flynn, 1987). IQ gains amounted to a median of 15 points in one generation, described by Flynn (1987) as "massive." An extension of Flynn's (1984a) work documented a mean rate of IQ gain equaling approximately 0.31 IQ points per year across 12 comparisons of the SB, WAIS, and WISC standardization samples (Flynn, 2007), a value highly consistent with earlier findings. Further, 14 comparisons of Stanford-Binet and Wechsler standardization samples, accounting for the recent publication of the WAIS-IV, revealed an annual rate of IQ gain equaling 0.31 (Flynn, 2009a). These latter findings, based on the simple averaging of IQ gains across studies, were supported by the only meta-analysis addressing the Flynn effect (Fletcher, Stuebing, & Hughes, 2010). For these 14 studies, Fletcher et al. (2010) calculated a weighted mean rate of IQ gain of 2.80 points per decade, 95% CI [2.50, 3.09], and a weighted mean rate of IQ gain of 2.86, 95% CI [2.50, 3.22], after excluding comparisons that included the WAIS-III because effect sizes produced by comparisons between the WAIS-III and another test differed considerably from the effect sizes produced by comparisons between other tests. The puzzling effects produced by comparisons including the WAIS-III were consistent with Flynn's (2006a) study, wherein he demonstrated that IQ score inflation on the WAIS-III was reduced because of differences in the range of possible scores at the lower end of the distribution.

Other notable investigations conducted by Flynn include the computation of a weighted average IQ gain per year of 0.29 between the WISC and WISC-R across 29 studies comprising 1,607 subjects (1985): a rate of IQ gain per year of 0.31 between the WISC-R and the WISC-III across test manual studies and a selection of studies carried out by independent researchers (1998a); and a rate of IQ gain per year of 0.20 between the WAIS-R and WAIS-III across test manual studies (1998a). Prior to these studies, Flynn (1984b) also reported SB gains across standardization samples, and both real and simulated gains for the WPPSI and the first two versions of the WISC and WAIS. Flynn (1988b) noted consistent gains between the WISC (N = 93) and WISC-R (N = 296) in Scottish children (1990); for the Matrices and Instructions tests in an Israeli military sample totaling approximately 26,000 subjects per year between 1971 and 1984; between the WISC-III and an earlier version of the test in samples from the United States, West Germany, Austria, and Scotland totaling 3,190 subjects (2000); and for the Coloured Progressive Matrices in British standardization samples totaling 1,833 participants (2009b). The existence of the Flynn effect is rarely disputed. However, a working magnitude and measurement error associated

with the Flynn effect are not well established, leaving unanswerable the question of how much of a correction – if any – to apply to IQ test scores to account for the norming date of the test. Further, there is considerable contention over factors that may cause the Flynn effect (Flynn, 2007, 2012; Neisser, 1998).

Proposed Causes of the Flynn Effect

There are multiple hypotheses about the basis for the Flynn effect, including genetic and environmental factors, and measurement issues.

Genetic hypotheses

Mingroni (2007) hypothesized that IQ gains are the result of increasingly random mating, termed heterosis (or hybrid vigor), a phenomenon that produces changes in traits governed by the combination of dominant and recessive alleles. However, Lynn (2009) noted that the Flynn effect in Europe has mirrored the effect in the United States despite evidence of minimal migration to Europe prior to 1950 and limited inter-mating between native and immigrant populations since then. A more comprehensive argument against a genetic cause for the Flynn effect has been made by Woodley (2011).

Environmental factors

Woodley (2011) argued that “The [Flynn] effect only concerns the non-*g* variance unique to specific cognitive abilities” (p. 691), presumably bringing environmental explanations for the Flynn effect to the forefront. Environmental factors hypothesized as moderators of the Flynn effect include sibship size (Sundet, Borren, & Tambs, 2008) and pre-natal and early post-natal nutrition (Lynn, 2009). In Norway, Sundet et al. demonstrated that an increase in IQ scores paralleled a decrease in sibship size, with the greatest increase in IQ scores occurring between cohorts with the greatest decrease in sibship size. For example, between birth cohort 1938–1940 and 1950–1952, the percentage of sibships composed of 6+ children decreased from 20% to 5%, and IQ score increased by 6 points.

With rates of Development Quotient score gains in infants mirroring IQ score gains of preschool children, school-aged children, and adults, Lynn (2009) questioned the validity of explanations whose effects would emerge later in development, such as improvements in child rearing (Elley, 1969) and education (Tuddenham, 1948); increased environmental complexity (Schooler, 1998), test sophistication (Tuddenham, 1948), and test-taking confidence (Brand, 1987); and the effects of genetics (Jensen, 1998) and the individual and social multiplier phenomena (Dickens & Flynn, 2001a; Dickens & Flynn, 2001b). Lynn (2009) proposed improvements in pre- and post-natal nutrition as likely causes of the Flynn effect, citing a parallel increase in infants of other nutrition-related characteristics, including height, weight, and head circumference. Improvement to the prenatal environment is also supported by trends in the reduction of alcohol and tobacco use during pregnancy (Bhuvaneswar, Chang, Epstein, & Stern, 2007; Tong, Jones, Dietz, D’Angelo, & Bombard, 2009).

Neisser (1998) suggested that increasing IQ scores have mirrored socioenvironmental changes in developing countries. If IQ test score changes are a product of

socioenvironmental improvements, then as living conditions optimize, IQ scores should plateau. This suggestion has been echoed by Sundet, Barlaug, and Torjussen (2004), who documented a plateau in IQ scores in Norway (Sundet et al., 2004) and speculated that changes in family life factors (e.g., family size, parenting style, and child care) might be partly responsible for this pattern. A decline in IQ scores has even been noted in Denmark (Teasdale & Owen, 2008; Teasdale & Owen, 2005), a pattern that the authors suggested might be due to a shift in educational priorities toward more practical skills manifest in the increasing popularity of vocational programs for post-secondary education.

Although Flynn (2010) acknowledged that his “scientific spectacles” hypothesis may no longer explain current IQ gains, he maintained that there was a period of time when it was the foremost contributor. Putting on “scientific spectacles” refers to the tendency of contemporary test takers to engage in formal operational thinking, as evidenced by a massive gain of 24 IQ points on the Similarities subtest of the WISC, a measure of abstract reasoning, between 1947 and 2002, a gain unparalleled by any other subtest (Flynn & Weiss, 2007). Conceptualizing IQ gains as a shift in thinking style from concrete operational to formal operational rather than an increase in intelligence per se would explain why previous generations thrived despite producing norms on IQ tests that overestimated the intellectual abilities of future generations (Flynn, 2007). However, this difference may be more simply attributed to changes across different versions of Similarities and other verbal subtests (Kaufman, 2010) of the WISC. Nonetheless, Dickinson and Hiscock (2010) reported a Flynn effect for WAIS Similarities of 4.5 IQ points per decade for WAIS to WAIS-R and 2.6 IQ points per decade for WAIS-R to WAIS-III. The average was 3.6 IQ points per decade or 0.36 IQ points per year. This change in adult performance is only moderately less than Flynn’s 0.45 points per year for the WISC between 1947 and 2002.

Measurement issues

Tests of verbal ability, compared with performance-based measures, have been reported to be less sensitive to the Flynn effect (Flynn, 1987; Flynn, 1994; Flynn, 1998b; Flynn, 1999), which may be related to changes in verbal subtests. Beaujean and Osterlind (2008) and Beaujean and Sheng (2010) used Item Response Theory (IRT) to determine whether increases in IQ scores over time reflect changes in the measurement of intellectual functioning rather than changes in the underlying construct, i.e., the latent variable of cognitive ability. Although changes in Peabody Picture Vocabulary Test-Revised scores were negligible (Beaujean & Osterlind, 2008), it is a verbal test that differs in many respects from Wechsler and Stanford-Binet tests. Wicherts et al. (2004) found that intelligence measures were not factorially invariant, such that the measures displayed differential patterns of gains and losses that were unexpected given each test’s common factor means. Taken together, these studies suggest that increases in IQ scores over time may be at least partly a result of changes in the measurement of intellectual functioning. Moreover, Dickinson and Hiscock (2010) reported that published norms for age-related changes in verbal and performance subtests do not take into account the Flynn effect. In comparisons of subtest scores from the WAIS-R and WAIS-III in 20-year-old and 70-year-old cohorts, the Flynn-corrected difference in Verbal IQ between 20-year-olds and 70-yearolds was 8.0 IQ points favoring the 70-year-olds (equivalent to 0.16 IQ points per year). In contrast, the

younger group outscored the older group in Performance IQ by a margin of 9.5 IQ points (equivalent to 0.19 IQ points per year). These findings suggested that apparent age-related declines in Verbal IQ between the ages of 20 and 70 years are largely artifacts of the Flynn effect and that, even though age-related declines in Performance IQ are real declines, the magnitudes of those declines are amplified substantially by the Flynn effect.

Some studies have examined intercorrelations among subtests of IQ measures to determine the variance in IQ scores explained by *g*, with preliminary evidence suggesting that IQ gains have been associated with declines in measurement of *g* (Kane & Oakland, 2000; Te Nijenhuis & van der Flier, 2007). Flynn (2007), on the other hand, has discounted the association between *g* and increasing IQ scores, and a dissociation between *g* and the Flynn effects has been claimed by Rushton (2000). However, Raven's Progressive Matrices, renowned for its *g*-loading, has demonstrated a rate of IQ gain of 7 points per decade, more than double the rate of the Flynn effect as manifested on WAIS, SB, and other multifactorial intellectual tests (Neisser, 1997).

What is Rising?

The theories highlighted above offer explanations for the Flynn effect but leave an important question unanswered: What exactly does the Flynn effect capture (i.e., what is rising)? Although much of the previous research on the Flynn effect has focused on the rise of mean IQ scores over time, studies distinguishing rates of gain among elements of IQ tests more readily answer the question of what is rising. Relative to scores produced by verbal tests, there have been greater gains in scores produced by nonverbal, performance-based measures like Raven's Progressive Matrices (Neisser, 1997) and Wechsler performance subtests (Dickinson & Hiscock, 2011; Flynn, 1999). These types of tests are strongly associated with fluid intelligence, suggesting less of a rise in crystalized intelligence that reflects the influence of education, such as vocabulary. A notable exception is the increasing scores produced by the Wechsler verbal subtest Similarities (Flynn, 2007; Flynn & Weiss, 2007), although this subtest taps into elements of reasoning not required by the other subtests comprising the Wechsler Verbal IQ composite.

Dickens and Flynn (2001b) provided a framework for understanding the rise in more fluid versus crystallized cognitive abilities. They identified social multipliers as elements of the sociocultural milieu that contributed to rising IQ scores among successive cohorts of individuals. Flynn (2006b) highlighted two possible sociocultural contributions to the Flynn effect, one related to patterns of formal education and the other to the influence of science. Specifically, years of formal education increased in the years prior to World War II, whereas priorities in formal education shifted from rote learning to problem solving in the years following World War II. As time continued to pass, the value placed on problem solving in the workplace and leisure time spent on cognitively engaging activities continued to exert an effect on skills assessed by nonverbal, performance-based measures. The second sociocultural contributor, science, refers to the simultaneous rise in the influence of scientific reasoning and the abstract thinking and categorization required to perform well on nonverbal, performance-based measures.

The Current Study

The primary objective of this meta-analysis was to determine whether the Flynn effect could be replicated and more precisely estimated across a wide range of individually administered, multifactorial intelligence tests used at different ages and levels of performance. Answers to these research questions will assist in determining the confidence with which a correction for the Flynn effect can be applied across a variety of intelligence tests, ages, ability levels, and samples. By completing the meta-analysis, we also hoped to provide evidence evaluative of existing explanations for the Flynn effect, thus contributing to theory.

With the exceptions of the Flynn (1984a, 2009a) and Flynn and Weiss (2007) analyses of gains in IQ scores across successive versions of the Stanford-Binet and Wechsler intelligence tests, most research comparing IQ test scores has focused on correlations between two tests and/or average mean difference between two successive versions of the same test. This study will expand the literature on estimates of the Flynn effect by computing more precisely the magnitude of the effect over multiple versions of several widely-used, individually administered, multifactorial intelligence tests, viz., Kaufman, Stanford-Binet, and Wechsler tests and versions of the Differential Ability Scales, McCarthy Scales of Children's Abilities, and the Woodcock-Johnson Tests of Cognitive Abilities. The data for these computations were obtained from validity studies conducted by test publishers or independent research teams. In addition to providing more precise weighted meta-analytic means, meta-analysis allows estimates of the standard error and evaluation of potential moderators.

This study deliberately focused on sources of heterogeneity (i.e., moderators) that could be readily identified through meta-analytic searches and that helped explain variability in estimates of the magnitude of the Flynn effect. Investigation of these moderators is needed to advance understanding of variables that might limit or promote confidence in applying a correction for the Flynn effect in high stakes decisions. Here the IQ tests that are used are variable in terms of test and normative basis, with the primary focus on the composite score. The tests are given to a broad age range and to people who vary in ability. It is not clear that the standard Flynn effect estimate can be applied among individuals of all ability levels and ages who took any of a number of individually-administered, multifactorial tests. In addition, there may be special circumstances related to test administration setting that might influence the numerical value of the Flynn effect. If the selected moderators (i.e., ability level, age, IQ tests administered, test administration setting, and test administration order) influence the estimate of the Flynn effect, the varying estimates will contribute to the tenability of the theories offered above for the existence and meaning of the Flynn effect.

The evidence for influences of these moderators is mixed, with no clear directions. Recent evidence has suggested that middle and lower ability groups (IQ = 79–109) demonstrate the customary 0.31–0.37-point increase per year, whereas higher ability groups (IQ = 110+) demonstrate a minimal increase of 0.06–0.15 points per year (Zhou, Zhu, Weiss, & Pearson, 2010). Whereas some previous studies have supported this finding (e.g., Lynn & Hampson, 1986; Teasdale & Owen, 1989), others have not. Two studies found the opposite pattern (Graf & Hinton, 1994; Sanborn, Truscott, Phelps, & McDougal, 2003), and one study

indicated smaller gains at intelligence levels both above and below average, with the highest gains evident in people at the lowest end of the ability spectrum (Spitz, 1989). Little research has been conducted to investigate the relation between age and gains in IQ score. Cross-sectional research has indicated no difference among young children, older children, and adults (Flynn, 1984b) and no difference among adult cohorts ranging in age from 35–80 years (Ronnlund & Nilsson, 2008).

Research on the Flynn effect has focused almost exclusively on the effect produced from administrations of the Stanford-Binet and Wechsler tests. This study expanded the scope by including a wider range of individually administered, largely multifactorial intelligence tests. Comparisons of older and more recently normed versions of the Stanford-Binet and Wechsler tests were conducted to facilitate comparisons with previous work and help determine if the Flynn effect has remained constant over time.

Another potential moderator pertains to study sample. Study data were collected by test publishers or independent researchers for validation purposes, or by mental health professionals for clinical decision-making purposes. Validation studies conducted by test publishers likely employed the most rigorous procedures with regard to sampling, selection of administrators, and adherence to administration and scoring protocols. However, the more homogenous samples examined in the research and clinical studies (e.g., children suspected of having an intellectual disability or juvenile delinquents) may produce results that are more generalizable to specific populations and permit comparison of Flynn effect values across those special populations.

Another set of moderators involves measurement issues, such as changes in subtest configuration and order effects. These issues were addressed by Kaufman (2010), who pointed out that changes in the instructions and content of specific Wechsler subtests (e.g., Similarities) could make comparing older and newer versions akin to comparing apples and oranges. However, other research has shown that estimates of the size of the Flynn effect based on changes in subtest scores yield values similar to estimates from the composite scores (Agbayani & Hiscock, 2013; Dickinson & Hiscock, 2010). Kaufman's concern related to interpretations of the basis of the Flynn effect and not to its existence, and we did not pursue this question because it has been addressed in other studies (Dickinson & Hiscock, 2011). Subtest coding of a larger corpus of tests was difficult because the data were often not available. However, Kaufman also suggested that the Flynn effect could be the result of prior exposure when taking the newer version of an IQ test first and then transferring a learned response style to the older IQ test, thus receiving higher scores when the older test is given second. In order for order effects to occur, the interval between the administration of the new and old tests would have to be short enough for the examinee to demonstrate learning, which is often the case in studies comparing different versions of an IQ test, the basis for determination of the Flynn effect.

Although the Flynn effect has been well documented during the 20th century, the meta-analytic method used during the current study is a novel approach to documenting this phenomenon. The method of the current study aligns with a key research proposal identified by Rodgers (1999) as important in advancing our understanding of the Flynn effect; viz., a

formal meta-analysis. Although many of Rodgers' (1999) proposals have since been implemented, there remains room for understanding the meaning of the Flynn effect, how the Flynn effect is reflected in batteries of tests over time, and how the Flynn effect manifests itself across subsamples defined by ability level or other characteristics.

Method

Inclusion and Exclusion Criteria

Studies identified from test manuals or peer-reviewed journals were included if they reported sample size and mean IQ score for each test administered; these variables were required for computation of the meta-analytic mean. All English-speaking participant populations from the United States and the United Kingdom were included. Variations in study design were acceptable. Administration of both tests must have occurred within one year of one another. Studies could have been conducted at any point prior to the completion date of the literature search in 2010.

We limited our primary investigation to comparisons between tests with greater than five years between norming periods, which is consistent with Flynn's (2009) work. The rationale for this decision was that any difference in IQ scores from a short interval, even seemingly insignificant ones, would be magnified when converted to a value per decade (see Flynn, 2012). As a secondary analysis, we expanded our investigation to all comparisons between tests with at least one year between norming periods to assess whether our decision to limit our investigation to comparisons between tests with greater than five years between norming periods affected the results of the meta-analysis. We did not include comparisons between tests with one year or less between norming periods since years between norming periods served as the denominator of our effect size. A value of zero, representing no difference in years between norming periods, produced an error in the effect size estimate. Finally, we did not include single construct tests, such as the Peabody Picture Vocabulary Test or the Test of Nonverbal Intelligence. There may be other multifactorial tests to consider, but the 27 we chose represent the major IQ tests in use over the past few decades.

Search Strategies

Twenty-seven intelligence test manuals for multifactorial measures were obtained, one for each version of the Differential Ability Scales (Elliot, 1990; Elliot, 2007), Kaufman Adolescent and Adult Intelligence Test (Kaufman & Kaufman, 1993), Kaufman Assessment Battery for Children (Kaufman & Kaufman, 1983; Kaufman & Kaufman, 2004a), Kaufman Brief Intelligence Test (Kaufman & Kaufman, 1990; Kaufman & Kaufman, 2004b), McCarthy Scales of Children's Abilities (McCarthy, 1972), Stanford-Binet Intelligence Scale (Roid, 2003; Terman & Merrill, 1937; Terman & Merrill, 1960; Terman & Merrill, 1973; Thorndike, Hagen, & Sattler, 1986), Wechsler Abbreviated Scale of Intelligence (Wechsler, 1999), Wechsler Adult Intelligence Scale (Wechsler, 1955; Wechsler, 1981; Wechsler, 1997; Wechsler, 2008), Wechsler Intelligence Scale for Children (Wechsler, 1949; Wechsler, 1974; Wechsler, 1991; Wechsler, 2003), Wechsler Preschool and Primary Scale of Intelligence (Wechsler, 1967; Wechsler, 1989; Wechsler, 2002), and Woodcock-

Johnson Tests of Cognitive Ability (Woodcock & Johnson, 1977; Woodcock & Johnson, 1989; Woodcock, McGrew, & Mather, 2001).

Also, a systematic literature review was completed using PsycINFO®, crossing the keywords *comparison*, *correlation*, and *validity* with the full and abbreviated titles of the measures. The first author reviewed each study in full unless abstract review determined the study was not relevant (e.g., some test validation studies included comparisons between tests not under consideration in this meta-analysis). A formal search for unpublished studies was not undertaken; it was presumed that the results of test validation studies would provide important information irrespective of the findings and would therefore constitute publishable data.

Coding Procedures

The first author, who had prior training and experience in coding studies for meta-analyses, coded all of the studies in the current meta-analysis. Two undergraduate volunteers were trained by the first author, and each volunteer coded half the studies. Agreement between the first author and the volunteers on each variable was calculated for blocks of ten studies. These estimates ranged from 90.5–99.1% per block, with an average agreement of 95.8% per block. Discrepancies were resolved through discussion, during which the first author and volunteers referred to the original article. Discrepancies were commonly the result of a coder typo or failure of a coder to locate a particular value in an article.

Moderator Analyses

Moderators included ability level, age, test set, order of administration, and sample. Ability level was coded as the sample's score on the most recently normed test, and age was coded as the sample's age in months. Each comparison was assigned to a test set, as follows. First, due to Flynn's focus on the Stanford-Binet and Wechsler tests, these tests were grouped together and were further separated into an old set and a modern set. The old set included comparisons of only Wechsler and Stanford-Binet tests normed before 1972, with the modern set representing versions normed since 1972. The latter set aligned with comparisons published in Flynn and Weiss (2007) and Flynn (2009). If a modern test was compared to an old test, the comparison was coded old. The Differential Ability Scales, Kaufman Adolescent and Adult Intelligence Test, and Woodcock-Johnson Tests of Cognitive Abilities were grouped together as non-Wechsler/Binet tests with modern standardization samples. The Kaufman Brief Intelligence Test and the Wechsler Abbreviated Scale of Intelligence were grouped together as screening tests. The Kaufman Assessment Battery for Children was separately analyzed due to its grounding in Luria's model of information processing that addressed differences in simultaneous and sequential processing. Fourteen effects remained from the original set of 285 after sorting effects into these groupings. All of these comparisons contained the McCarthy Scales, but with multiple old and modern tests.

Order of administration was included as a moderator variable. Tests were frequently counterbalanced so that approximately half of the sample got each test first. However, in a substantial number of the studies, one test was uniformly given first. We coded these by the

percentage of examinees given the old test first: 100 means that 100% of the examinees got the old test first; 0 means that all examinees got the new test first; 50 means that the tests were counterbalanced. In 7 of these effects, a different value was reported and these were rounded to 0, 0.50 or 100. For example, 14% (given the old test first) was rounded to 0, and 94% was rounded to 100.

Each comparison was also grouped by study sample. *Standardization* studies were completed during standardization and were reported in test manuals. *Research* studies appeared in peer-reviewed journals and examined comparisons among a small selection of intelligence tests. *Clinical* studies reported results from assessments completed of clinical samples, including determination of special education needs.

Statistical Methods

Effect size metric—Comprehensive Meta Analysis software (Borenstein, Hedges, Higgins, & Rothstein, 2005) was used for the core set of analyses. Specifically, we employed the module that requires input of an effect size and its variance for each study. Effects were coded as the difference between the old test mean and the new test mean. Positive effects reflect a positive Flynn effect with the score on the old test higher than the score on the new test despite being taken by the same individuals at approximately the same time. The effect size calculated from each study was the raw difference between the mean score on the old and new tests divided by the number of years between the norming dates of the two tests. This metric is directly interpretable as the estimated magnitude of the Flynn effect per year. Since the scales used by all of the tests were virtually the same (M = 100, SD = 15 or 16), no further standardization (such as dividing by population standard deviation [SD]) was required (Borenstein, Hedges, Higgins, & Rothstein, 2009). The actual SD for each test was used in computing the variance of the effects.

Effect size weighting—The variance for each effect is required for computation of the weight given to each effect in the overall analysis. The weight is the inverse of the variance, so studies with the smallest variance are given the most weight. Small variance (high precision) for an effect is achieved via (a) large Ns, (b) high reliabilities for both tests and high content overlap between tests which are jointly reflected in the correlation between the tests, and (c) long intervals between the norming periods of the two tests. The formula (Borenstein, Hedges, Higgins, & Rothstein, 2009) used for the variance of typical pretest-posttest effects in meta-analysis is:

$$\text{Variance}=\frac{SD^2_{New}+SD^2_{Old}-2rSD_{New}SD_{Old}}{N} \quad (1)$$

Where SD^2_{New} is the variance of the more recently normed test, SD^2_{Old} is the variance of the less recently normed test, r is the reported correlation between the two tests, and N is the total sample size. In the numerator, actual reported correlations were used when available. For 54 of the 285 studies, no correlation was reported. In these cases, if there were other studies that compared the same two tests, the correlations from the other studies were converted to Fisher's *z*. These were then averaged and converted back to a correlation and used in place of the missing value. If no other studies compared the same two tests, the mean

correlation for the entire set of studies was computed and substituted in for the missing value. This occurred for two study results. The mean correlation for each pair of tests was also retained and used in a parallel analysis to determine the impact of using the sample-specific correlation rather than a population correlation in the estimator of the effect variance.

To allow for the differential precision in effects due to the years between norming periods of the two tests being compared, we adapted a formula from Raudenbush and Xiao-Feng (2001) that allows calculation of the change in variance as a function of the change in duration in years of the period between the norming of the two tests, holding number of time points constant. Using D to represent a duration of 1 year, D' to represent a different duration, either longer or shorter, and ω=D'/D to represent the factor of increase or decrease from one year, then the proportion of the variances is equal to:

$$\frac{V'}{V}=\frac{1}{\omega^2} \quad (2)$$

In other words, the variance (V') for an effect with a 5 year duration between norming periods will be 1/25th the size of the variance (V) of an effect with a one year duration between norming periods, all other things being equal. Thus, the variance we entered into the CMA software for each effect size was:

$$\text{Variance}=\frac{SD^2_{New}+SD^2_{Old}-2rSD_{New}SD_{Old}}{N\omega^2} \quad (3)$$

The numerator of the above formula is the variance of the difference between the two tests being compared. The denominator adjusts this variance by the sample size (N) and by the duration in years of the period between the norming of the two tests.

Credibility intervals—In a random effects model, the true variance of effects is estimated. The standard deviation of this distribution is represented by Tau [τ]. Tau is used to form a credibility interval around the mean effect, capturing 95% of the distribution of true effects by extending out 1.96τ from the mean in both positive and negative directions. The credibility interval acknowledges that there is a distribution of true effects rather than one true effect. In interpreting the credibility interval, it is helpful to consider width as well as location. Even a distribution of true effects that is centered near 0 (where the mean effect might not be significant) may contain many members that might be meaningfully large in either direction. Moderator analysis may be used to try to find subsets of effects within this distribution, to narrow the uncertainty about how large the effect might be in a given situation; however, in the case of true random effects, each causal variable might explain a very small portion of the variance and moderator analysis might not improve prediction substantially.

Selection of random effects model—A random effects analytic model was employed because the studies were not strict replications of each other, in which case it would make

sense to expect a single underlying fixed effect. Rather, the studies varied in multiple ways, each of which was expected to have some impact on the observed Flynn effect. These factors include, but are not limited to (a) the specific test pair being compared, (b) the unique population being tested, (c) the age of the sample (which was not always reported quantitatively), (d) the interval between the presentation of the old and new test, (e) the order of presentation of the tests, (f) unusual administration practices (e.g., Spruill, 1988), and (g) interactions among these factors. The result of these multiple causes is a distribution of true effects, rather than a single effect.

In a random effects model, the mean effect is ultimately interpreted as the mean of a distribution of true population effects. Additionally, in a random effects model, the variance of the effects has two variance components. One is due to the true variance in population effects and the second is due to sampling variance around the population mean effect. The result is that the weight given each study is a function of both within-study precision due to sample size and between-study variability. Sample size thus has less effect in the precision of each study. Large sample size studies are given less weight than they would have been in a fixed effects study, and studies with smaller samples are given more weight (Borenstein et al., 2009).

Heterogeneity in effect sizes—Heterogeneity describes the degree to which effect sizes vary between studies. The *Q* statistic is employed to capture the significance of this variance and is calculated by summing the squared differences between individual study effect sizes and the mean effect size. It is distributed as a chi-square statistic with *k*-1 degrees of freedom, where *k* is the number of studies. In addition, I^2 is employed to capture the extent to which detected heterogeneity is due not to chance but to true, identifiable variation between studies. I^2 is calculated:

$$I^2=(Q-df)/Q \quad (4)$$

and once multiplied by 100 is directly interpretable as the proportion of variance due to true heterogeneity.

Publication bias—We did not expect to find evidence for publication bias in this meta-analysis. The descriptive data collected from each study in the form of sample sizes, means, and correlations between tests is not typically the type of data that is subject to tests of significance and thus would not be a direct cause of failure to publish due to non-significance. Additionally, many of the effects were gleaned from the technical manuals of the tests being compared where no publication bias is expected. However, we did evaluate the distributions of effects within each portion of our analysis via funnel plots.

Results

Citations

The literature review produced a total of 4,383 articles. This total does not reflect unique articles, since each article would often appear in multiple keyword searches. One hundred and fifty-four empirical studies and 27 test manuals met inclusion criteria, from which 378

comparisons were extracted, 285 of which were normed more than 5 years apart. The chronological range of the Flynn effect data collected was from 1951 upon publication of Weider, Noller, and Schramm's (1951) comparison study of the WISC and SB to 2010, the year in which the literature review was completed. Table 1 shows the effect size produced by each of the 378 comparisons and includes information pertaining to sample size and age in months.

Overall Model

The mean effect over 285 total studies (n = 14,031) in the random effects model was 0.231 IQ points per year, 95% CI [0.20, 0.26], z = 14.10, $p < .0001$, with a confidence interval and p-value indicating that the Flynn effect is different from zero[1]. The effects were significantly heterogeneous, ($Q_{(284)}$ = 4710, $p < .0001$). The estimated I^2, or proportion of the total variance due to true study variance, was I^2 = 0.94. The Tau, or estimated standard deviation of the true effects, was τ = 0.25, resulting in a credibility interval of −0.26 to +0.72. Eighty-two percent of the distribution of true effects was above zero.

Distribution of Effects

The effects were plotted against their standard error in a funnel plot (Figure 1). There is no apparent publication bias, which would be represented by a gap on the lower left side of the plot. A similar absence of a gap is seen on the lower right side of the plot. What is most apparent in the funnel plot is that many effects fall outside the 1.96 standard error line, suggesting that there is important true heterogeneity in these effects that is not consistent with sampling error alone.

Moderator Analysis

We first modeled the significant heterogeneity in the effect sizes as a function of test set. There was a significant between-test group effect, $Q_{(5)}$ = 231, $p < .0001$, with test group explaining 5.2% of the explainable variance in effects. We then regressed all effects on ability level using Unrestricted Maximum Likelihood for mixed meta-regression within Comprehensive Meta-Analysis software (Borenstein et al., 2005). The range of ability means in the set of effects was 40.6–132.7 standard score points. The intercept was significant (a = 0.38, z = 2.58, $p < .01$), but the slope was not (b = −.002, z = −1.08, $p < .28$), indicating that the effect did not change significantly over the range of ability levels represented in this set of effects.

Further Analysis within Test Groups

We completed separate meta-analyses within test groups to place the results of the modern tests within the context of this larger set. This was done so we could meaningfully compare our results to Flynn's (1984a, 2009a) and Flynn and Weiss' (2007) results, which were

[1]A systematic literature search for manual and empirical studies published since 2010 produced five new studies (Wechsler, 2011 [WASI-II vs. KBIT-2, WASI-II vs. WAIS-IV, WASI-II vs. WASI, WASI-II vs. WISC-IV]; Wilson & Gilmore, 2012 [WISC-IV vs. SB5]), three of which included tests with norming dates at least five years apart. The mean effect over three studies with norming dates at least five years apart in the random effects model was 0.297 IQ points per year, 95% CI [.09, .51]. The mean effect over all five studies in the random effects model was 0.283 IQ points per year, 95% CI [.01, .47]. These results are consistent with the overall results.

based on data published after 1972. Because our focus is on the modern set, we conducted moderator analyses only within that set.

Older Wechsler/Binet tests—The mean effect (k = 152, n = 5,550) of studies involving Wechsler/Binet scales normed before 1972 (and including other IQ tests with an older normative basis) in the random effects model was 0.23 IQ points per year, 95% CI [0.19, 0.27], z =11.12, $p < .0001$. The effects were significantly heterogeneous, ($Q_{(151)} = 3237$, $p < .0001$). The estimated I^2, or proportion of the total variance due to true study variance, was $I^2 = .95$, indicating that very little of the variance in observed effects was attributable to sampling error or unreliability in the tests. The Tau, or estimated standard deviation of the true effects, was $\tau = 0.24$, indicating a 95% credibility interval of −0.23 to +0.70. In other words, approximately 84% of the distribution of true effects was above zero.

Screening tests—The mean effect (k = 17, n = 1,325) in the random effects model was 0.02 IQ points per year, 95% CI [−0.15, 0.19], z = 0.21, $p < .84$. Although the mean effect was not significantly different from 0, the effects were significantly heterogeneous ($Q_{(16)}$ = 232, $p < .0001$). The estimated I^2, or proportion of the total variance due to true study variance, was $I^2 = .93$. The Tau, or estimated standard deviation of the true effects, was τ = 0.33, indicating a 95% credibility interval of −0.63 to +0.66, indicating that more than half of the true effects were above zero.

KABC tests—The mean effect (k = 34, n = 1,611) in the random effects model was 0.02 IQ points per year, 95% CI [−0.16, 0.19], z = 0.19, p = .85. Although the mean effect was not significantly different from zero, the effects were significantly heterogeneous ($Q_{(33)}$ = 295, $p < .0001$). The estimated I^2, or proportion of the total variance due to true study variance, was $I^2 = .89$. The Tau, or estimated standard deviation of the true effects, was τ = 0.47, indicating a 95% credibility interval of −0.90 to +0.93. Again, more than half of the true effects were positive.

Other modern tests—The mean effect (k = 12, n = 925) for the modern tests other than Wechsler and Binet pairs normed since 1972 in the random effects model was 0.30 IQ points per year, 95% CI [0.21, 0.40], z = 6.13, $p < .0001$. Although the mean effect was significantly different from zero, the effects were significantly heterogeneous ($Q_{(11)} = 44$, $p < .0001$). The estimated I^2, or proportion of the total variance due to true study variance, was $I^2 = .75$. The Tau, or estimated standard deviation of the true effects, was $\tau = 0.14$, indicating a credibility interval of 0.03 to +0.57. For the other modern effects, 98.6% of the true effects were positive.

McCarthy test comparisons—The mean effect (k = 14, n = 557) in the random effects model involving the McCarthy was 0.33 IQ points per year, 95% CI [0.15, 0.51], z = 3.60, $p < .0001$. Although the mean effect was significantly different from zero, the effects were significantly heterogeneous ($Q_{(13)} = 74$, $p < .0001$). The estimated I^2, or proportion of the total variance due to true study variance, was $I^2 = .83$. The Tau, or estimated standard deviation of the true effects, was $\tau = 0.28$, indicating a credibility interval of −0.23 to +0.89. For this set of tests, 87.8% of the true effects were positive.

Modern Wechsler/Binet tests—The mean effect (k =56, n = 4,063) for the Wechsler and Binet tests normed since 1972 in the random effects model was 0.35 IQ points per year, 95% CI [0.28, 0.42], z = 10.06, p < .00001. Although the mean effect was significantly different from zero, the effects were significantly heterogeneous ($Q_{(55)}$ = 597.34, p < .0001). The estimated I^2, or proportion of the total variance due to true study variance, was I^2 = .91. The Tau, or estimated standard deviation of the true effects, was τ = 0.23, indicating a credibility interval of −0.10 to +0.80. For the modern effects, 93.5% of the true effects were positive.

Moderator Analyses of the Modern Tests

Ability level—The first moderator selected to explore the significant heterogeneity of the modern tests was ability level. The significant mixed effects meta-regression slope of effect size on ability level was b = −.01, 95% CI [−.016, −.004), z = −3.37, p < .0007. The Q for the model in this analysis was 11.38, accounting for 15.8% of the total variability as estimated by the Unrestricted Likelihood method.

Inspection of Figure 2 revealed an unusual bimodal pattern in the effects representing samples with the lowest ability. This pattern indicates that some of the lower ability samples had higher than average Flynn effects whereas others had lower than average Flynn effects. In order to understand this pattern and its apparent contribution to the heterogeneity of the set of effects, we looked carefully at each of the ten lowest ability studies. Of the five studies with the highest effect sizes in this group (Gordon, Duff, Davidson, & Whitaker, 2010; Nelson & Dacey, 1999; Spruill, 1991; Thorndike, Hagen, & Sattler, 1986), four were comparisons between Stanford-Binet-4 (SB-4) and Wechsler Adult Intelligence Scales-Revised (WAIS-R). The lowest possible score on the SB-4 is 36, and the lowest possible score on the WAIS-R is 45. Individuals who obtain the lowest possible score on both tests will still have an apparent difference in their standard scores of 9 points. Consistent with the plot, as the scores get closer to the mean of 100, the differences in the scales become smaller, and the effects become smaller.

A different factor was noted in the three unusually low effects at the low ability side of the plot. For two of these effects, the administration of the tests was not counterbalanced. All subjects received the old test first. It is possible that for these comparisons, the participants performed better on the second (newer) test than on the first due to an order effect (see below). Effects for the two non-counterbalanced studies fall below the regression line and are the second and fourth from the lowest in ability in that cluster. One (Thorndike, Hagen, & Sattler, 1986) was a comparison of SB4 with Stanford-Binet L-M (floor = 36 points on both tests) and the other (Thorndike, Hagen, & Sattler, 1986) was a comparison of SB-4 with the Wechsler Intelligence Scales for Children-Revised (WISC-R). To evaluate the influence of these potentially highly influential but atypical effects to the analysis, we ran a cumulative analysis of the meta-analytic effect. We arranged all modern effects in descending order by ability level and then added them to the meta-analysis one at a time.

Figure 3 depicts a cumulative chart of all of the effects produced from the modern set, with scores ordered from left to right with ability on the horizontal axis and average effect size on the vertical axis. After including the one study with the highest level of ability, the effect

was approximately −0.05. With the addition of the second study, the average effect was about 0.45. By the time approximately 20 studies had been included, the effect stabilized and once all but the lowest ability 10 studies were included, the estimate was 0.28. The addition of the last effects did indeed have a large impact, bringing the overall mean back up to 0.35. Eliminating the three lowest ability effects results in a mean estimate of the remaining 53 effects (n = 3,951) of 0.293 points per year, 95% CI [0.23, 0.35], and the regression of effect on ability is no longer significant. The other five studies that are part of the bimodal distribution in Figure 2 do not appear to have significant impact on the overall estimate.

Age—Effect size was regressed on the average age of each sample in the set of 53 effects (n = 3,951) retained in the ability analysis above. The regression of effect size on age was nonsignificant, accounting for less than one percent of the variance in effect sizes.

Sample type—Each modern study (k=53) was coded for sample type, which included clinical (k = 1, n = 24), research (k = 22, n = 902) and manuals (k = 30 n = 3,025). Because there was only 1 effect from a clinical sample, the moderator analysis was done on the remaining 52 effects. Although each group mean effect was significantly different from zero (Table 2), type of sample was not significant in the random effects analysis, $Q_{(1)} = 3.14$, $p < .076$.

Order effects—Table 3a summarizes estimated Flynn effects (random effects model) by test group for studies that were counterbalanced. The pattern of effect sizes paralleled the overall study results for each test group. For the modern tests, summarized in Table 3b, the estimate of 0.28 is close to the estimate of 0.29 for all 53 effects. Within the 53 modern effects, 50 provided information on test order. Most studies either uniformly gave the tests in the same order or counterbalanced so that half got the old test first and half got the new test first. The order effect was not significant in the random effects analysis, $Q_{(2)} = 4.30$ $p < .17$. The mean effects for the counterbalanced group (k = 30, n = 2,912) (M = 0.29, 95% CI [0.23, 0.36]) and the group of effects where the old test was given second (k = 8, n = 505) (M = 0.54, 95% CI [0.16, 0.91]) were significantly different from zero. The mean effect for the studies where the older test was given first (k = 12, n = 396) was not significantly different from zero (M = 0.14, 95% CI [−.04, 0.32]).

For the effects coded 100 where the old test was uniformly given first, negative effects due to prior exposure would be expected. In this ordering, Table 3b shows that prior exposure reduces the Flynn effect (.14 per year, n.s.). For effects coded 0, we would expect the mean effect to be amplified, reflecting a Flynn effect plus a prior exposure effect. Table 3b shows that the Flynn effect estimate is indeed larger (.54 per year). Finally, if the order was counterbalanced, the estimate should reflect the Flynn effect with less bias than either of the other two estimates. The estimate for the 30 counterbalanced groups is .29 per year. Although the order effect was not statistically significant, the estimates are different from 0 and the order test may not have been adequately powered. The patterns are consistent with hypothesis by Kaufman (2010).

Effect of pairing—Examining the counterbalanced tests permitted a comparison controlling for order effects when pairing Binet/Binet tests (k = 8, n = 545), Wechsler/Wechsler tests (k = 18, n = 2,023), and Wechsler/Binet tests (k = 4, n = 344). These comparisons yielded similar estimates close to the overall estimate of 0.293 per year: Binet/Binet: M = .291, 95% CI [0.14, 0.45]; Wechsler/Wechsler: M = 0.296, 95% CI [0.22, 0.38]; Wechsler/Binet: M = 0.292, 95% CI [0.17, 0.42].

Sensitivity Analysis

Finally, we explored the effect of our decisions on the results of the meta-analysis. First, the formula for the variance of each study included the sample-specific correlation between the two tests being compared in a given study. This correlation, however, is subject to sampling variance and to possible restriction of range within the sample studied. It is also potentially attenuated below the population correlation between the two tests if the administration is done in such a way as to affect the actual reliability of the tests as given. For example, test directions might be misunderstood or misread, the testing environment might introduce distractions, or there might be inaccuracies in scoring. As an alternative, we calculated the average r for each pair of tests by converting all observed correlations to Fisher's z and averaging within test pairs, or by using the overall r, as above, if the specific study was missing the correlation and there were no other studies with the same test pair. For the overall analyses and within the test groups, mean effects differed by no more than 0.03 points per year. All significance tests and tests of heterogeneity resulted in the same conclusions reached above.

In addition to the 285 effects analyzed above, there were an additional 93 effects with norming gaps of 5 years or less. The mean effect over the combined 378 studies in the random effects model was 0.28 IQ points per year, 95% CI [0.25, 0.31], z = 16.83, p < .0001. The effects were significantly heterogeneous, ($Q_{(377)}$ = 5581, p < .0001). The estimated I^2, or proportion of the total variance due to true study variance, was I^2 = .93, so very little of the variance in observed effects was attributable to sampling error or unreliability in the tests. The Tau, or estimated standard deviation of the true effects, was τ = 0.26, indicating a 95% credibility interval of −0.23 to +0.79. In other words, approximately 86% of the distribution of true effects was above zero. The funnel plot for the entire set of effects can be seen in Figure 4. Note that the 285 effects captured in Figure 1 comprise the tip of this pyramid. The range of standard errors in Figure 1 is from 0.0 to +0.6, whereas in Figure 4, the range is 0.0 to +20.0.

Discussion

Major Findings

The overall Flynn effect of 2.31 produced by this meta-analysis was lower than Flynn's (2009a) value of 3.11 and Fletcher et al.'s (2010) value of 2.80. It also fell below Dickinson and Hiscock's (2010) estimate of 2.60, which was the average of separate calculations for each of the 11 Wechsler subtests. However, our overall comparisons included all identified studies back to 1951. When a meta-analytic mean was calculated for the modern set (composed exclusively of 53 comparisons involving the Wechsler/Binet and excluding 3

atypical comparisons, and more comparable to the studies from Flynn [2009]), the Flynn effect was 2.93 points per decade, a value larger than estimates based on studies that included older data. This value is the most reasonable estimate of the Flynn effect for Wechsler/Binet tests normed since 1972 and is similar to the 3 points per decade rule of thumb commonly recommended in practice. The standard error of this estimate is less than 1 point (*SE* = 0.35).

Moderator Analyses

Ability level—Defined as the score produced by the most recently normed IQ test, ability level did not explain a significant amount of variance in the Flynn effect in the overall model. Although the literature has produced inconsistent evidence with regard to the direction and/or linearity of the relation between ability level and mean Flynn effect (Zhou et al., 2010; Lynn & Hampson, 1986; Teasdale & Owen, 1989; Graf & Hinton, 1994; Sanborn et al., 2003; Spitz, 1989), the present data revealed no relation between these two variables in the overall analysis. This finding may be the result of a methodological difference between our meta-analysis, which treated ability level as a continuous variable, and previous studies, many of which treated ability level as a categorical variable.

Within the set of modern tests, ability level did explain a significant amount of variance in the Flynn effect, with lower ability samples producing higher Flynn effects. However, this was not a clearly reliable finding. The distribution of effects at lower ability levels was bimodal, with a subsample of comparisons producing higher than anticipated Flynn effects and another subsample of comparisons producing lower than anticipated Flynn effects. When the three effects with the lowest level of ability were deleted, ability was no longer a significant predictor of effect size. Thus, estimating the magnitude of the Flynn effect in lower ability individuals, for whom testing may have the greatest ramifications, appears to be more complex than estimating the magnitude of the Flynn effect in the remainder of the ability distribution. As noted previously, the distribution of Flynn effects that we observed at lower ability levels might be the result of artifacts found in studies of groups within this range of ability. When studies were added one at a time, we obtained stability at about 0.27–0.30 points per year, with a mean of 0.293 points per year (excluding the three atypical low ability studies). These findings suggest that the mean magnitude of the Flynn effect may not change significantly with level of ability and that the correction can be applied to scores across the spectrum of ability level.

Age—Results revealed no difference in the Flynn effect based on participant age, suggesting that the Flynn effect is consistent across age cohorts. This finding is consistent with previous research (Flynn, 1984, 1987).

Sample type—Although the sample type effect was not statistically significant, it was based on a small number of effects and the means were different from zero, with the patterns showing lower Flynn effect estimates for test manual than research studies. We might expect for standardization samples to exercise the most control over variables related to participant selection, testing environment, and test administration procedures, so that the Flynn effect

increases as control over these variables is relaxed. Because the sample size constituting the clinical set is so small (k = 1, n = 24), future research with a larger set of studies is needed.

Order of test administration—Test order was not a statistically significant moderator. However, the number of effects per comparison was small and the patterns were consistent with hypotheses by Kaufman (2010). For all test sets that were counterbalanced, the Flynn effect estimates were similar in magnitude and pattern across test sets to the overall estimates. In the modern set, where order varied, the effect for counterbalanced administrations only (M = 0.293, k = 30, n = 2,912) was the same as the overall estimate for the full set of modern tests (M = 0.293, k = 53, n = 3,951, excluding the three atypical low ability studies), reflecting the fact that the bulk of the effects (k = 30) were derived from counterbalanced studies. However, if the new test was given first, the estimate (0.54) was larger, reflecting the additive effects of prior exposure and norms obsolescence. If the old test was given first, the estimate (0.14) was smaller, reflecting the opposing influences of prior exposure and norms obsolescence. Our data do not address Kaufman's (2010) more specific concern about asymmetric order effects such that taking the newer test first increased subsequent performance on the older test more than taking the older test first increases subsequent performance on the newer test. This putative pattern might be expected when the content or administration of an IQ test or subtest (e.g., Similarities subtest of the WISC-R) is changed in ways that could benefit a child who subsequently encounters the previous version of the same subtest. Given the variety of subtests underlying the IQ scores included in our meta-analyses, and the convergence of Flynn effect estimates around 0.29 for the modern tests, the order effect tends to be transitive with a mean magnitude of approximately ± .20. When the newer test is administered first, the Flynn effect estimate is approximately 0.29 + .20 and, when the older test is administered first, the Flynn effect estimate is approximately 0.35 – .20.

Pairing—Examining just the modern tests administered in a counterbalanced order and excluding the three atypical studies showed that the estimates for pairings of Wechsler/Wechsler, Binet/Binet, and Wechsler/Binet tests (all about 0.29) were remarkably similar to the overall estimate of 0.293 per year. These results suggest that similar corrections can be made to different versions of the Wechsler and Binet tests normed since 1972.

Implications of the Flynn Effect for Theory and Practice

Theory

Genetic hypotheses: As discussed above, there are multiple hypotheses about the basis of the Flynn effect, including genetic and environmental factors, and measurement issues. Although genetic hypotheses have not gained much tractability, they make predictions about relations with age and cohort that can be compared to these results. The larger Flynn estimate in our study for newer than older tests provides no compelling support for the heterosis hypothesis.

Environmental factors: Our finding that the Flynn effect has not diminished over time and may be larger for modern than older tests is not consistent with Sundet et al.'s (2008)

hypothesis relating increasing IQ scores and decreasing family size, although we do not have data for a direct evaluation.

The larger effect for modern than older tests could be regarded as consistent with Lynn's (2009) hypothesis pertaining to pre- and early postnatal nutrition. However, although we cannot directly address cohort effects in this meta-analysis, we note that the magnitude of increases in Wechsler and SB scores has remained close to the nominal value of 3 IQ points per decade since 1984 (Flynn, 2009). Deviations from this constant value--such as the difference we found between modern and old tests--might indicate an IQ difference between older and younger cohorts, but it also might reflect other differences that have occurred over time, such as scaling changes, ceiling effects, or differences in the sampling of study participants (e.g., Kaufman, 2010; Hiscock, 2007).

Our study did not find evidence for the plateauing or decline of the Flynn effect in the United States, as has been documented in Norway (Sundet et al., 2004) and Denmark (Teasdale & Owen, 2008; Teasdale & Owen, 2005), respectively. Table 5.6 in the WAIS-IV manual (Wechsler, 2008) summarizes an excellent planned comparison of the WAIS-III (standardized in 1995) and the WAIS-IV (standardized in 2005) scores administered in counterbalanced order to 240 examinees. This table shows results similar to our meta-analysis, with average WAIS-III scores about 3 points higher than WAIS-IV scores. In addition, the effect was similar across age and ability level cohorts. To the extent that the United States and Scandinavia differ on at least the variables proposed to be related to the plateauing of scores in Scandinavia (e.g., family life factors [Sundet et al., 2004] and educational priorities [Teasdale & Owen, 2008; Teasdale & Owen, 2005]), we might anticipate the difference in IQ score patterns noted. For example, Scandinavia's parental leave and subsidized childcare might be indices of optimal socioenvironmental conditions and are generous relative to the United States. With regard to educational priorities, the relative value of a liberal arts education persists in the United States.

Measurement issues: Different types of tests yield different estimates of the Flynn effect. The effects were most apparent for multifactorial tests like the Wechsler and Binet scales, and extend to other modern tests with the exception of the KABC, which yielded little evidence of a Flynn effect. This is surprising because the KABC minimizes the need for verbal responses, and Flynn effects tend to be relatively large for nonverbal tests such as the Wechsler Digit Symbol subtest (Dickinson & Hiscock, 2010). In addition, the variability of estimates for the KABC was very high, 95% CI [−0.16, +0.19], 95% credibility interval [−.90, +.93]. Mean estimates were negligible for screening tests, which is surprising because most screening tests include matrix problem-solving tests, which historically have yielded large estimates for norms obsolescence. Again, the variability is high, 95% CI [−0.15, +0.19], 95% credibility interval [−.63, +.66]). Altogether, these results suggest caution in estimating the degree of norms obsolescence for the KABC and different screening tests.

Practice

Assessment and decision-making: The results of this meta-analysis support the persistent findings of a significant and continuous elevation of IQ test norms as described by Flynn

(1984, 1987, 1998, 1999, 2007). The rate of change obtained from the overall model was somewhat less pronounced than the 3 IQ points per decade typically cited. Nevertheless, when only the modern Wechsler/Binet tests were considered in isolation, the magnitude of the effect appears to be close to 3 points per decade and showed no evidence of reducing in magnitude. Our support for a robust Flynn effect, manifested across various tests in nearly 300 studies, underscores the importance of considering this factor in high stakes decisions where the cut point on an IQ test is a salient criterion. These decisions include assessments for intellectual disability, which have implications for educational services received in schools, the death penalty, and financial assistance in cases where the individual is not competent to work.

Intellectual disability professionals have debated the necessity of correcting IQ scores for the Flynn effect in decisions about intellectual disability (e.g., Greenspan, 2006; Moore, 2006; Young, Boccaccini, Conroy, & Lawson, 2007). The present findings, which demonstrate the pervasiveness and stability of the Flynn effect across multiple tests and many decades, support the feasibility of correcting IQ according to the interval between norming and administration of the test, i.e., according to the degree to which the norms have become obsolete (Flynn, 2006a, 2009a). A precise correction, however, cannot be assured in all circumstances because the Flynn effect, as it applies to a given test, may strengthen or weaken at any time in the future. Moreover, the exact size of the Flynn effect may vary from one sample to another. Nonetheless, the rough approximation of 3 points per decade (plus or minus about 1 point based on the standard error and a 95% confidence interval) is consistent with the results of the modern studies in this meta-analysis.

Correction for the Flynn effect, although it increases the validity of the measured IQ (Flynn, 2006a, 2007, 2009a), does not justify using a conventional cut point as the sole criterion for determining intellectual disability (cf. Flynn & Widaman, 2008). In other words, increasing the validity of the measured IQ does not diminish the importance of other factors, including adaptive behavior. These include skills related to interpersonal effectiveness, activities of daily living, and the understanding of concepts such as money (AAIDD, 2010). Research has demonstrated a positive relation between IQ and measures of adaptive behavior (Schatz & Hamdan-Allen, 1995; Bolte & Poustka, 2002), and this supports the potential importance of considering both kinds of information when high stakes decisions must be made (Flynn & Widaman, 2008).

The results of this meta-analysis suggest that examiners be mindful about the particular tests administered in situations where an individual is retested to assess for progress and to determine the necessity of special education services. The significant Flynn effect means that, when individuals are tested near the release of a newly normed assessment, the difference in IQ scores produced by the newer test and the older test would indicate that the individual is performing more poorly than what earlier testing may have suggested. A critical implication was highlighted in a recent article by Kanaya and Ceci (2012), who observed that children administered the WISC-R during a special education assessment and administered the WISC-III during a reevaluation were less likely to be rediagnosed with a learning disorder than children administered the WISC-R on both occasions. Unawareness of the Flynn effect on the part of test examiners can compound this problem. For example,

Gregory and Gregory (1994) raised concerns that at the time of its publication, the Revised Neale Analysis of Reading Ability was producing lower scores than the older British Ability Scales (BAS) Word Reading scale. A critique of Gregory and Gregory's (1994) concerns by Halliwell and Feltham (1994) and possible explanations for the findings ensued, yet no mention of the possibility of norms obsolescence was presented. Our data show that norms obsolescence could have significant ramifications for the test results of students.

Further, in cases where an individual is assessed at two different sites (e.g., when a child moves and is assessed in a different school district), it may be possible for the child to have completed the newer version of a test first, especially if the assessments are occurring near to the release of a newly normed assessment. In this case, the IQ score produced by the second assessment may be particularly inflated due to both the Flynn effect and prior exposure. This child may be more likely to receive a diagnosis of a learning disability during this second assessment than a recommendation of special education services. This example underscores the importance of correcting for the Flynn effect in high stakes decisions, a directive consistent with AAIDD's (2010) recommendation, but addressed in few state special education standards for determining intellectual disability

Future research: The need for better estimates of the Flynn effect in research pertains to attempts to assess the breadth of the Flynn effect across cognitive domains. Several recent studies indicate that the Flynn effect is not limited to intelligence tests but may be measured in tests of memory (Baxendale, 2010; Rönnlund & Nilsson, 2008, 2009) and object naming (Connor, Spiro, Obler, & Martin, 2004), as well as certain commonly used neuropsychological tests (Dickinson & Hiscock, 2011). As Flynn effect estimates become more precise, it should be possible to differentiate not only the presence or absence of the effect but also gradations in the strength of the effect. Being able to quantify the magnitude of the Flynn effect in various domains would constitute an important advance toward answering the ultimate Flynn effect question, i.e., the underlying mechanism of the phenomenon.

From differences in the rates at which scores from the various Wechsler subtests have risen over time, Flynn (2007) has inferred characteristics of the intellectual skills that are rising rapidly and of the skills that are relatively static. We did not address this issue in this metaanalysis, partly because of the focus on the impact and precision of Flynn effect estimates for high stakes decisions across a range of tests and because the greater impact of the Flynn effect on fluid versus crystallized intelligence is well-established. More relevant would be additional knowledge about the strength of the Flynn effect on tests of memory and language and various neuropsychological tests, which would facilitate a more complete characterization of other higher mental functions that are susceptible to the Flynn effect in varying degrees. The data available from tests other than IQ tests are not likely to be sufficient in quality or quantity to yield precise Flynn effect estimates, but precise estimates for IQ tests will provide a reliable standard against which data from other tests can be evaluated.

Limitations

The objective of the current study was to build upon Flynn's (2009a) foundational work and Fletcher et al.'s (2010) meta-analytic study on the rate of IQ gain among modern Wechsler-Binet tests per test manual validation studies by expanding the scope of investigation to other tests, eras, and samples. As such, the approach to the current study replicates the method of Flynn (2009a) and Fletcher et al. (2010) by examining intragroup change in IQ score as a function of the norming date of the test. An alternate approach, taken by Flynn (1987) and others since (e.g., Sundet et al., 2004; Sundet et al., 2008) broadens the perspective from intragroup to intergroup change by focusing on draft board test performance within countries in the practice of administering IQ tests to all young men being assessed for suitability for conscription. For the study of a cohort phenomenon like the Flynn effect, this approach is appropriate. Unfortunately, no comparable data exist for American young men. Whereas the Raven's test administered to Scandinavian young men has not changed in format or content since its development, this is not the case for the Armed Services Vocational Aptitude Battery (arguably a measure of literacy rather than intelligence per se [Marks, 2010]) administered to potential conscripts in the United States. In addition, the data collected from Scandinavian young men, most of whom are evaluated for suitability for the armed services, are more representative of the Scandinavian population than potential conscripts in the United States who self-select into the armed services are of the American population.

There are drawbacks to studying the Flynn effect on the basis of IQ test validation studies per the method of Flynn (2009a) and Fletcher et al. (2010): sample sizes tend to be small; the earlier and later versions of the same test may differ significantly in format or content (e.g., Kaufman, 2010); there may be significant order effects; many tests are never re-normed and therefore lie beyond the reach of this method; and direct within-examinee comparisons have not been made for many tests even if the tests have been re-normed. In addition, validation studies rely on group-level data and presuppose a representative normative basis for the derivation of a standardized IQ score.

Even in the absence of speculation about the representativeness of a normative sample (see Flynn [2009] and Fletcher et al. [2010] for a discussion of the representativeness of the WAIS-III normative sample), normative sample sizes are significantly reduced once stratified by age. For example, 2,200 children constituted the WISC-IV standardization sample, from which were derived norms for subsets of 11 age groups. Similarly, 4,800 individuals constituted the SB5 standardization sample, from which were derived norms for subsets of 23 age groups.

Our alternative method involves relating mean scores on a test to the interval between norming and testing. This third method is capable of detecting changes in test performance over time without the need to track scores over many years or to restrict our analysis to tests for which repeated- measures data have been collected by test publishers. Our method is not as direct as Flynn's tracking of raw scores on Raven's Matrices, nor does it provide the detailed information that can be obtained by comparing old and new versions of the Wechsler and Stanford-Binet batteries in the same individuals. On the other hand, our method has the advantage of being applicable to a very large number of informative

samples. Our study not only confirms the findings for the Wechsler and Stanford-Binet tests that were obtained using the second method, but it also expands those findings to include numerous tests on which the Flynn effect could not otherwise be assessed. The results show that the IQ increase is pervasive, not only with respect to geography and time, but also with respect to the tests used to measure IQ. Our findings also suggest that the typical 6 IQ points per decade rise in Raven's Matrices score is unrepresentative of the Flynn effect magnitude measured with most other tests. Most of the tests included in our meta-analysis show rates of increase that are comparable to those measured for the Wechsler and Stanford-Binet batteries. Additionally, the large number of studies included in our meta-analysis provides a strong empirical basis for concluding that comparable IQ increases are evident in samples ranging from preschool children to elderly adults.

Relying on one numerical value to represent a continuous variable, including IQ score and age, results in a significant loss of information. For example, mean values can be greatly influenced by the number and magnitude of extreme values such that the resulting value may not be an adequate measure of central tendency nor an effective illustration of the relation between IQ score and the moderators assessed. Nonetheless, because the correction for the Flynn effect is not a correction to an individual score, but to the normative basis to which individual scores are compared, concerns about applying group data to individual scores do not really apply (Flynn, 2006a).

The usefulness of a meta-analysis depends to a great extent on the accessibility of studies meeting inclusion criteria. Although a thorough review was conducted on PsycINFO® and in test manuals, possibly there were studies meeting inclusion criteria that were not accessed. However, the number of comparisons included in this review appears more than sufficient to assess the magnitude of the Flynn effect and the precision of the obtained value, and to address the additional research questions under consideration. Further, there was no dearth of effect sizes at the lower end of the distribution of effect sizes (Figure 1), which suggests there was no oversampling of studies producing higher Flynn effects.

The homogeneity analysis indicated that there were sources of substantial heterogeneity among the studies included in the meta-analysis. In fact, 91% of the variance in the Flynn effect was due to true variance among studies. The selected moderator variables explained small amounts of the true variance in the modern set, suggesting that additional factors that explain variance in the Flynn effect have yet to be identified.

Conclusions

For the present, the need to correct IQ test scores for norms obsolescence in high stakes decision-making is abundantly clear. At average levels of IQ, a score difference of 95 and 98 is not critical. However, in capital punishment cases, life and death may reside on a 3-point difference of 76 versus 73, or 71 versus 68. This becomes especially important when comparing IQ test scores across a broad period of time and when IQ test scores obtained in childhood are brought to bear on an adult obtained score. Correcting for norms obsolescence is a form of scaling to the same standard. Weight standards often are adjusted each decade because people get larger over time. For these changes, the critical decision points are changed for obesity. For intellectually disability, we could (in theory) use the same test over

time. Thus, if a child were assessed in 2013 with the WISC-R standardized in 1973, we could adjust the mean to 109 (SD = 15) and the cut point for intellectual disability to 79 (3 points). Because the convention in our society is to use a cut point of 70, corrections for norms obsolescence, i.e., the Flynn effect, must be made.

The existence of unknown factors that influence the Flynn effect should not obscure the major findings of this study: the mean value of the Flynn effect within the modern set centered around 3 points per decade, most of the estimated distribution of true effects was larger than zero, and the standard error of this estimate is 0.35 (resulting in a 95% CI that extends about .7, rounded to 1 point, on either side of 3 points per decade). These findings are consistent with previous research and with the argument that it is feasible and advisable to correct IQ scores for the Flynn effect in high stakes decisions.

References

Agbayani KA, Hiscock M. Age-related change in Wechsler IQ norms after adjustment for the Flynn effect: Estimates from three computational models. Journal of Clinical and Experimental Neuropsychology. 2013; 35(6):642–654. [PubMed: 23767697]

American Association on Intellectual and Developmental Disabilities. Intellectual disability: Definition, classification, and systems of supports. Washington DC: American Association on Intellectual and Developmental Disabilities; 2010.

American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, fourth edition, text revision (DSM-IV-TR). Washington, DC: American Psychiatric Association; 2000.

*. Appelbaum AS, Tuma JM. Social class and test performance: Comparative validity of the Peabody with the WISC and WISC-R for two socioeconomic groups. Psychological Reports. 1977; 40:139–145.

*. Arffa S, Rider LH, Cummings JA. A validity study of the Woodcock-Johnson Psycho-Educational Battery and the Stanford-Binet with black preschool children. Journal of Psychoeducational Assessment. 1984; 2:73–77.

*. Arinoldo CG. Concurrent validity of McCarthy's Scales. Perceptual and Motor Skills. 1982; 54:1343–1346. [PubMed: 7110874]

*. Arnold FC, Wagner WK. A comparison of Wechsler children's scale and Stanford-Binet scores for eight-and nine year olds. The Journal of Experimental Education. 1955; 24(1):91–94.

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

*. Axelrod BN. Validity of the Wechsler Abbreviated Scale of Intelligence and other very short forms estimating intellectual functioning. Assessment. 2002; 9:17–23. [PubMed: 11911230]

*. Axelrod BN, Naugle RI. Evaluation of two brief and reliable estimates of the WAIS-R. International Journal of Neuroscience. 1998; 94:85–91. [PubMed: 9622802]

*. Barclay A, Yater AC. Comparative study of the Wechsler Preschool and Primary Scale of Intelligence and the Stanford-Binet Intelligence Scale, Form L-M, among culturally deprived children. Journal of Consulting and Clinical Psychology. 1969; 33(2):257. [PubMed: 5783267]

*. Barratt ES, Baumgarten DL. The relationship of the WISC and Stanford-Binet to school achievement. Journal of Consulting Psychology. 1957; 21(2):144. [PubMed: 13416433]

Baxendale S. The Flynn effect and memory function. Journal of Clinical and Experimental Neuropsychology. 2010; 32:699–703. [PubMed: 20119877]

Beaujean AA, Osterlind SJ. Using item response theory to assess the Flynn effect in the National Longitudinal Study of Youth 79 Children and Young Adults data. Intelligence. 2008; 36:455–463.

Beaujean AA, Sheng Y. Examining the Flynn effect in the General Social Survey Vocabulary test using item response theory. Personality and Individual Differences. 2010; 48:294–298.

Bhuvaneswar CG, Chang G, Epstein LA, Stern T. Alcohol use during pregnancy: Prevalence and impact. The Primary Care Companion to the Journal of Clinical Psychiatry. 2007; 9(6):455–460.

Blume J. Defendants whose death penalties have been reduced because of a finding of "mental retardation" since. Atkins v. Virginia. 2008 (2002). Retrieved from http://www.deathpenaltyinfo.org/sentence-reversals-intellectual-disability-cases.

Bolte S, Poustka F. The relation between general cognitive level and adaptive behavior domains in individuals with autism with and without co-morbid mental retardation. Child Psychiatry and Human Development. 2002; 33(2):165–172. [PubMed: 12462353]

Borenstein, M.; Hedges, L.; Higgins, J.; Rothstein, H. Comprehensive Meta-analysis Version 2. Englewood NJ: Biostat; 2005.

Borenstein, M.; Hedges, LV.; Higgins, JPT.; Rothstein, HR. Introduction to Meta-Analysis. United Kingdom: John Wiley & Sons, Ltd; 2009.

*. Bower A, Hayes A. Relations of scores on the Stanford Binet Fourth Edition and Form L-M: Concurrent validation study with children who have mental retardation. American Journal on Mental Retardation. 1995; 99(5):555–563. [PubMed: 7779350]

*. Bracken BA, Prasse DP, Breen MJ. Concurrent validity of the Woodcock-Johnson Psycho-Educational Battery with regular and learning-disabled students. Journal of School Psychology. 1984; 22:185–192.

*. Bradway KP, Thompson CW. Intelligence at adulthood: A twenty-five year follow-up. Journal of Educational Psychology. 1962; 53(1):1–14.

*. Brengelmann JC, Renny JT. Comparison of Leiter, WAIS, and Stanford-Binet IQ's in retardates. Journal of Clinical Psychology. 1961; 17(3):235–238.

Brand CR. Bryter still and bryter? Nature. 1987; 328:110. [PubMed: 3600785]

*. Brooks CR. WISC, WISC-R, S-B L & M, WRAT: Relationships and trends among children ages six to ten referred for psychological evaluation. Psychology in the Schools. 1977; 14:30–33.

*. Byrd PD, Buckhalt JA. A multitrait-multimethod construct validity study of the Differential Ability Scales. Journal of Psychoeducational Assessment. 1991; 9:121–129.

*. Carvajal H, Gerber J, Hewes P, Weaver KA. Correlations between scores on Stanford-Binet IV and Wechsler Adult Intelligence Scale-Revised. Psychological Reports. 1987; 61(1):83–86.

*. Carvajal H, Hardy K, Smith KL, Weaver KA. Relationships between scores on Stanford-Binet IV and Wechsler Preschool and Primary Scale of Intelligence. Psychology in the Schools. 1988; 25(2):129–131.

*. Carvajal HH, Hayes JE, Lackey KL, Rathke ML, Wiebe DA, Weaver KA. Correlations between scores on the Wechsler Intelligence Scale for Children-III and the General Purpose Abbreviated Battery of the Stanford-Binet IV. Psychological Reports. 1993; 72(3):1167–1170. [PubMed: 8337322]

*. Carvajal H, Karr SK, Hardy KM, Palmer BL. Relationships between scores on Stanford-Binet IV and scores on McCarthy's Scales of Children's Abilities. Bulletin of the Psychonomic Society. 1988; 26(4):349.

*. Carvajal HH, Parks JP, Bays KJ, Logan RA, Lujano CI, Page GL, Weaver KA. Relationships between scores on the Wechsler Preschool and Primary Scale of Intelligence – Revised and Stanford-Binet IV. Psychological Reports. 1991; 69(1):23–26. [PubMed: 1961799]

*. Carvajal H, Weyand K. Relationships between scores on Stanford-Binet IV and Wechsler Intelligence Scale for Children-Revised. Psychological Reports. 1986; 59(2):963–966. [PubMed: 3809352]

*. Chelune GJ, Eversole C, Kane M, Talbott R. WAIS versus WAIS-R subtest patterns: A problem of generalization. The Clinical Neuropsychologist. 1987; 1(3):235–242.

*. Clark RD, Wortman S, Warnock S, Swerdlik M. A correlational study of Form L-M and the 4th Edition of the Stanford-Binet with 3- to 6-year olds. Diagnostique. 1987; 12(2):118–120.

*. Cohen BD, Collier MJ. A note on the WISC and other tests of children six and eight years old. Journal of Consulting Psychology. 1952; 16(3):226–227. [PubMed: 14946292]

*. Coleman MC, Harmer WR. The WISC-R and Woodcock-Johnson Tests of Cognitive Ability: A comparative study. Psychology in the Schools. 1985; 22:127–132.

Connor LT, Spiro A III, Obler LK, Albert ML. Change in object naming ability during adulthood. Journals of Gerontology: Series B: Psychological Sciences and Social Sciences. 2004; 59B:P203–P209.

*. Covin TM. Comparability of WISC and WISC-R scores for 38 8- and 9-year-old institutionalized Caucasian children. Psychological Reports. 1977; 40:382. [PubMed: 859966]

*. Craft NP, Kronenberger EJ. Comparability of WISC-R and WAIS IQ scores in educable mentally handicapped adolescents. Psychology in the Schools. 1979; 16(4):502–504.

Cumming G, Finch S. Inference by eye: Confidence intervals and how to read pictures of data. American Psychologist. 2005; 60:170–180. [PubMed: 15740449]

*. Davis EE. Concurrent validity of the McCarthy Scales of Children's Abilities. Measurement and Evaluation in Guidance. 1975; 8:101–104.

*. Davis EE, Walker C. McCarthy Scales and WISC-R. Perceptual and Motor Skills. 1977; 44:966.

Dickens WT, Flynn JR. Great leap forward: A new theory of intelligence. New Scientist. 2001a Apr 21.:44–47.

Dickens WT, Flynn JR. Heritability estimates versus large environmental effects: The IQ paradox resolved. Psychological Review. 2001b; 108:346–369. [PubMed: 11381833]

Dickinson MD, Hiscock M. Age-related IQ decline is reduced markedly after adjustment for the Flynn effect. Journal of Clinical and Experimental Neuropsychology. 2010; 32:865–870. [PubMed: 20349385]

Dickinson MD, Hiscock M. The Flynn effect in neuropsychological assessment. Applied Neuropsychology. 2011; 18:136–142. [PubMed: 21660765]

*. Doll B, Boren R. Performance of severely language-impaired students on the WISC-III, language scales, and academic achievement measures. Journal of Psychoeducational Assessment, Advances in Psychological Assessment Monograph Series. 1993:77–86.

*. Dumont R, Willis JO, Farr LP, McCarthy T, Price L. The relationship between the Differential Ability Scales (DAS) and the Woodcock-Johnson Tests of Cognitive Ability-Revised (WJ-R COG) for students referred for special education evaluations. Journal of Psychoeducational Assessment. 2000; 18:27–38.

*. Edwards BR, Klein M. Comparison of the WAIS and the WAIS-R with Ss of high intelligence. Journal of Clinical Psychology. 1984; 40(1):300–302.

*. Eisenstein N, Engelhart CI. Comparison of the K-BIT with short forms of the WAIS-R in a neuropsychological population. Psychological Assessment. 1997; 9(1):57–62.

Elley WB. Changes in mental ability in New Zealand. New Zealand Journal of Educational Studies. 1969; 4:140–155.

*. Elliot, CD. Differential ability scales. San Diego, CA: Harcourt Brace Jovanovich; 1990.

*. Elliot, CD. Differential Ability Scales, Second Edition, Technical manual. San Antonio, TX: The Psychological Corporation; 2007.

*. Estabrook GE. A canonical correlation analysis of the Wechsler Intelligence Scale for Children-Revised and the Woodcock-Johnson Tests of Cognitive Ability in a sample referred for suspected learning disabilities. Journal of Educational Psychology. 1984; 76(6):1170–1177.

*. Fagan J, Broughton E, Allen M, Clark B, Emerson P. Comparison of the Binet and WPPSI with lower-class five-year-olds. Journal of Consulting and Clinical Psychology. 1969; 33(5):607–609.

*. Faust DS, Hollingsworth JO. Concurrent validation of the Wechsler Preschool and Primary Scale of Intelligence-Revised (WPPSI-R) with two criteria of cognitive abilities. Journal of Psychoeducational Assessment. 1991; 9:224–229.

*. Field GE, Sisley RC. IQ score differences between the WAIS and the WAIS-R: Confirmation with a New Zealand sample. Journal of Clinical Psychology. 1986; 42(6):986–988.

Fletcher, JM.; Lyon, GR.; Fuchs, LS.; Barnes, MA. Learning disabilities: From identification to intervention. New York, NY: The Guilford Press; 2007.

Fletcher JM, Stuebing KK, Hughes LC. IQ scores should be corrected for the Flynn effect in high stakes decisions. Journal of Psychoeducational Assessment. 2010; 28(5):469–473.

Flynn JR. The mean IQ of Americans: Massive gains 1932–1978. Psychological Bulletin. 1984a; 95(1):29–51.

Flynn JR. IQ gains and the Binet decrements. Journal of Educational Measurement. 1984b; 21(3):283–290.

Flynn JR. Wechsler intelligence tests: Do we really have a criterion of mental retardation? American Journal of Mental Deficiency. 1985; 90(3):236–244. [PubMed: 4083304]

Flynn JR. Massive IQ gains in 14 nations: What IQ tests really measure. Psychological Bulletin. 1987; 101(2):171–191.

Flynn JR. Massive IQ gains on the Scottish WISC: Evidence against Brand et al.'s hypothesis. Irish Journal of Psychology. 1990; 11(1):41–51.

Flynn, JR. IQ gains over time. In: Sternberg, RJ., editor. The encyclopedia of human intelligence. New York: Macmillan; 1994. p. 617-623.

Flynn JR. WAIS-III and WISC-III gains in the United States from 1972–1995: How to compensate for obsolete norms. Perceptual and Motor Skills. 1998a; 86:1231–1239.

Flynn JR. Israeli military IQ tests: Gender differences small; IQ gains large. Journal of Biosocial Science. 1998b; 30:541–553. [PubMed: 9818560]

Flynn JR. Searching for justice: The discovery of IQ gains over time. American Psychologist. 1999; 54:5–20.

Flynn JR. The hidden history of IQ and special education: Can the problems be solved? Psychology, Public Policy, and Law. 2000a; 6:191–198.

Flynn JR. IQ gains, WISC subtests and fluid *g*: *g* theory and the relevance of Spearman's hypothesis to race. Novartis Foundation Symposium. 2000b; 233:202–216. [PubMed: 11276904]

Flynn JR. Tethering the elephant: Capital cases IQ, and the Flynn effect. Psychology, Public Policy, and Law. 2006a; 12:170–189.

Flynn, JR. Efeito Flynn: Repensando a inteligência e seus efeitos [The Flynn effect: Rethinking intelligence and what affects it]. In: Flores-Mendoza, C.; Colom, R., editors. Introdução à psicologia das diferenças individuais [Introduction to the psychology of individual differences]. Porto Alegre, Brasil: ArtMed; 2006b. p. 387-411.(English trans. available from jim.flynn@stonebow.otago.ac.nz).

Flynn, JR. What is intelligence?. Cambridge: Cambridge University Press; 2007.

Flynn JR. The WAIS-III and WAIS-IV: Daubert motions favor the certainly false over the approximately true. Applied Neuropsychology. 2009a; 16:98–104. [PubMed: 19430991]

Flynn JR. Requiem for nutrition as the cause of IQ gains: Raven's gains in Britain 1938–2008. Economics and Human Biology. 2009b; 7:18–27. [PubMed: 19251490]

Flynn JR. Problems with IQ gains: The huge Vocabulary gap. Journal of Psychoeducational Assessment. 2010; 28(5):412–433.

Flynn, JR. Are we getting smarter? Rising IQ in the twenty-first century. Cambridge: Cambridge University Press; 2012.

Flynn JR, Weiss LG. American IQ gains from 1932 to 2002: The WISC subtests and educational progress. International Journal of Testing. 2007; 7(2):209–224.

Flynn JR, Widaman KF. 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. 2008; 35:121–149.

*. Fourqurean JM. A K-ABC and WISC-R comparison for Latino learning-disabled children of limited English proficiency. Journal of School Psychology. 1987; 25(1):15–21.

*. Frandsen AN, Higginson JB. The Stanford-Binet and the Wechsler Intelligence Scale for Children. Journal of Consulting Psychology. 1951; 15(3):236–238. [PubMed: 14841297]

*. Gehman IH, Matyas RP. Stability of the WISC and Binet tests. Journal of Consulting Psychology. 1956; 20(2):150–152. [PubMed: 13306847]

*. Gerken KC, Hodapp AF. Assessment of preschoolers at-risk with the WPPSI-R and the Stanford-Binet L-M. Psychological Reports. 1992; 71:659–664. [PubMed: 1410125]

*. Giannell AS, Freeburne CM. The comparative validity of the WAIS and the Stanford-Binet with college freshmen. Educational and Psychological Measurement. 1963; 23(3):557–567.

*. Gordon S, Duff S, Davidson T, Whitaker S. Comparison of the WAIS-III and WISC-IV in 16-year-old special education students. Journal of Applied Research in Intellectual Disabilities. 2010; 23:197–200.

Graf MH, Hinton RN. A 3-year comparison study of WISC-R and WISC-III IQ scores for a sample of special education students. Educational and Psychological Measurement. 1994; 14:128–133.

Greenspan S. Issues in the use of the “Flynn Effect” to adjust IQ scores when diagnosing MR. Psychology. Mental Retardation and Developmental Disabilities. 2006; 31:3–7. doi:

Greenspan, S.; Switzky, HN. Lessons learned from the Atkins decision in the next AAMR *Manual*. In: Switzky, HN.; Greenspan, S., editors. What is mental retardation? Ideas for an evolving disability in the 21st century. Washington, DC: American Association on Mental Retardation; 2006. p. 283-302.

*. Gregg N, Hoy C. A comparison of the WAIS-R and the Woodcock-Johnson Tests of Cognitive Ability with learning-disabled college students. Journal of Psychoeducational Assessment. 1985; 3:267–274.

Gregory HM, Gregory AH. A comparison of the Neale and the BAS reading tests. Educational Psychology in Practice. 1994; 10:15–18.

*. Gunter CM, Sapp GL, Green AC. Comparison of scores on WISC-III and WISC-R of urban learning disabled students. Psychological Reports. 1995; 77(2):473–474. [PubMed: 8559871]

Hagen LD, Drogin EY, Guilmette TJ. Adjusting IQ scores for the Flynn effect: Consistent with the standards of practice? Professional Psychology: Research and Practice. 2008; 39(6):619–625.

Halliwell M, Feltham R. Comparing the Neale and BAS reading tests: A reply to Gregory and Gregory. Educational Psychology in Practice. 1995; 10(4):228–230.

*. Hamm H, Wheeler J, McCallum S, Herrin M, Hunter D, Catoe C. A comparison between the WISC and WISC-R among educable mentally retarded children. Psychology in the Schools. 1976; 13:4–8.

*. Hannon JE, Kicklighter R. WAIS versus WISC in adolescents. Journal of Consulting and Clinical Psychology. 1970; 35(2):179–182.

*. Harrington RG, Kimbrell J, Dai X. The relationship between the Woodcock-Johnson Psycho-Educational Battery-Revised (Early Development) and the Wechsler Preschool and Primary Scale of Intelligence-Revised. Psychology in the Schools. 1992; 29(2):116–125.

*. Hartlage LC, Boone KE. Achievement test correlates of Wechsler Intelligence Scale for Children and Wechsler Intelligence Scale for Children-Revised. Perceptual and Motor Skills. 1977; 45:1283–1286.

*. Hartwig SS, Sapp GI, Clayton GA. Comparison of the Stanford-Binet Intelligence Scale: Form L-M and the Stanford-Binet Intelligence Scale Fourth Edition. Psychological Reports. 1987; 60(3): 1215–1218.

*. Hayden DC, Furlong MJ, Linnemeyer S. A comparison of the Kaufman Assessment Battery for Children and the Stanford-Binet IV for the assessment of gifted children. Psychology in the Schools. 1988; 25(3):239–243.

*. Hays JR, Reas DH, Shaw JB. Concurrent validity of the Wechsler Abbreviated Scale of Intelligence and the Kaufman Brief Intelligence Test among psychiatric inpatients. Psychological Reports. 2002; 90(2):335–359.

*. Hendershott J, Searight HR, Hatfield JI, Rogers BJ. Correlations between the Stanford-Binet, Fourth Edition and the Kaufman Assessment Battery for Children for a preschool sample. Perceptual and Motor Skills. 1990; 71(3):819–825.

Hiscock M. The Flynn effect and its relevance to neuropsychology. Journal of Clinical and Experimental Neuropsychology. 2007; 29(5):514–529. [PubMed: 17564917]

*. Holland GA. A comparison of the WISC and Stanford-Binet IQ’s of normal children. Journal of Consulting Psychology. 1953; 17(2):147–152. [PubMed: 13044892]

*. Ingram GF, Hakari LJ. Validity of the Woodcock-Johnson Tests of Cognitive Ability for gifted children: A comparison study with the WISC-R. Journal for the Education of the Gifted. 1985; 9(1):11–23.

*. Ipsen SM, McMillan JH, Fallen NH. An investigation of the reported discrepancy between the Woodcock-Johnson Tests of Cognitive Ability and the Wechsler Intelligence Scale for Children-Revised. Diagnostique. 1983; 9:32–44.

Jensen, AR. The g Factor. Westport, CT: Praeger; 1998.

*. Jones S. The Wechsler Intelligence Scale for Children applied to a sample of London primary school children. British Journal of Educational Psychology. 1962; 32(2):119–133.

Kanaya T, Ceci SJ. Are all IQ scores created equal? The differential costs of IQ cutoff scores for at-risk children. Child Development Perspective. 2007; 1(1):52–56.

Kanaya T, Ceci SJ, Scullin MH. The rise and fall of IQ in special ed: Historical trends and their implications. Journal of School Psychology. 2003a; 41:453–465.

Kanaya T, Scullin MH, Ceci SJ. The Flynn effect and U.S. policies: The impact of rising IQ scores on American society via mental retardation diagnoses. American Psychologist. 2003b; 58(10):778–790. [PubMed: 14584994]

Kane H, Oakland TD. Secular declines in Spearman's *g*: Some evidence from the United States. The Journal of Genetic Psychology. 2000; 16(3):337–345. [PubMed: 10971912]

*. Kangas J, Bradway K. Intelligence at middle age: A thirty-eight-year follow-up. Developmental Psychology. 1971; 5(2):333–337.

*. Kaplan CH, Fox LM, Paxton L. Bright children and the Revised WPPSI: Concurrent validity. Journal of Psychoeducational Assessment. 1991; 9:240–246.

*. Karr SK, Carvajal H, Elser D. Concurrent validity of the WPPSI-R and the McCarthy Scales of Children's Abilities. Psychological Reports. 1993; 72(3):940–942.

*. Karr SK, Carvajal H, Palmer BL. Comparison of Kaufman's short form of the McCarthy Scales of Children's Abilities and the Stanford-Binet Intelligence Scales – Fourth Edition. Perceptual and Motor Skills. 1992; 74(3):1120–1122. [PubMed: 1501979]

Kaufman AS. "In what way are apples and oranges alike?" A critique of Flynn's interpretation of the Flynn Effect. Journal of Psychoeducational Assessment. 2010; 28(5):382–398.

*. Kaufman, AS.; Kaufman, NL. Kaufman Assessment Battery for Children. Circle Pines, MN: American Guidance Service; 1983.

*. Kaufman, AS.; Kaufman, NL. Kaufman Brief Intelligence Test. Circle Pines, MN: American Guidance Service; 1990.

*. Kaufman, AS.; Kaufman, NL. Kaufman Adolescent and Adult Intelligence Test. Sydney, Australia: PsychCorp; 1993.

*. Kaufman, AS.; Kaufman, NL. Kaufman Assessment Battery for Children, Second Edition, Manual. San Antonio, TX: The Psychological Corporation; 2004a.

*. Kaufman, AS.; Kaufman, NL. Kaufman Brief Intelligence Test, Second Edition, Manual. San Antonio, TX: The Psychological Corporation; 2004b.

*. King JD, Smith RA. Abbreviated forms of the Wechsler Preschool and Primary Scale of Intelligence for a kindergarten population. Psychological Reports. 1972; 30:539–542.

*. Klanderman J, Devine J, Mollner C. The K-ABC: A construct validity study with the WISC-R and Stanford-Binet. Journal of Clinical Psychology. 1985; 41(2):273–281.

*. Klinge V, Rodziewicz T, Schwartz L. Comparison of the WISC and WISC-R on a psychiatric adolescent inpatient sample. Journal of Abnormal Psychology. 1976; 4(1):73–81.

*. Knight BC, Baker BH, Minder CC. Concurrent validity of the Stanford-Binet: Fourth Edition and the Kaufman Assessment Battery for Children with learning-disabled students. Psychology in the Schools. 1990; 27(2):116–120.

*. Krohn EJ, Lamp RE. Concurrent validity of the Stanford-Binet Fourth Edition and K-ABC for Head Start children. Journal of School Psychology. 1989; 27(1):59–67.

*. Krohn EJ, Lamp RE, Phelps CG. Validity of the K-ABC for a black preschool population. Psychology in the Schools. 1988; 25(1):15–21.

*. Krohn EJ, Traxler AJ. Relationship of the McCarthy Scales of Children's Abilities to other measures of preschool cognitive, motor, and perceptual development. Perceptual and Motor Skills. 1979; 49:783–790.

*. Krugman JI, Justman J, Wrightstone JW, Krugman M. Pupil functioning on the Stanford-Binet and the Wechsler Intelligence Scale for Children. Journal of Consulting Psychology. 1951; 15(6): 475–483. [PubMed: 14888753]

*. Kureth G, Muhr JP, Weisgerber CA. Some data on the validity of the Wechsler Intelligence Scale for Children. Child Development. 1952; 23(4):281–287. doi: [PubMed: 13042894]

*. Lamp RE, Krohn EJ. A longitudinal predictive validity investigation of the SB:FE and K-ABC with at-risk children. Journal of Psychoeducational Measurement. 2001; 19:334–349.

*. Larrabee GJ, Holroyd RG. Comparison of WISC and WISC-R using a sample of highly intelligent children. Psychological Reports. 1976; 38:1071–1074.

*. Lavin C. The Wechsler Intelligence Scale for Children – Third Edition and the Stanford-Binet Intelligence Scale: Fourth Edition: A preliminary study of validity. Psychological Reports. 1996; 78(2):491–496.

*. Law JG, Faison L. WISC-III and KAIT results in adolescent delinquent males. Journal of Clinical Psychology. 1996; 52(6):699–703. [PubMed: 8912113]

*. Levinson BM. A comparative study of the verbal and performance ability of monolingual and bilingual native born Jewish preschool children of traditional parentage. Journal of Genetic Psychology. 1960; 97:93–112. [PubMed: 14416301]

*. Levinson BM. A comparison of the performance of bilingual and monolingual native born Jewish preschool children of traditional parentage on four intelligence tests. Journal of Clinical Psychology. 1959; 15(1):74–76. [PubMed: 13611073]

*. Lippold S, Claiborn JM. Comparison of the Wechsler Adult Intelligence Scale and the Wechsler Adult Intelligence Scale-Revised. Journal of Consulting and Clinical Psychology. 1983; 51(2): 315. [PubMed: 6841778]

Lipsey, MW.; Wilson, DB. Practical meta-analysis. Thousand Oaks, CA: Sage Publications, Inc.; 2001.

*. Lukens J, Hurrell RM. A comparison of the Stanford-Binet IV and the WISC-III with mildly mentally retarded children. Psychology in the Schools. 1996; 33:24–27.

Lynn R. The role of nutrition in secular increases in intelligence. Personality and Individual Differences. 1990; 11:273–285.

Lynn R. What has caused the Flynn effect? Secular increases in the Development Quotients of infants. Intelligence. 2009; 37:16–24.

Lynn R, Hampson S. The rise of national intelligence: Evidence from Britain, Japan, and the U.S.A. Personality and Individual Differences. 1986; 7:23–32.

Marks DF. IQ variations across time, race, and nationality: An artifact of differences in literacy skills. Psychological Reports. 2010; 106:643–664. [PubMed: 20712152]

*. McCarthy, D. Manual: McCarthy Scales of Children's Abilities. San Antonio, TX: The Psychological Corporation; 1972.

*. McCrowell KL, Nagle RJ. Comparability of the WPPSI-R and the S-B:IV among preschool children. Journal of Psychoeducational Assessment. 1994; 12:126–134.

*. McGinley P. A comparison of WISC and WISC-R test results. The Irish Journal of Psychology. 1981; 1:23–24. doi:

*. McKerracher DW, Scott J. I.Qscores and the problem of classification: A comparison of the W.A.I.S. and S-B, Form L-M in a group of subnormal and psychopathic patients. British Journal of Psychiatry. 1966; 112:537–541. [PubMed: 5964257]

*. Milrod RJ, Rescorla L. A comparison of the WPPSI-R and WPPSI with high-IQ children. Journal of Psychoeducational Assessment. 1991; 9:255–262.

Mingroni MA. Resolving the IQ paradox: Heterosis as a cause of the Flynn effect and other trends. Psychological Bulletin. 2007; 114(3):806–829.

*. Mishra SP, Brown KH. The comparability of WAIS and WAIS-R IQs and subtest scores. Journal of Clinical Psychology. 1983; 39(5):754–757. [PubMed: 6630552]

*. Mitchell RE, Grandy TG, Lupo JV. Comparison of the WAIS and the WAIS-R in the upper ranges of IQ. Professional Psychology: Research and Practice. 1986; 17(1):82–83.

Moore RB Jr. Modification of individual IQ scores is not accepted professional practice. Psychology in Mental Retardation and Developmental Disabilities. 2006; 32(2)

*. Munford PR. A comparison of the WISC and WISC-R on black psychiatric outpatients. Journal of Clinical Psychology. 1978; 34(4):938–943. [PubMed: 711888]

*. Munford PR, Munoz A. A comparison of the WISC and WISC-R on Hispanic children. Journal of Clinical Psychology. 1980; 36(2):452–457.

*. Nagle RJ, Lazarus SC. The comparability of the WISC-R and WAIS among 16-year-old EMR children. Journal of School Psychology. 1979; 17(4):362–367.

*. Naglieri JA. Concurrent and predictive validity of the Kaufman Assessment Battery for Children with a Navajo sample. Journal of School Psychology. 1984; 22(4):373–380.

*. Naglieri JA. Normal children's performance on the McCarthy Scales, Kaufman Assessment Battery, and Peabody Individual Achievement Test. Journal of Psychoeducational Assessment. 1985; 3:123–129.

*. Naglieri JA, Harrison PL. Comparison of McCarthy General Cognitive Indexes and Stanford-Binet IQs for educable mentally retarded children. Perceptual and Motor Skills. 1979; 48:1251–1254. [PubMed: 492899]

*. Naglieri JA, Jensen AR. Comparison of black-white differences on the WISC-R and the K-ABC: Spearman's hypothesis. Intelligence. 1987; 11(1):21–43.

*. Naugle RI, Chelune GJ, Tucker GD. Validity of the Kaufman Brief Intelligence Test. Psychological Assessment. 1993; 5(2):182–186.

Neisser U. Rising scores on intelligence tests. American Scientist. 1997; 85(5):440–447.

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

*. Nelson WM III, Dacey CM. Validity of the Stanford-Binet Intelligence Scale-IV: Its use in young adults with mental retardation. Mental Retardation. 1999; 37(4):319–325. [PubMed: 10463026]

*. Oakland RD, King JD, White LA, Eckman R. A comparison of performance on the WPPSI, WISC, and SB with preschool children: Companion studies. Journal of School Psychology. 1971; 9(2): 144–149.

*. Obrzut A, Nelson RB, Obrzut JE. Construct validity of the Kaufman Assessment Battery for Children with mildly mentally retarded students. American Journal of Mental Deficiency. 1987; 92(1):74–77. [PubMed: 3618659]

*. Obrzut A, Obrzut JE, Shaw D. Construct validity of the Kaufman Assessment Battery for Children with learning disabled and mentally retarded. Psychology in the Schools. 1984; 21(4):417–424.

*. Pasewark RA, Rardin MW, Grice JE Jr. Relationship of the Wechsler Pre-School and Primary Scale of Intelligence and the Stanford-Binet (L-M) in lower class children. Journal of School Psychology. 1971; 9(1):43–50.

*. Phelps L, Leguori S, Nisewaner K, Parker M. Practical interpretations of the WISC-III with language-disordered children. Journal of Psychoeducational Assessment WISC-III Monograph. 1993:71–76.

*. Phelps L, Rosso M, Falasco SL. Correlations between the Woodcock-Johnson and the WISC-R for a behavior disordered population. Psychology in the Schools. 1984; 21:442–446.

*. Phillips BL, Pasewark RA, Tindall RC. Relationship among McCarthy Scales of Children's Abilities, WPPSI, and Columbia Mental Maturity Scale. Psychology in the Schools. 1978; 15(3): 352–356.

*. Pommer LT. Seriously emotionally disturbed children's performance on the Kaufman Assessment Battery for Children: A concurrent validity study. Journal of Psychoeducational Assessment. 1986; 4:155–162.

*. Prewett PN. The relationship between the Kaufman Brief Intelligence Test (K-BIT) and the WISC-R with incarcerated juvenile delinquents. Educational and Psychological Measurement. 1992; 52(4):977–982.

*. Prewett PN, Matavich MA. A comparison of referred students' performance on the WISC-III and the Stanford-Binet Intelligence Scale: Fourth Edition. Journal of Psychoeducational Assessment. 1994; 12:42–48.

*. Prifitera A, Ryan JJ. WAIS-R/WAIS comparisons in a clinical sample. Clinical Neuropsychology. 1983; 5(3):97–99.

*. Prosser NS, Crawford VD. Relationship of scores on the Wechsler Preschool and Primary Scale of Intelligence and the Stanford-Binet Intelligence Scale Form LM. Journal of School Psychology. 1971; 9(3):278–283.

*. Quereshi MY. The comparability of WAIS and WISC subtest scores and IQ estimates. The Journal of Psychology. 1968; 68:73–82. [PubMed: 5636182]

*. Quereshi MY, Erstad D. A comparison of the WAIS and the WAIS-R for ages 61–91 years. Psychological Assessment: A Journal of Consulting and Clinical Psychology. 1990; 2(3):293–297.

*. Quereshi MY, McIntire DH. The comparability of the WISC, WISC-R, and WPPSI. Journal of Clinical Psychology. 1984; 40(4):1036–1043.

*. Quereshi MY, Miller JM. The comparability of the WAIS, WISC, and WBII. Journal of Educational Measurement. 1970; 7(2):105–111.

*. Quereshi MY, Ostrowski MJ. The comparability of three Wechsler Adult Intelligence Scales in a college sample. Journal of Clinical Psychology. 1985; 41(3):397–407.

*. Quereshi MY, Seitz R. Non-equivalence of WPPSI, WPPSI-R, and WISC-R scores. Current Psychology. 1994; 13(3):210–225.

*. Quereshi MY, Treis KM, Riebe AL. The equivalence of the WAIS-R and the WISC-R at age 16. Journal of Clinical Psychology. 1989; 45(4):633–641. [PubMed: 2768503]

*. Rabourn RE. The Wechsler Adult Intelligence Scale (WAIS) and the WAIS-Revised: A comparison and a caution. Professional Psychology: Research and Practice. 1983; 14(3):357–361.

Raudenbush SW, Xiao-Feng L. Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. Psychological Methods. 2001; 6(4): 387–401. [PubMed: 11778679]

*. Reeve RE, Hall RJ, Zakreski RS. The Woodcock-Johnson Tests of Cognitive Ability: Concurrent validity with the WISC-R. Learning Disability Quarterly. 1979; 2(2):63–69.

*. Reilly TP, Drudge OW, Rosen JC, Loew DE, Fischer M. Concurrent and predictive validity of the WISC-R, McCarthy Scales, Woodcock-Johnson, and academic achievement. Psychology in the Schools. 1985; 22:380–382.

*. Rellas AJ. The use of the Wechsler Preschool and Primary Scale (WPPSI) in the early identification of gifted students. The California Journal of Educational Research. 1969; 20:1171–119.

*. Reynolds CR, Hartlage L. Comparison of WISC and WISC-R regression lines for academic prediction with black and with white referred children. Journal of Consulting and Clinical Psychology. 1979; 47(3):589–591.

*. Robinson NM, Dale PS, Landesman S. Validity of Stanford-Binet IV with linguistically precocious toddlers. Intelligence. 1990; 14(2):173–186.

*. Robinson EL, Nagle RJ. The comparability of the Test of Cognitive Skills with the Wechsler Intelligence Scale for Children-Revised and the Stanford-Binet: Fourth Edition with gifted children. Psychology in the Schools. 1992; 29(2):107–112.

Rodgers JL. A critique of the Flynn effect: Massive IQ gains, methodological artifacts, or both? Intelligence. 1999; 26(4):337–356.

*. Rohrs FW, Haworth MR. The 1960 Stanford-Binet, WISC, and Goodenough tests with mentally retarded children. The American Journal of Mental Deficiency. 1962; 66(6):853–859.

*. Roid, GH. Stanford-Binet Intelligence Scales, Fifth Edition, Technical manual. Itasca, IL: Riverside Publishing; 2003.

Rönnlund M, Nilsson L-G. The magnitude, generality, and determinants of Flynn effects on forms of declarative memory and visuospatial ability: Time-sequential analyses of data from a Swedish cohort study. Intelligence. 2008; 36:192–209.

Rönnlund M, Nilsson L-G. Flynn effects on subfactors of episodic and semantic memory: Parallel gains over time and the same set of determining factors. Neuropsychologia. 2009; 47:2174–2180. [PubMed: 19056409]

*. Ross RT, Morledge J. Comparison of the WISC and WAIS at chronological age sixteen. Journal of Consulting Psychology. 1967; 31(3):331–332. [PubMed: 6046591]

*. Rothlisberg BA. Comparing the Stanford-Binet, Fourth Edition to the WISC-R: A concurrent validation study. Journal of School Psychology. 1987; 25(2):193–196.

*. Rowe HAH. Borderline versus mentally deficient: A study of the performance of educable mentally retarded adolescents on WISC-R and WISC. Australian Journal of Mental Retardation. 1977; 4:11–14.

Rushton JP. Flynn effects not genetic and unrelated to race differences. American Psychologist. 2000; 55(5):542–543. [PubMed: 10842435]

*. Rust JO, Lindstrom A. Concurrent validity of the WISC-III and Stanford-Binet IV. Psychological Reports. 1996; 79(2):618–620.

*. Rust JO, Yates AG. Concurrent validity of the Wechsler Intelligence Scale for Children – Third Edition and the Kaufman Assessment Battery for Children. Psychological Reports. 1997; 80(1): 89–90.

*. Sabatino DA, Spangler RS. The relationship between the Wechsler Intelligence Scale for Children-Revised and the Wechsler Intelligence Scale for Children-III scales and subtests with gifted children. Psychology in the Schools. 1995; 32:18–23.

Sanborn KJ, Truscott SD, Phelps L, McDougal JL. Does the Flynn effect differ by IQ level in samples of students classified as learning disabled? Journal of Psychoeducational Assessment. 2003; 21:145–159.

*. Sandoval J, Sassenrath J, Penaloza M. Similarity of WISC-R and WAIS-R scores at age 16. Psychology in the Schools. 1988; 25(4):374–379.

SAS Institute, Inc.. SAS Release 9.2. Cary, NC: SAS Institute Inc.; 2008.

Schatz J, Hamdan-Allen G. Effects of age and IQ on adaptive behavior domains for children with autism. Journal of Autism and Developmental Disorders. 1995; 25(1):51–60. [PubMed: 7608034]

Schooler, C. Environmental complexity and the Flynn effect. In: Neisser, U., editor. The Rising Curve. Washington, DC: American Psychological Association; 1998.

Schull, WJ.; Neel, JV. The effects of inbreeding on Japanese children. New York: Harper & Row; 1965.

*. Schwarting FG. A comparison of the WISC and WISC-R. Psychology in the Schools. 1976; 13:139–141.

*. Sevier RC, Bain SK. Comparison of WISC-R and WISC-III for gifted students. Roeper Review. 1994; 17(1):39–42.

*. Sewell TE. A comparison of the WPPSI and Stanford-Binet Intelligence Scale (1972) among lower SES black children. Psychology in the Schools. 1977; 14(2):158–161.

*. Sewell T, Manni J. Comparison of scores of normal children on the WISC-R and Stanford-Binet, Form LM, 1972. Perceptual and Motor Skills. 1977; 45:1057–1058.

*. Shahim S. Correlations for Wechsler Intelligence Scale for Children-Revised and the Wechsler Preschool and Primary Scale of Intelligence for Iranian children. Psychological Reports. 1992; 70:27–30. [PubMed: 1565731]

*. Sherrets S, Quattrocchi M. WISC – WISC-R differences – Fact or artifact? Journal of Pediatric Psychology. 1979; 4(2):119–127.

*. Simon CS, Clopton JR. Comparison of WAIS and WAIS-R scores of mildly and moderately mentally retarded adults. American Journal of Mental Deficiency. 1984; 89(3):301–303. [PubMed: 6517112]

*. Simpson RL. Study of the comparability of the WISC and WAIS. Journal of Consulting and Clinical Psychology. 1970; 34(2):156–158.

*. Simpson M, Carone DA Jr, Burns WJ, Seidman T, Montgomery D, Sellers A. Assessing giftedness with the WISC-III and the SB-IV. Psychology in the Schools. 2002; 39(5):515–524.

*. Skuy M, Taylor M, O'Carroll S, Fridjhon P, Rosenthal L. Performance of black and white South African children on the Wechsler Intelligence Scale for Children-Revised and the Kaufman Assessment Battery. Psychological Reports. 2000; 86(3):727–737. [PubMed: 10876320]

*. Smith RP. A comparison study of the Wechsler Adult Intelligence Scale and the Wechsler Adult Intelligence Scale-Revised in a college population. Journal of Consulting and Clinical Psychology. 1983; 51(3):414–419.

*. Smith DK, St. Martin ME, Lyon MA. A validity study of the Stanford-Binet: Fourth Edition with students with learning disabilities. Journal of Learning Disabilities. 1989; 22(4):260–261. [PubMed: 2738463]

Social Security Administration. Disability Evaluation under Social Security. 2008. SSA Publication No. 54-039, section 12.05. www.ssa.gov/disability/professionals/bluebook.

*. Solly DC. Comparison of WISC and WISC-R scores of mentally retarded and gifted children. Journal of School Psychology. 1977; 15(3):255–258.

Spitz HH. Variations in Wechsler interscale IQ disparities at different levels of IQ. Intelligence. 1989; 13:157–167.

*. Spruill J. A comparison of the Wechsler Adult Intelligence Scale-Revised with the Stanford-Binet Intelligence Scale (4th Edition) for mentally retarded adults. Psychological Assessment: A Journal of Consulting and Clinical Psychology. 1991; 3(1):133–135.

*. Spruill J, Beck BL. Comparison of the WAIS and WAIS-R: Different results for different IQ groups. Professional Psychology: Research and Practice. 1988; 19(1):31–34.

*. Stokes EH, Brent D, Huddleston NJ, Rozier JS. A comparison of WISC and WISC-R scores of sixth grade students: Implications for validity. Educational and Psychological Measurement. 1978; 38(2):469–473.

Sundet JM, Borren I, Tambs K. The Flynn effect is partly caused by changing fertility patterns. Intelligence. 2008; 36:183–191.

Sundet JM, Barlaug DG, Torjussen TM. The end of the Flynn effect? A study of secular trends in mean intelligence test scores of Norwegian conscripts during half a century. Intelligence. 2004; 32:349–362.

*. Swerdlik ME. Comparison of WISC and WISC-R scores of referred black, white and Latino children. Journal of School Psychology. 1978; 16(2):110–125.

Teasdale TW, Owen DR. Continuing secular increases in intelligence and stable prevalence of high intelligence levels. Intelligence. 1989; 13:255–262.

Teasdale TW, Owen DR. A long-term rise and recent decline in intelligence test performance: The Flynn effect in reverse. Personality and Individual Differences. 2005; 39:837–843.

Teasdale TW, Owen DR. Secular declines in cognitive test scores: A reversal of the Flynn effect. Intelligence. 2008; 36:1212–126.

Te Nijenhuis J, van der Flier H. The secular rise in IQs in the Netherlands: Is the Flynn effect on *g*? Personality and Individual Differences. 2007; 43:1259–1265.

*. Templer DI, Schmitz SP, Corgiat MD. Comparison of the Stanford-Binet with the Wechsler Adult Intelligence Scale-Revised: Preliminary report. Psychological Reports. 1985; 57(1):335–336.

*. Terman, LM.; Merrill, MA. Measuring intelligence. Boston, MA: Houghton Mifflin; 1937.

*. Terman, LM.; Merrill, MA. Stanford-Binet Intelligence Scale: Manual for the Third Revised Form L-M. Boston, MA: Houghton Mifflin; 1960.

*. Terman, LM.; Merrill, MA. Stanford-Binet Intelligence Scale: Manual for the Third Revision Form L-M (1972 Normal Tables by R.L. Thorndike). Boston, MA: Houghton Mifflin; 1973.

*. Thompson PL, Brassard MR. Validity of the Woodcock-Johnson Tests of Cognitive Ability: A comparison with the WISC-R in LD and normal elementary students. Journal of School Psychology. 1984; 22:201–208.

*. Thompson AP, Sota DD. Comparison of WAIS-R and WISC-III scores with a sample of 16-year-old youth. Psychological Reports. 1998; 82(3):1339–1347. [PubMed: 9709537]

*. Thorndike, RL.; Hagen, EP.; Sattler, JM. Stanford-Binet Intelligence Scale, Fourth Edition, Technical manual. Chicago, IL: The Riverside Publishing Company; 1986.

*. Triggs FO, Cartee JK. Pre-school pupil performance on the Stanford-Binet and the Wechsler Intelligence Scale for Children. Journal of Clinical Psychology. 1953; 9(1):27–29. [PubMed: 13022792]

Tong VT, Jones JR, Dietz PM, D'Angelo D, Bombard JM. Trends in smoking before, during, and after pregnancy – The Pregnancy Risk Assessment Monitoring System (PRAMS), United States, 31 sites, 2000–2005. Morbidity and Mortality Weekly Report. 2009; 58(SS-4):1–30. [PubMed: 19145219]

Tuddenham RD. Soldier intelligence in world wars I and II. American Psychologist. 1948; 3:54–56. [PubMed: 18911933]

*. Tuma JM, Appelbaum AS, Bee DE. Comparability of the WISC and the WISC-R in normal children of divergent socioeconomic backgrounds. Psychology in the Schools. 1978; 15(3):339–346.

*. Urbina SP, Clayton JP. WPPSI-R/WISC-R: A comparative study. Journal of Psychoeducational Assessment. 1991; 9:247–254.

*. Urbina SP, Golden CJ, Ariel RN. WAIS/WAIS-R: Initial comparisons. Clinical Neuropsychology. 1982; 4(4):145–146. doi:

*. Valencia RR. Concurrent validity of the Kaufman Assessment Battery for Children in a sample of Mexican-American children. Educational and Psychological Measurement. 1984; 44(2):365–372.

*. Valencia RR, Rothwell JG. Concurrent validity of the WPPSI with Mexican-American preschool children. Educational and Psychological Measurement. 1984; 44(4):955–961.

*. Vo DH, Weisenberger JL, Becker R, Jacob-Timm S. Concurrent validity of the KAIT for students in grade six and eight. Journal of Psychoeducational Assessment. 1999; 17:152–162.

Walker v. True, 399 F.3d 315, 322–23 (4th Cir. 2005).

*. Walters SO, Weaver KA. Relationships between the Kaufman Brief Intelligence Test and the Wechsler Adult Intelligence Scale – Third Edition. Psychological Reports. 2003; 92(3):1111–1115. [PubMed: 12931928]

*. Wechsler, D. Wechsler Intelligence Scale for Children. New York, NY: The Psychological Corporation; 1949.

*. Wechsler, D. Wechsler Adult Intelligence Scale, Manual. New York, NY: The Psychological Corporation; 1955.

*. Wechsler, D. Wechsler Preschool and Primary Scale of Intelligence, Manual. New York, NY: The Psychological Corporation; 1967.

*. Wechsler, D. Wechsler Intelligence Scale for Children – Revised, Manual. San Antonio, TX: The Psychological Corporation; 1974.

*. Wechsler, D. Wechsler Adult Intelligence Scale – Revised, Manual. New York, NY: The Psychological Corporation; 1981.

*. Wechsler, D. Wechsler Preschool and Primary Scale of Intelligence-Revised, Manual. San Antonio, TX: The Psychological Corporation; 1989.

*. Wechsler, D. Wechsler Intelligence Scale for Children, Third Edition, Manual. San Antonio, TX: The Psychological Corporation; 1991.

*. Wechsler, D. Wechsler Adult Intelligence Scale, Third Edition, Technical manual. San Antonio, TX: The Psychological Corporation; 1997.

*. Wechsler, D. Wechsler Abbreviated Scale of Intelligence, Manual. San Antonio, TX: The Psychological Corporation; 1999.

*. Wechsler, D. Wechsler Preschool and Primary Scale of Intelligence, Third Edition, Technical manual. San Antonio, TX: The Psychological Corporation; 2002.

*. Wechsler, D. Wechsler Intelligence Scale for Children, Fourth Edition, Technical and interpretive manual. San Antonio, TX: The Psychological Corporation; 2003.

*. Wechsler, D. Wechsler Adult Intelligence Scale, Fourth Edition, Technical manual. San Antonio, TX: The Psychological Corporation; 2008.

*. Weider A, Noller PA, Schramm TA. The Wechsler Intelligence Sale for Children and the Revised Stanford-Binet. Journal of Consulting Psychology. 1951; 15(4):330–333. [PubMed: 14861332]

*. Weiner SG, Kaufman AS. WISC-R versus WISC for black children suspected of learning or behavioral disorders. Journal of Learning Disabilities. 1979; 12(2):100–105. [PubMed: 438637]

*. Wheaton PJ, Vandergriff AF, Nelson WH. Comparability of the WISC and WISC-R with bright elementary school students. Journal of School Psychology. 1980; 18(3):271–275.

*. Whitworth RH, Chrisman SM. Validation of the Kaufman Assessment Battery for Children comparing Anglo and Mexican-American preschoolers. Educational and Psychological Measurement. 1987; 47(3):695–702.

*. Whitworth RH, Gibbons RT. Cross-racial comparison of the WAIS and WAIS-R. Educational and Psychological Measurement. 1986; 46(4):1041.

Wicherts JM, Dolan CV, Hessen DJ, Oosterveld P, van Baal GCM, Boomsma DI, Span MM. Are intelligence tests measurement invariant over time? Investigating the nature of the Flynn effect. Intelligence. 2004; 32:509–537.

*. Woodcock, RW.; Johnson, MB. Woodcock-Johnson Psycho-Educational Battery. Allen, TX: DLM Teaching Resources; 1977.

*. Woodcock, RW.; Johnson, MB. Woodcock-Johnson Psycho-Educational Battery-Revised. Allen, TX: DLM Teaching Resources; 1989.

*. Woodcock, RW.; McGrew, KS.; Mather, N. WJ III NU Tests of Cognitive Ability, Examiner's Manual. Rolling Meadows, IL: Riverside Publishing; 2001.

Woodley MA. Heterosis doesn't cause the Flynn effect: A critical examination of Mingroni (2007). Psychological Review. 2011; 118:689–693. [PubMed: 22003846]

*. Yater AC, Boyd M, Barclay A. A comparative study of WPPSI and WISC performances of disadvantaged children. Journal of Clinical Psychology. 1975; 31(1):78–80.

Young B, Boccaccini MT, Conroy MA, Lawson K. Four practical and conceptual assessment issues that evaluators should address in capital case mental retardation evaluations. Professional Psychology: Research and Practice. 2007; 38:169–178.

*. Ysseldyke J, Shinn M, Epps S. A comparison of the WISC-R and the Woodcock-Johnson Tests of Cognitive Ability. Psychology in the Schools. 1981; 18:15–19.

Zhou X, Zhu J, Weiss LG, Pearson. Peeking inside the "blackbox" of the Flynn effect: Evidence from three Wechsler instruments. Journal of Psychoeducational Assessment. 2010; 28(5):399–411.

*. Zimmerman IL, Woo-Sam J. The utility of the Wechsler Preschool and Primary Scale of Intelligence in the public school. Journal of Clinical Psychology. 1970; 26(4):472.

*. Zimmerman IL, Woo-Sam J. A note on the current validity of the renormed (1974) Stanford Binet LM. 1974

*. Zins JE, Barnett DW. A validity study of the K-ABC, the WISC-R, and the Stanford-Binet with nonreferred children. Journal of School Psychology. 1984; 22:369–371.

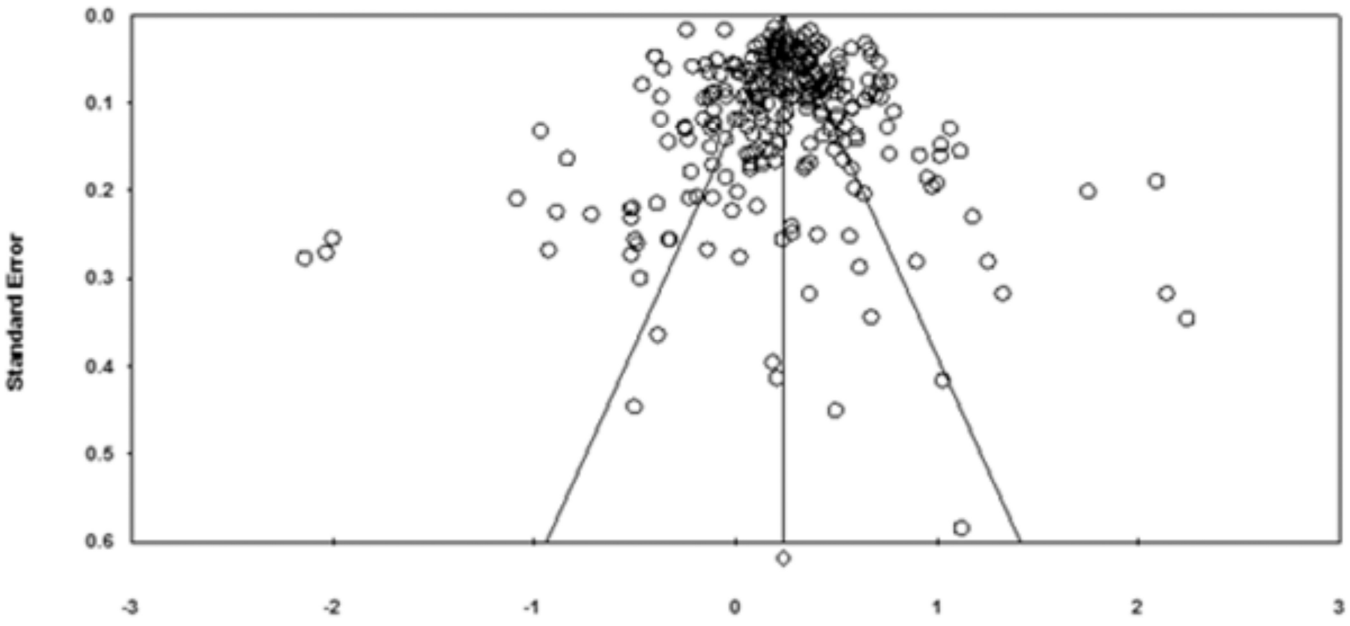

Figure 1.
Study effect sizes and standard errors included in the overall model.

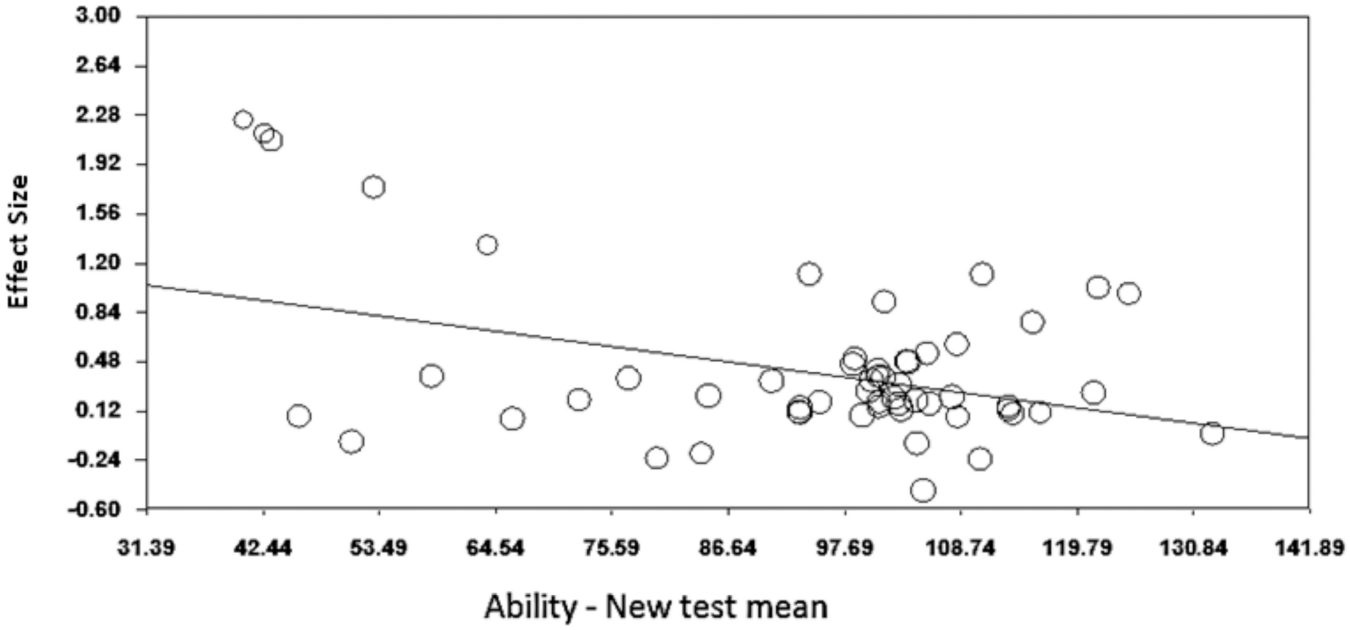

Figure 2.
Study effect size regressed on sample ability in the modern set.

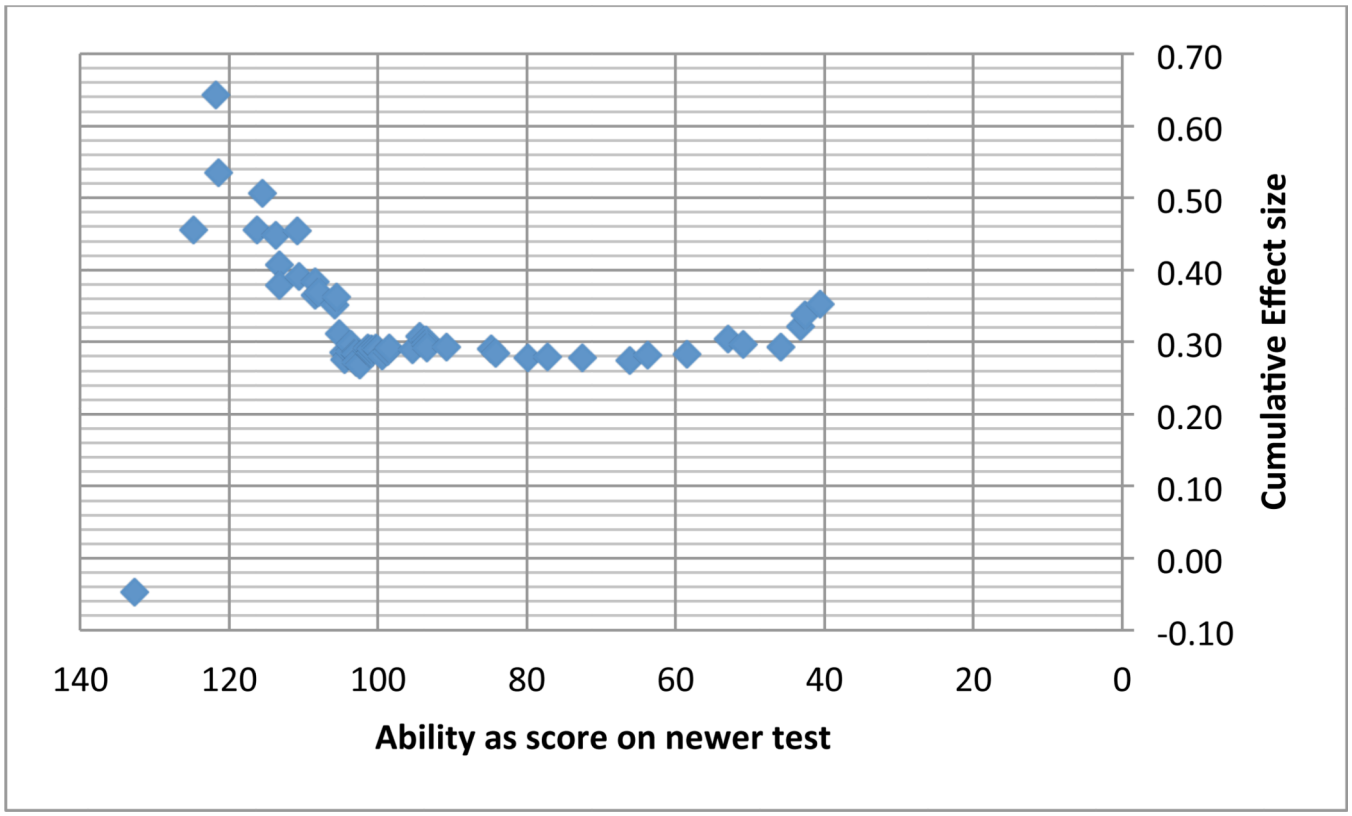

Figure 3.
Cumulative Flynn effect by decreasing sample ability.

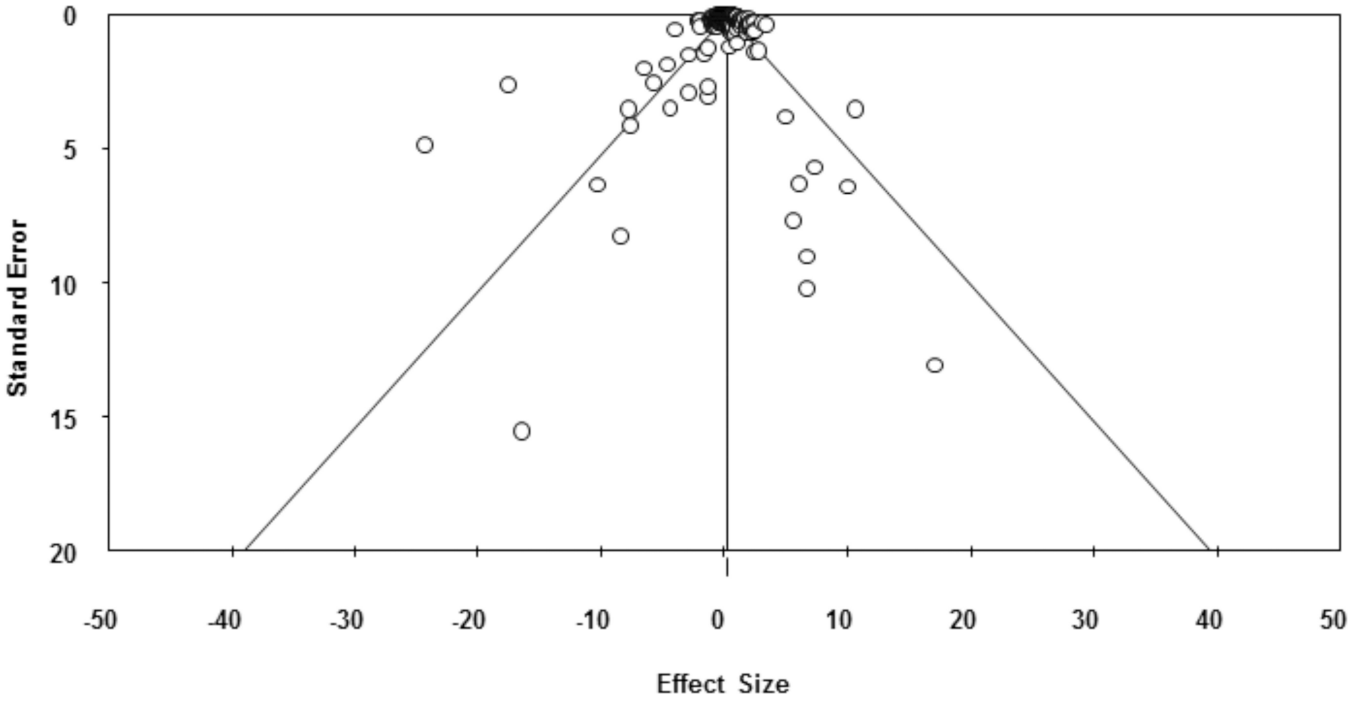

Figure 4.
Complete set of study effect sizes and their standard errors.

Table 1

Sample Size, Sample Age, Tests Administered, and Effect Sizes by Study

	Source	N	Age[a]	Newer Test	Older Test	Effect size
			Modem ≥ 5[b]			
1	Bower & Hayes, 1995	26	132.88	SB4[c]	SB 72[d]	0.08
2	Carvajal & Weyand, 1986	23	109.5	SB4	WISC-R[e]	0.13
3	Carvajal et al., 1987	32	227	SB4	WAIS-R[f]	0.37
4	Clark et al., 1987	47	63	SB4	SB 72	0.07
5	Doll & Boren, 1993	24	114	WISC-III[g]	WISC-R	0.35
6	Gordon et al., 2010	17	194	WISC-IV[h]	WAIS-III[i]	1.75
7	Gunter et al., 1995	16	132	WISC-III	WISC-R	−0.19
8	Krohn & Lamp, 1989	89	59	SB4	SB 72	0.11
9	Lamp & Krohn, 2001	89	59	SB4	SB 72	0.10
10	Nelson & Dacey, 1999	42	248.04	SB4	WAIS-R	2.09
11	Quereshi & Seitz, 1994	72	75.1	WPPSI-R[j]	WISC-R	0.34
12	Quereshi et al., 1989	36	197.9	WAIS-R	WISC-R	0.1
13	Quereshi et al., 1989	36	197.9	WAIS-R	WISC-R	1.11
14	Quereshi et al., 1989	36	197.05	WAIS-R	WISC-R	0.18
15	Quereshi et al., 1989	36	197.05	WAIS-R	WISC-R	1.11
16	Robinson & Nagle, 1992	75	111	SB4	WISC-R	0.25
17	Robinson et al., 1990	28	30	SB4	SB 72	0.97
18	Roid, 2003	87	744	SB5[k]	WAIS-III	0.91
19	Roid, 2003	66	132	SB5	WISC-III	0.41
20	Roid, 2003	71	48	SB5	WPPSI-R	−0.46
21	Roid, 2003	104	108	SB5	SB4	0.22
22	Roid, 2003	80	84	SB5	SB L-M[l]	0.12
23	Rothlisberg, 1987	32	93.19	SB4	WISC-R	0.53
24	Sabatino & Spangler, 1995	51	163.2	WISC-III	WISC-R	−0.04
25	Sandoval et al., 1988	30	197.5	WAIS-R	WISC-R	0.18
26	Sevier & Bain, 1994	35	110	WISC-III	WISC-R	0.76

	Source	N	Age[a]	Newer Test	Older Test	Effect size
27	Spruill, 1991	32	.	SB4	WAIS-R	2.24
28	Spruill, 1991	38	.	SB4	WAIS-R	2.14
29	Thompson & Sota, 1998	23	196	WISC-III	WAIS-R	0.60
30	Thompson & Sota, 1998	23	196	WISC-III	WAIS-R	−0.23
31	Thorndike et al., 1986	21	234	SB4	WAIS-R	1.32
32	Thorndike et al., 1986	47	233	SB4	WAIS-R	0.5
33	Thorndike et al., 1986	205	113	SB4	WISC-R	0.21
34	Thorndike et al., 1986	19	155	SB4	WISC-R	0.10
35	Thorndike et al., 1986	90	132	SB4	WISC-R	0.23
36	Thorndike et al., 1986	61	167	SB4	WISC-R	0.06
37	Thorndike et al., 1986	139	83	SB4	SB 72	0.17
38	Thorndike et al., 1986	82	88	SB4	SB 72	1.01
39	Thorndike et al., 1986	14	100	SB4	SB 72	−0.22
40	Thorndike et al., 1986	22	143	SB4	SB 72	−0.10
41	Urbina & Clayton, 1991	50	79	WPPSI-R	WISC-R	0.48
42	Wechsler, 1981	80	192	WAIS-R	WISC-R	0.15
43	Wechsler, 1989	50	79	WPPSI-R	WISC-R	0.47
44	Wechsler, 1991	189	192	WISC-III	WAIS-R	0.36
45	Wechsler, 1991	206	132	WISC-III	WISC-R	0.31
46	Wechsler, 1997	184	192	WAIS-III	WISC-III	−0.11
47	Wechsler, 1997	24	219.6	WAIS-III	WISC-R	0.33
48	Wechsler, 1997	26	343.2	WAIS-III	SB4	0.15
49	Wechsler, 1997	192	522	WAIS-III	WAIS-R	0.17
50	Wechsler, 1997	88	583.2	WAIS-III	WAIS-R	0.14
51	Wechsler, 2002	176	60	WPPSI-III[m]	WPPSI-R	0.08
52	Wechsler, 2003	183	192	WISC-IV	WAIS-III	0.45
53	Wechsler, 2003	233	132	WISC-IV	WISC-III	0.19
54	Wechsler, 2008	238	632.4	WAIS-IV[n]	WAIS-III	0.26
55	Wechsler, 2008	24	386.4	WAIS-IV	WAIS-III	0.37
56	Wechsler, 2008	24	348	WAIS-IV	WAIS-III	0.2

	Source	N	Age[a]	Newer Test	Older Test	Effect size
			Other ≥ 5[o]			
57	Appelbaum & Tuma, 1977	20	121	WISC-R	WISC[p]	0.07
58	Appelbaum & Tuma, 1977	20	120	WISC-R	WISC	0.15
59	Arinoldo, 1982	20	57	MSCA[q]	WPPSI[r]	0.20
60	Arnold & Wagner, 1955	50	102	WISC	SB 32[s]	0.07
61	Axelrod & Naugle, 1998	200	519.6	KBIT[t]	WAIS-R	−0.4
62	Barratt & Baumgarten, 1957	30	126	WISC	SB 32	0.58
63	Barratt & Baumgarten, 1957	30	126	WISC	SB 32	0.08
64	Bradway & Thompson, 1962	111	354	WAIS[u]	SB 32	0.66
65	Brengelmann & Renny, 1961	75	442.92	WAIS	SB 32	−0.35
66	Brooks, 1977	30	96	WISC-R	WISC	0.29
67	Brooks, 1977	30	96	SB 72	WISC	0.37
68	Byrd & Buckhalt, 1991	46	149	DAS[v]	WISC-R	0.12
69	Carvajal et al., 1988	21	69	SB4	MSCA	−0.1
70	Carvajal et al., 1988	20	66	SB4	WPPSI	0.05
71	Chelune et al., 1987	43	576	WAIS-R	WAIS	0.40
72	Cohen & Collier, 1952	51	89	WISC	SB 32	0.32
73	Covin, 1977	30	102	WISC-R	WISC	−0.00
74	Craft & Kronenberger, 1979	15	196.44	WISC-R	WAIS	0.72
75	Craft & Kronenberger, 1979	15	196.8	WISC-R	WAIS	0.54
76	Davis, 1975	53	69	MSCA	SB 60[w]	0.57
77	Edwards & Klein, 1984	19	451.2	WAIS-R	WAIS	0.13
78	Edwards & Klein, 1984	19	451.2	WAIS-R	WAIS	0.37
79	Eisenstein & Engelhart, 1997	64	500.4	KBIT	WAIS-R	−0.25
80	Elliot, 1990	23	54	WPPSI-R	K-ABC[x]	0.5
81	Elliot, 1990	49	41.5	DAS	MSCA	0.42
82	Elliot, 1990	40	42.5	DAS	MSCA	0.45
83	Elliot, 1990	66	110	DAS	WISC-R	0.50
84	Elliot, 1990	60	180	DAS	WISC-R	0.35
85	Elliot, 1990	23	54	DAS	K-ABC	0.67

	Source	N	Age[a]	Newer Test	Older Test	Effect size
86	Elliot, 1990	27	72	DAS	K-ABC	1.25
87	Faust & Hollingsworth, 1991	33	53.9	WPPSI-R	MSCA	0.07
88	Field & Sisley, 1986	17	360	WAIS-R	WAIS	0.22
89	Field & Sisley, 1986	25	360	WAIS-R	WAIS	0.25
90	Fourqurean, 1987	42	116	K-ABC	WISC-R	−0.71
91	Frandsen & Higginson, 1951	54	116	WISC	SB 32	0.21
92	Gehman & Matyas, 1956	60	182	WISC	SB 32	−0.10
93	Gehman & Matyas, 1956	60	133	WISC	SB 32	−0.12
94	Gerken & Hodapp, 1992	16	54	WPPSI-R	SB 60	0.08
95	Giannell & Freeburne, 1963	38	218.88	WAIS	SB 32	0.55
96	Giannell & Freeburne, 1963	36	219.96	WAIS	SB 32	0.50
97	Giannell & Freeburne, 1963	35	224.28	WAIS	SB 32	0.35
98	Hamm et al., 1976	22	121.68	WISC-R	WISC	0.32
99	Hamm et al., 1976	26	153.73	WISC-R	WISC	0.29
100	Hannon & Kicklighter, 1970	13	192	WAIS	WISC	−0.04
101	Hannon & Kicklighter, 1970	13	192	WAIS	WISC	−2.03
102	Hannon & Kicklighter, 1970	32	192	WAIS	WISC	0.95
103	Hannon & Kicklighter, 1970	33	192	WAIS	WISC	−0.50
104	Hannon & Kicklighter, 1970	11	192	WAIS	WISC	1.12
105	Hannon & Kicklighter, 1970	18	192	WAIS	WISC	1.03
106	Harrington et al., 1992	10	48	WPPSI-R	WJTCA[y]	−0.66
107	Harrington et al., 1992	10	60	WPPSI-R	WJTCA	−0.16
108	Hartlage & Boone, 1977	42	126	WISC-R	WISC	0.20
109	Hartwig et al., 1987	30	135.6	SB4	SB 60	−0.05
110	Hays et al., 2002	85	408	WASI	KBIT	0.22
111	Holland, 1953	23	.	WISC	SB 32	0.10
112	Holland, 1953	29	.	WISC	SB 32	0.10
113	Jones, 1962	80	96	WISC	SB 32	0.54
114	Jones, 1962	80	108	WISC	SB 32	0.46
115	Jones, 1962	80	120	WISC	SB 32	0.38
116	Kangas & Bradway, 1971	48	498	SB 60	WAIS	−2

	Source	N	Age[a]	Newer Test	Older Test	Effect size
117	Kaplan et al., 1991	30	57	WPPSI-R	WPPSI	0.36
118	Karr et al., 1992	21	69	SB4	MSCA	−0.17
119	Karr et al., 1993	32	63.6	WPPSI-R	MSCA	0.07
120	Kaufman & Kaufman, 1990	64	257	KBIT	WAIS-R	−0.11
121	Kaufman & Kaufman, 1990	41	66	KBIT	K-ABC	−0.13
122	Kaufman & Kaufman, 1990	35	128	KBIT	WISC-R	0.35
123	Kaufman & Kaufman, 1990	70	100	KBIT	K-ABC	0.07
124	Kaufman & Kaufman, 1990	39	136	KBIT	K-ABC	−0.48
125	Kaufman & Kaufman, 1993	118	156	KAIT	WISC-R	0.23
126	Kaufman & Kaufman, 1993	71	208.8	KAIT	WAIS-R	0.14
127	Kaufman & Kaufman, 1993	108	312	KAIT	WAIS-R	0.21
128	Kaufman & Kaufman, 1993	90	494.4	KAIT	WAIS-R	0.47
129	Kaufman & Kaufman, 1993	74	747.6	KAIT	WAIS-R	0.25
130	Kaufman & Kaufman, 1993	124	135.6	KAIT	K-ABC	0.47
131	Kaufman & Kaufman, 2004b	54	68	KBIT-II[z]	K-BIT	0.10
132	Kaufman & Kaufman, 2004a	48	120	K-ABC-II[aa]	K-ABC	0.30
133	Kaufman & Kaufman, 2004a	119	126	K-ABC-II	WISC-III	0.09
134	Kaufman & Kaufman, 2004a	29	174	K-ABC-II	KAIT[bb]	0.13
135	Kaufman & Kaufman, 2004b	53	135	KBIT-II	K-BIT	0.24
136	Kaufman & Kaufman, 2004b	74	383	KBIT-II	K-BIT	0.16
137	Kaufman & Kaufman, 2004b	43	122	KBIT-II	WISC-III	0.24
138	Kaufman & Kaufman, 2004b	67	384	KBIT-II	WAIS-III	0.78
139	King & Smith, 1972	24	72	WPPSI	WISC	−0.15
140	King & Smith, 1972	24	72	SB 60	WISC	−0.51
141	Klanderman et al., 1985	41	102	K-ABC	SB 72	0.56
142	Klanderman et al., 1985	41	102	K-ABC	WISC-R	0.40
143	Klinge et al., 1976	16	169.32	WISC-R	WISC	−0.12
144	Klinge et al., 1976	16	169.32	WISC-R	WISC	0.40
145	Krohn et al., 1988	38	51	K-ABC	SB 72	−0.32
146	Krohn & Lamp, 1989	89	59	K-ABC	SB 72	−0.12
147	Krugman et al., 1951	38	60	WISC	SB 32	0.72

	Source	N	Age[a]	Newer Test	Older Test	Effect size
148	Krugman et al., 1951	20	174	WISC	SB 32	0.24
149	Krugman et al., 1951	38	72	WISC	SB 32	0.64
150	Krugman et al., 1951	43	84	WISC	SB 32	0.25
151	Krugman et al., 1951	44	96	WISC	SB 32	0.39
152	Krugman et al., 1951	31	108	WISC	SB 32	0.66
153	Krugman et al., 1951	29	120	WISC	SB 32	0.36
154	Krugman et al., 1951	37	132	WISC	SB 32	0.42
155	Krugman et al., 1951	22	144	WISC	SB 32	0.42
156	Krugman et al., 1951	30	156	WISC	SB 32	0.42
157	Kureth et al., 1952	50	60	WISC	SB 32	0.72
158	Kureth et al., 1952	50	72	WISC	SB 32	0.36
159	Lamp & Krohn, 2001	89	59	K-ABC	SB 72	−0.11
160	Larrabee & Holroyd, 1976	24	129	WISC-R	WISC	0.25
161	Larrabee & Holroyd, 1976	14	129	WISC-R	WISC	0.50
162	Levinson, 1959	57	65.54	WISC	SB 32	0.76
163	Levinson, 1959	60	66.65	WISC	SB 32	0.66
164	Levinson, 1960	117	66.1	WISC	SB 32	0.71
165	Lippold & Claiborn, 1983	30	619.56	WAIS-R	WAIS	0.34
166	McCarthy, 1972	35	75	MSCA	SB 60	1.02
167	McCarthy, 1972	35	75	MSCA	WPPSI	0.36
168	McGinley, 1981	12	141	WISC-R	WISC	0.17
169	McGinley, 1981	9	141	WISC-R	WISC	0.37
170	McKerracher & Scott, 1966	31	384	SB 60	WAIS	0.64
171	Milrod & Rescorla, 1991	50	59	WPPSI-R	WPPSI	0.38
172	Milrod & Rescorla, 1991	30	59	WPPSI-R	WPPSI	0.05
173	Mishra & Brown, 1983	88	359.76	WAIS-R	WAIS	0.19
174	Mitchell et al., 1986	35	.	WAIS-R	WAIS	0.15
175	Munford, 1978	10	141	WISC-R	WISC	0.04
176	Munford, 1978	10	141	WISC-R	WISC	−0.36
177	Munford & Munoz, 1980	11	150.5	WISC-R	WISC	−0.07
178	Munford & Munoz, 1980	9	150.5	WISC-R	WISC	0.34

	Source	N	Age[a]	Newer Test	Older Test	Effect size
179	Nagle & Lazarus, 1979	30	197.5	WISC-R	WAIS	0.69
180	Naglieri, 1984	35	105	K-ABC	WISC-R	−0.92
181	Naglieri, 1984	33	105	K-ABC	WISC-R	0.59
182	Naglieri, 1985	37	117	K-ABC	WISC-R	−0.83
183	Naglieri, 1985	51	91	K-ABC	MSCA	−0.11
184	Naglieri & Jensen, 1987	86	128.4	K-ABC	WISC-R	0.43
185	Naglieri & Jensen, 1987	86	129.6	K-ABC	WISC-R	0.08
186	Naugle et al., 1993	200	519.6	KBIT	WAIS-R	−0.39
187	Oakland et al., 1971	24	72	SB 60	WISC	−0.52
188	Oakland et al., 1971	24	74	WPPSI	WISC	0.21
189	Oakland et al., 1971	24	72	WPPSI	WISC	−0.15
190	Oakland et al., 1971	24	74	SB 60	WISC	0.02
191	Obrzut et al., 1984	19	110.06	K-ABC	WISC-R	0.28
192	Obrzut et al., 1984	13	111.06	K-ABC	WISC-R	−0.47
193	Obrzut et al., 1987	29	114.96	K-ABC	SB 72	−0.38
194	Obrzut et al., 1987	29	114.96	K-ABC	WISC-R	−0.88
195	Phelps et al., 1993	40	108	WISC-III	K-ABC	1.00
196	Phillips et al., 1978	60	73.92	MSCA	WPPSI	1.17
197	Pommer, 1986	56	87.86	K-ABC	WISC-R	−1.08
198	Prewett, 1992	40	189	KBIT	WISC-R	0.02
199	Prifitera & Ryan, 1983	32	529.08	WAIS-R	WAIS	0.31
200	Quereshi, 1968	124	180.1	WAIS	WISC	0.6
201	Quereshi & Miller, 1970	72	208.65	WAIS	WISC	0.49
202	Quereshi & McIntire, 1984	24	74.5	WPPSI	WISC	0
203	Quereshi & McIntire, 1984	24	74.5	WPPSI	WISC	0.50
204	Quereshi & McIntire, 1984	24	74.5	WPSSI	WISC	0.16
205	Quereshi & McIntire, 1984	24	74.5	WISC-R	WISC	−0.14
206	Quereshi & McIntire, 1984	24	74.5	WISC-R	WISC	0.25
207	Quereshi & McIntire, 1984	24	74.5	WISC-R	WISC	0.18
208	Quereshi & McIntire, 1984	24	74.5	WISC-R	WPPSI	−0.49
209	Quereshi & McIntire, 1984	24	74.5	WISC-R	WPPSI	−0.33

	Source	N	Age[a]	Newer Test	Older Test	Effect size
210	Quereshi & McIntire, 1984	24	74.5	WISC-R	WPPSI	0.23
211	Quereshi & Ostrowski, 1985	72	230.9	WAIS-R	WAIS	0.15
212	Quereshi & Erstad, 1990	36	891.6	WAIS-R	WAIS	0.64
213	Quereshi & Erstad, 1990	36	891.6	WAIS-R	WAIS	0.43
214	Quereshi & Erstad, 1990	18	1032	WAIS-R	WAIS	0.67
215	Quereshi & Erstad, 1990	27	906	WAIS-R	WAIS	0.57
216	Quereshi & Erstad, 1990	27	786	WAIS-R	WAIS	0.41
217	Quereshi & Seitz, 1994	72	75.1	WPPSI-R	WPPSI	0.40
218	Quereshi & Seitz, 1994	72	75.1	WISC-R	WPPSI	0.53
219	Raboum, 1983	52	308.4	WAIS-R	WAIS	0.27
220	Reilly et al., 1985	26	84	WJTCA	MSCA	−0.05
221	Reynolds & Hartlage, 1979	66	152.4	WISC-R	WISC	0.18
222	Rohrs & Haworth, 1962	46	149.88	SB 60	WISC	−0.33
223	Ross & Morledge, 1967	30	192	WAIS	WISC	−0.36
224	Rowe, 1977	20	170.5	WISC-R	WISC	0.016
225	Rowe, 1977	24	170.5	WISC-R	WISC	0.34
226	Rust & Yates, 1997	67	102	WISC-III	K-ABC	0.01
227	Schwarting, 1976	58	126	WISC-R	WISC	0.30
228	Sewell, 1977	35	62.29	SB 72	WPPSI	0.61
229	Shahim, 1992	40	74.4	WISC-R	WPPSI	−0.22
230	Sherrets & Quattrocchi, 1979	13	141.6	WISC-R	WISC	0.05
231	Sherrets & Quattrocchi, 1979	15	141.6	WISC-R	WISC	0.20
232	Simon & Clopton, 1984	29	354	WAIS-R	WAIS	−0.08
233	Simpson, 1970	120	192	WAIS	WISC	−0.96
234	Skuy et al., 2000	21	114	K-ABC	WISC-R	−2.13
235	Skuy et al., 2000	35	100.8	K-ABC	WISC-R	−0.38
236	Smith, 1983	35	247.2	WAIS-R	WAIS	−0.21
237	Smith, 1983	35	247.2	WAIS-R	WAIS	0.51
238	Solly, 1977	12	124	WISC-R	WISC	0.50
239	Solly, 1977	12	124	WISC-R	WISC	0.43
240	Spruill & Beck, 1988	23	306	WAIS-R	WAIS	0.37

	Source	N	Age[a]	Newer Test	Older Test	Effect size
241	Spruill & Beck, 1988	35	306	WAIS-R	WAIS	0.19
242	Spruill & Beck, 1988	25	306	WAIS-R	WAIS	−0.05
243	Spruill & Beck, 1988	25	306	WAIS-R	WAIS	−0.24
244	Stokes et al., 1978	59	147	WISC-R	WISC	0.10
245	Swerdlik, 1978	100	108	WISC-R	WISC	0.23
246	Swerdlik, 1978	64	163.2	WISC-R	WISC	0.20
247	Templer et al., 1985	15	347.16	WAIS-R	SB 60	0.75
248	Thorndike et al., 1986	75	66	SB4	WPPSI	0.24
249	Triggs & Cartee, 1953	46	60	WISC	SB 32	1.06
250	Tuma et al., 1978	9	119	WISC-R	WISC	0.12
251	Tuma et al., 1978	9	119	WISC-R	WISC	0.29
252	Tuma et al., 1978	9	123	WISC-R	WISC	−0.04
253	Tuma et al., 1978	9	123	WISC-R	WISC	0.27
254	Urbina et al., 1982	68	505.92	WAIS-R	WAIS	0.21
255	Valencia & Rothwell, 1984	39	54.9	MSCA	WPPSI	0.18
256	Valencia, 1984	42	59.5	K-ABC	WPPSI	−0.10
257	Walters & Weaver, 2003	20	278.4	WAIS-III	KBIT	−0.51
258	Wechsler, 1955	52	252	WAIS	SB 32	0.23
259	Wechsler, 1974	40	203	WISC-R	WAIS	0.33
260	Wechsler, 1974	50	72	WISC-R	WPPSI	0.34
261	Wechsler, 1981	72	474	WAIS-R	WAIS	0.30
262	Wechsler, 1989	61	63.5	WPPSI-R	WPPSI	0.50
263	Wechsler, 1989	83	63.5	WPPSI-R	WPPSI	0.20
264	Wechsler, 1989	93	62.5	WPPSI-R	MSCA	0.14
265	Wechsler, 1989	59	61	WPPSI-R	K-ABC	0.9
266	Wechsler, 1999	176	137.52	WASI	WISC-III	0.02
267	Weider et al., 1951	44	77.5	WISC	SB 32	0.47
268	Weider et al., 1951	62	119.5	WISC	SB 32	0.00
269	Weiner & Kaufman, 1979	46	110	WISC-R	WISC	0.32
270	Wheaton et al., 1980	25	119.76	WISC-R	WISC	−0.01
271	Wheaton et al., 1980	25	116.16	WISC-R	WISC	0.36

	Source	N	Age[a]	Newer Test	Older Test	Effect size
272	Whitworth & Gibbons, 1986	25	252	WAIS-R	WAIS	0.18
273	Whitworth & Gibbons, 1986	25	252	WAIS-R	WAIS	0.30
274	Whitworth & Gibbons, 1986	25	252	WAIS-R	WAIS	0.21
275	Whitworth & Chrisman, 1987	30	58	K-ABC	WPPSI	0.35
276	Whitworth & Chrisman, 1987	30	58	K-ABC	WPPSI	0.13
277	Woodcock et al., 2001	150	117.5	WJTCA-III	WISC-III	0.57
278	Woodcock et al., 2001	122	120.6	WJTCA-III	DAS	0.42
279	Yater et al., 1975	20	80.5	WPPSI	WISC	−0.11
280	Yater et al., 1975	20	63.45	WPPSI	WISC	0.23
281	Yater et al., 1975	20	68.15	WPPSI	WISC	−0.24
282	Zimmerman & Woo-Sam, 1974	22	72	SB 72	WPPSI	−0.01
283	Zimmerman & Woo-Sam, 1974	22	66	SB 72	WPPSI	−0.5
284	Zins & Barnett, 1984	40	111	K-ABC	SB 72	0.28
285	Zins & Barnett, 1984	40	111	K-ABC	WISC-R	0.58
	Modem < 5[cc]					
286	Brooks, 1977	30	96	WISC	SB 72	−7.76
287	Carvajal et al., 1991	51	68.4	WPPSI	SB4	2.36
288	Carvajal et al., 1993	32	123	WISC-III	SB4	−0.74
289	Klanderman et al., 1985	41	102	WISC-R	SB 72	6.16
290	Lavin, 1996	40	127.2	WISC-III	SB4	0.28
291	Lukens & Hurrell, 1996	31	161	WISC-III	SB4	2.05
292	McCrowell & Nagle, 1994	30	60	WPPSI-R	SB4	0.63
293	Obrzut et al., 1987	29	114.96	WISC-R	SB 72	17.2
294	Prewett & Matavich, 1994	73	116	WISC-III	SB4	2.23
295	Rust & Lindstrom, 1996	57	111.6	WISC-III	SB4	−0.37
296	Sewell & Manni, 1977	33	84	WISC-R	SB 72	7.4
297	Sewell & Manni, 1977	73	144	WISC-R	SB 72	5.08
298	Simpson et al., 2002	20	108	WISC-III	SB4	1.86

	Source	N	Age[a]	Newer Test	Older Test	Effect size
299	Simpson et al., 2002	20	111	WISC-III	SB4	0.88
300	Wechsler, 1974	29	114	WISC-R	SB 72	6.8
301	Wechsler, 1974	27	150	WISC-R	SB 72	6.8
302	Wechsler, 1974	29	198	WISC-R	SB 72	−8.4
303	Wechsler, 1974	33	72	WISC-R	SB 72	10
304	Wechsler, 1989	115	70	WPPSI-R	SB4	0.69
305	Wechsler, 1991	188	72	WISC-III	WPPSI-R	−4
306	Wechsler, 2003	254	132	WISC-IV	WASI	0.85
307	Wechsler, 2008	141	198	WAIS-IV	WISC-IV	0.28
308	Zins & Barnett, 1984	40	111	WISC-R	SB 72	−10.24
	Other < 5[dd]					
309	Arffa et al., 1984	60	55	WJTCA	SB 72	−0.86
310	Arinoldo, 1982	20	93	WISC-R	MSCA	−6.5
311	Axelrod, 2002	72	644.4	WASI	WAIS-III	−0.98
312	Barclay, 1969	50	63.84	WPPSI	SB 60	1.51
313	Bracken et al., 1984	99	143	WJTCA	WISC-R	2.14
314	Bracken et al., 1984	37	143	WJTCA	WISC-R	1.44
315	Coleman & Harmer, 1985	54	108	WJTCA	WISC-R	1.32
316	Davis, 1975	53	69	SB 72	MSCA	0.4
317	Davis & Walker, 1977	51	97	WISC-R	MSCA	−1.6
318	Dumont et al., 2000	81	148	DAS	WJTCA-R[ee]	−2.8
319	Elliot, 1990	62	63	DAS	WPPSI-R	10.8
320	Elliot, 1990	23	54	DAS	WPPSI-R	5.6
321	Elliot, 1990	58	60	DAS	SB4	0.8
322	Elliot, 1990	55	119	DAS	SB4	1.16
323	Elliot, 1990	29	103	DAS	SB4	1.93
324	Elliot, 2007	95	57.6	DAS-II[ff]	WPPSI-III	0.72
325	Estabrook, 1984	152	120	WJTCA	WISC-R	1.38
326	Fagan et al., 1969	32	65	WPPSI	SB 60	1.62
327	Gregg & Hoy, 1985	50	268.8	WAIS-R	WJTCA	1.06

	Source	N	Age[a]	Newer Test	Older Test	Effect size
328	Harrington et al., 1992	10	36	WPPSI-R	WJTCA-R	−16.4
329	Hayden et al., 1988	32	111.6	SB4	K-ABC	−1.85
330	Hendershott et al., 1990	36	48	SB4	K-ABC	1.81
331	Ingram & Hakari, 1985	33	124.8	WJTCA	WISC-R	0.70
332	Ipsen et al., 1983	27	108	WJTCA	WISC-R	0.68
333	Ipsen et al., 1983	19	108	WJTCA	WISC-R	0.65
334	Ipsen et al., 1983	14	108	WJTCA	WISC-R	0.60
335	Kaufman & Kaufman, 1993	79	204	KAIT	SB4	0.14
336	Kaufman & Kaufman, 2004	86	138	K-ABC-II	WJTCA-III[gg]	−0.09
337	Kaufman & Kaufman, 2004	56	138	K-ABC-II	WISC-IV	−4.6
338	Kaufman & Kaufman, 2004	36	42	K-ABC-II	WPPSI-III	−2.8
339	Kaufman & Kaufman, 2004	39	66	K-ABC-II	WPPSI-III	−7.6
340	Kaufman & Kaufman, 2004	80	136	KBIT-II	WASI[hh]	0.76
341	Kaufman & Kaufman, 2004	62	512	KBIT-II	WASI	1
342	Kaufman & Kaufman, 2004	63	130	KBIT-II	WISC-IV	−1.3
343	King & Smith, 1972	24	72	WPPSI	SB 60	0.74
344	Knight et al., 1990	30	115	SB4	K-ABC	0.54
345	Krohn & Traxler, 1979	22	39	SB 72	MSCA	−1.2
346	Krohn & Traxler, 1979	24	54	SB 72	MSCA	−5.73
347	Krohn & Lamp, 1989	89	59	SB4	K-ABC	0.61
348	Lamp & Krohn, 2001	89	59	SB4	K-ABC	0.56
349	Lamp & Krohn, 2001	72	81	SB4	K-ABC	1.41
350	Lamp & Krohn, 2001	75	104	SB4	K-ABC	0.28
351	Law & Faison, 1996	30	182.4	KAIT	WISC-III	−17.4
352	Naglieri & Harrison, 1979	15	88	SB 72	MSCA	24.26
353	Oakland et al., 1971	24	74	WPPSI	SB 60	0.7
354	Oakland et al., 1971	24	72	WPPSI	SB 60	0.76
355	Pasewark et al., 1971	72	67.11	WPPSI	SB 60	0.78
356	Phelps et al., 1984	55	188	WJTCA	WISC-R	0.54
357	Prosser & Crawford, 1971	50	58	WPPSI	SB 60	1.5
358	Reeve et al., 1979	51	111	WJTCA	WISC-R	3.04

	Source	N	Age[a]	Newer Test	Older Test	Effect size
359	Reilly et al., 1985	26	84	WISC-R	MSCA	2.5
360	Reilly et al., 1985	26	84	WJTCA	WISC-R	−0.65
361	Rellas, 1969	26	76	WPPSI	SB 60	3.40
362	Roid, 2003	145	96	SB5	WJTCA-III	0.46
363	Smith et al., 1989	18	125	SB4	K-ABC	0.48
364	Thompson & Brassard, 1984	20	122.4	WJTCA	WISC-R	0.25
365	Thompson & Brassard, 1984	20	120	WJTCA	WISC-R	2.21
366	Thompson & Brassard, 1984	20	120	WJTCA	WISC-R	2.47
367	Thorndike et al., 1986	175	84	SB4	K-ABC	−0.09
368	Thorndike et al., 1986	30	107	SB4	K-ABC	0.4
369	Vo et al., 1999	30	147	KAIT	WISC-III	−1.34
370	Vo et al., 1999	30	175	KAIT	WISC-III	−4.28
371	Wechsler, 1967	98	66.5	WPPSI	SB 60	0.34
372	Wechsler, 1991	27	108	WISC-III	DAS	−2.8
373	Wechsler, 1999	248	623.76	WASI	WAIS-III	−0.14
374	Ysseldyke et al., 1981	50	123	WJTCA	WISC-R	1.80
375	Zimmerman & Woo-Sam, 1970	26	72	WPPSI	SB 60	1
376	Zimmerman & Woo-Sam, 1970	21	72	WPPSI	SB 60	2.54
377	Zimmerman & Woo-Sam, 1974	22	72	WPPSI	SB 60	1.2
378	Zimmerman & Woo-Sam, 1974	22	66	WPPSI	SB 60	2.54

[a] Age reported in months.

[b] Modern comparisons with at least five years between test norming periods.

[c] Stanford-Binet Intelligence Scales – Fourth Edition.

[d] Stanford-Binet Intelligence Scales – Form L-M (1972 norms ed.).

[e] Wechsler Intelligence Scale for Children-Revised.

[f] Wechsler Adult Intelligence Scale-Revised.

[g] Wechsler Intelligence Scale for Children – Third Edition.

[h] Wechsler Intelligence Scale for Children – Fourth Edition.

[i] Wechsler Adult Intelligence Scale – Third Edition.

[j] Wechsler Preschool and Primary Scale of Intelligence-Revised.

[k] Stanford-Binet Intelligence Scales – Fifth Edition.

[l] Stanford-Binet Intelligence Scales – Form L-M.

[m] Wechsler Preschool and Primary Scale of Intelligence – Third Edition.

[n] Wechsler Adult Intelligence Scale – Fourth Edition.

[o] All other comparisons with at least five years between test norming periods.

[p] Wechsler Intelligence Scale for Children.

[q] McCarthy Scales of Children's Abilities.

[r] Wechsler Preschool and Primary Scale of Intelligence.

[s] Stanford-Binet Intelligence Scales – Form L.

[t] Kaufman Brief Intelligence Test.

[u] Wechsler Adult Intelligence Scale.

[v] Differential Ability Scales.

[w] Stanford-Binet Intelligence Scales – Form L-M (1960).

[x] Kaufman Assessment Battery for Children.

[y] Woodcock-Johnson Tests of Cognitive Abilities.

[z] Kaufman Brief Intelligence Test – Second Edition.

[aa] Kaufman Assessment Battery for Children – Second Edition.

[bb] Kaufman Adolescent and Adult Intelligence Test.

[cc] Modern comparisons with less than five years between test norming periods.

[dd] All other comparisons with less than five years between test norming periods.

[ee] Woodcock-Johnson Tests of Cognitive Abilities-Revised.

[ff] Differential Ability Scales – Second Edition.

[gg] Woodcock-Johnson Tests of Cognitive Abilities – Third Edition.

[hh] Wechsler Abbreviated Scale of Intelligence.

Table 2

Flynn Effect by Sample Type

Sample	N	Mean	SE	Lower CI	Upper CI	*z*	*p*<
Clinical	1	0.36	0.11	0.15	0.57	3.34	0.001
Research	22	0.39	0.08	0.23	0.55	4.76	0.0001
Manuals	30	0.23	0.03	0.17	0.30	7.11	0.0001

Table 3

a

Flynn Effect by Test Group for Counterbalanced Administration Only

Group	N	Point estimate	SE	Lower limit	Upper limit
Modern SB/W[a]	30	0.29	0.03	0.23	0.36
Modern Other[b]	7	0.33	0.08	0.17	0.49
Old SB/W[c]	81	0.26	0.03	0.21	0.31
K-ABC[d]	20	−0.08	0.14	−0.36	0.20
Screening[e]	6	0.09	0.06	−0.02	0.20
McCarthy[f]	12	0.36	0.11	0.15	0.56

b

Flynn Effect by Test Group for Modern Tests with Known Administration Order

Group	N	Point estimate	SE	Lower limit	Upper limit
Flynn effect plus practice effect	8	0.54	0.19	0.16	0.91
Flynn effect less practice effect	12	0.14	0.09	−0.04	0.32
Counterbalanced order	30	0.29	0.03	0.23	0.36

Note. Atypical modern effects have been deleted from these analyses.

[a]Modern SB/W effects include only Stanford-Binet and Wechsler tests normed in 1972 or later.

[b]Modern Other includes other tests normed in 1972 or later.

[c]Old SB/W includes comparisons of Stanford-Binet and Wechsler tests only, where at least one test was normed before 1972.

[d]K-ABC includes comparisons with the K-ABC test.

[e]Screening includes effects on screening instruments.

[f]Remainder includes effects that do not fall into any of the other categories.

APPLIED NEUROPSYCHOLOGY, *16*: 114–123, 2009

ISSN: 0908-4282 print/1532-4826 online
DOI: 10.1080/09084280902864451

Psychology Press
Taylor & Francis Group

Adaptive Behavior Assessment and the Diagnosis of Mental Retardation in Capital Cases

Marc J. Tassé
University of South Florida, Tampa, Florida

There are essentially three main prongs to the definition and diagnosis of the condition known as mental retardation: deficits in intellectual functioning, deficits in adaptive behavior, and onset of these deficits during the developmental period. The U.S. Supreme Court ruled in 2002 in a decision known as *Atkins v. Virginia* that it was essentially cruel and unusual punishment to execute a person with mental retardation, thus violating the Eighth Amendment of the American Constitution. For the purpose of this article, we focused on the issues as they relate to the second prong of the definition of mental retardation, that is, adaptive behavior. We present and discuss the primary concerns and issues related to the assessment of adaptive behavior when making a diagnosis of mental retardation in an *Atkins* claim case. Issues related to standardized assessment instruments, self-report, selection of respondents, use of collateral information, malingering, and clinical judgment are discussed.

Key words: *adaptive behavior, assessment, Atkins, death penalty, diagnosis, forensic, intellectual disability, mental retardation*

INTRODUCTION

Mental retardation[1] is a condition that has been referenced in texts and writings since the dawn of man (Scheerenberger, 1983). There are essentially three main prongs to the definition and diagnosis of the condition known as mental retardation: deficits in intellectual functioning, deficits in adaptive behavior, and onset of these deficits during the developmental period. The U.S. Supreme Court ruled in 2002 in a decision known as *Atkins* that is was cruel and unusual punishment to execute a person with mental retardation, thus violating the Eighth Amendment of the American Constitution. Not surprisingly, there was a swell in the number of referrals and requests for mental retardation evaluations in death penalty cases immediately following this ruling. When making a mental retardation determination within a criminal justice context, the most challenging characteristic for attorneys, judges, and jurors to correctly understand, and for expert clinicians to adequately assess and interpret. is adaptive behavior. This article will focus on discussing the diagnostic issues around the construct of adaptive behavior.

Adaptive behavior is defined as the collection of conceptual, social, and practical skills that have been learned by people to function in their everyday lives (Luckasson, Borthwick-Duffy, Buntinx, Coulter, et al., 2002). Adaptive behavior is a required diagnostic criterion of all systems defining mental retardation (see American Psychiatric Association, 2000; Luckasson et al., 2002; World Health Organization, 1992). Some confusion once existed regarding problem behavior and adaptive behavior, largely because of the misnomer ''maladaptive

[1]The term ''mental retardation'' has acquired such a negative stigma over the years that most professional organizations (American Association on Intellectual and Developmental Disabilities, American Psychological Association) and governmental agencies (e.g., National Institutes of Health, President's Committee for Persons with Intellectual Disability) have adopted ''intellectual disability'' as the new terminology to designate the condition previously known as mental retardation.

Address correspondence to Marc J. Tassé, Phd, Florida Center for Inclusive Communities–UCEDP, University of South Florida, 13301 Bruce B. Downs Blvd., Tampa, FL 33612. E-mail: mtasse@fmhi.usf.edu

behavior'' that was once used to designate problem behaviors such as self-injurious behavior, aggression, stereotypies, destruction of property, etc. ''Maladaptive behavior'' is a separate and independent construct of adaptive behavior (Luckasson et al., 2002; Schalock, Buntinx, Borthwick-Duffy, Luckasson, et al. 2007). The presence or absence of ''maladaptive behaviors'' has little relationship to an individual's adaptive functioning. These behaviors can occur in individuals with poor adaptive behavior (e.g., someone bangs their head because they are unable to communicate that they have a headache), and they can occur in individuals with good adaptive behavior, but for whom they are associated to a cooccurring mental health problem (e.g., depression and aggressive behavior). ''Maladaptive behaviors'' are not part of the diagnostic criteria of mental retardation.

The American Association on Intellectual and Development Disabilities (AAIDD) was the first organization, almost 50 years ago, to introduce adaptive behavior as a diagnostic criteria of mental retardation (see Heber, 1959, 1961). In fact, Heber (1959) first defined adaptive behavior much the same way as it is currently defined in the most recent edition of the AAIDD Terminology and Classification manual (Luckasson et al., 2002). Heber (1959) first introduced the concept of adaptive behavior into the AAIDD terminology and classification manual as reflected by impairments in ''maturation, learning, and social adjustment.'' These three domains were later collapsed into a unitary construct identified as ''adaptive behavior'' (Heber, 1961). More than 40 years later, AAIDD has returned to Heber's (1959) original conceptualization of adaptive behavior as practical, conceptual, and social skills. Although the U.S. Supreme Court declined to provide a specific definition of mental retardation in their *Atkins* decision, they did cite both the American Psychiatric Association (2000; Diagnostic and Statistical Manual for Mental Disorders, Fourth Edition [DSM-IV-TR]) and the AAIDD (Luckasson Coulter, Polloway, Reiss, et al., 1992) diagnostic criteria. Writing for the majority, Justice Stevens stipulated that ''As discussed above, clinical definitions of mental retardation require not only subaverage intellectual functioning, but also significant limitations in adaptive skills such as communication, self-care, and self-direction, the become manifest before age 18'' (*Atkins v. Virginia*, 536 U.S. 304, 2002, p. 318).

There are two other large organizations that have conducted systematic reviews of the literature and proposed guidelines for defining mental retardation: the American Psychological Association's Division 33 and the Social Security Administration. The American Psychological Association's (APA) Division 33 (Intellectual and Developmental Disabilities) panel reviewed the literature and proposed a definition and diagnostic criteria for mental retardation. Their definition was published as a chapter in a handbook on mental retardation (see Barclay, Drotar, Favell, Foxx, et al., 1996). The APA Division 33 panel proposed a three-prong definition of mental retardation that was congruent with the American Psychiatric Association (2000), World Health Organization (1992), and AAIDD (Luckasson et al., 2002) definitions. The APA Division 33 definition included significant deficits in intellectual functioning, significant deficits in adaptive behavior, and an onset of these significant limitations during the developmental period (see Barclay et al., 1996).

The U.S. Social Security Administration (SSA) convened a panel of experts to review the existing literature and propose recommendations to the SSA regarding criteria to identify individuals as having mental retardation (see Reschly, Myers, & Hartel, 2002). Although not meant as a diagnostic system but as recommendations to develop the eligibility criteria to receive SSA benefits under the classification of mental retardation, Reschly and his colleagues proposed a definition that included the same three prongs (Intellectual functioning, adaptive behavior, and age of onset).

Although there remains minor discrepancies in how each of these systems has operationally defined each of the three prongs, the consensus regarding the diagnosis of mental retardation is that there needs to be the presence of deficits in both intellectual functioning and adaptive behavior, and these deficits must have originated during the developmental period. It should be noted that ''originated during the developmental period'' does not preclude making a first time diagnosis of mental retardation when an individual is an adult. The clinician must, however, adequately document that the deficits in intellectual and adaptive functioning were present before the end of the developmental period.

AAIDD (formerly, the American Association on Mental Retardation) is generally regarded as the leading authority in defining mental retardation. The APA publishes the main diagnostic manual for the field of psychiatry entitled the Diagnostic and Statistical Manual for Mental Disorders, which is currently in its fourth edition. It should be pointed out that the DSM-IV-TR contains information on almost 300 disorders, of which, mental retardation is one. The AAIDD has been solely focused for the past 100 years on defining mental retardation. It is not surprising that, historically, the APA and the DSM panel have largely adopted the AAIDD definition and diagnostic criteria of mental retardation in their revisions of the DSM. This is illustrated in the most recent revision of the DSM, the DSM-IV-TR. The DSM-IV-TR (American Psychiatric Association, 2000) adopted the AAIDD (Luckasson et al., 1992) definition and changed it's conceptualization of adaptive behavior to reflect Luckasson et al.'s (1992) definition of adaptive behavior, which consisted of 10 adaptive skill areas. The AAIDD

1992 manual (Luckasson et al., 1992) defined the second prong of the definition as ''limitations in two or more of the following adaptive skill areas: communication, self-care, home living, social skills, community use, self-direction, health and safety, functional academics, leisure, and work'' (p. 1).

Probably due to a misplaced comma, the DSM-IV-TR actually defined adaptive behavior deficits as limitations in two or more of 11 skill areas (instead of 10 skill areas), having placed a comma between ''health'' and ''safety'' (American Psychiatric Association, 2000, p, 49). The DSM-IV-TR can be cited as follows:

> Concurrent deficits or impairments in present adaptive functioning (i.e., the person's effectiveness in meeting the standards expected for his or her age by his or her cultural group) in at least two of the following areas: communication, self-care, home living, social/interpersonal skills, use of community resources, self-direction, functional academic skills, work, leisure, health, and safety (American Psychiatric Association, 2000, p. 49).

The DSM-IV-TR diagnostic criteria and, the then most current AAIDD diagnostic criteria (Luckasson et al., 1992) were virtually identical. In the 10th edition of its Terminology and Classification manual in 2002 (see Luckasson et al., 2002), AAIDD moved away from the 10 adaptive skill areas to a more psychometrically grounded definition of adaptive behavior consisting of three domains: conceptual, practical, and social adaptive skills. It should be noted that many had acknowledged that the previous 10 adaptive skill areas were not supported by the existing psychometric literature in the field (Heal & Tassé, 1999; Luckasson et al., 2002; Spreat, 1999; Thompson, McGrew, & Bruininks, 1999; Widaman & McGrew, 1996).

Although the assessment of intellectual functioning has a longer history than does the assessment of adaptive behavior, the psychometric properties of adaptive behavior instruments have improved significantly since the Vineland Social Maturity Scale (Doll, 1936) was first published. When Edgar Doll first published the Vineland Social Maturity Scale in 1936 (this test later evolved into the Vineland Adaptive Behavior Scales), he defined a construct that he labeled ''social competence.'' The first version of his instrument consisted of items organized into six broad domains (self-help: general, dressing, and eating; self-direction; communication; socialization; motor; and work). Doll (1953) defined social competence as ''the functional ability of the human organism for exercising personal independence and social responsibility'' (p. 10). Doll's vision of assessing social competence (now called adaptive behavior) remains ingrained in today's definitions of adaptive behavior and assessment instruments. For example, Doll wrote: ''Our task was to measure attainment in social competence considered as habitual performance rather than as latent ability or capacity'' (Doll, 1953, p. 5). This view is consistent with AAIDD's long standing position that adaptive behavior assessment must focus on the individual's typical performance and not maximal ability (see Luckasson et al., 2002).

The reliance on standardized measures of adaptive behavior as part of the mental retardation diagnostic process was first prescribed by Barclay et al. (1996) in their definition endorsed by APA's Division 33. AAIDD (Luckasson et al., 2002) and Reschly, Myers, and Hartel (2002) reiterated the importance of establishing that the individual has ''significant limitations'' in adaptive behavior based on the results of an individually administered measure of adaptive behavior. Luckasson et al. also emphasized the importance of using standardized adaptive measures that had been normed on the general population and assessed the broad array of adaptive behavior, including conceptual, practical, and social skills.

The use of a standardized adaptive behavior scale is often insufficient to capture all aspects of an individual's adaptive behavior. Elements of adaptive behavior that are related to adult social adaptive skills or higher order interpersonal skills are lacking from most existing adaptive behavior scales (Duffy, 2007; Luckasson et al., 2002; Reschly, Myers, & Hartel, 2002). Greenspan (Greenspan, 1981; Greenspan, 2006; Greenspan, 2008; Greenspan, Loughlin, & Black, 2001; Greenspan & Switzky, 2006) has devoted much of his career to studying and publishing on concepts that are often present in individuals with mild mental retardation, but under-represented in standardized adaptive behavior scales: social competence, gullibility, naïveté, and lack of wariness.

We will not provide an exhaustive review of the existing adaptive behavior instruments in this article. The interested reader is encouraged to consult previously published articles that have already provided excellent reviews (Luckasson et al., 2002; Reschly, Myers, & Hartel, 2002; Stevens & Price, 2006). Rather, we will focus on discussing measurement issues that are most relevant when assessing adaptive behavior for the purpose of making or ruling out a diagnosis of mental retardation.

Our recommendations are applicable to any clinical diagnosis of mental retardation but we will pay special attention to the particular challenges that are posed when the assessed individual is facing criminal charges and is incarcerated at the time of the evaluation. Several authors have long argued the mitigating circumstances of mental retardation for individuals involved in the criminal justice system (Ellis & Luckasson, 1985; Keyes, Edwards, & Perske, 1997). Since the *Atkins* decision there has been an increased interest in mental retardation in individuals in capital cases or in those awaiting the death penalty. Any case involving a diagnosis of mental retardation should be considered as ''high

stakes,'' and, as such, clinicians should always use the utmost prudence and rigor in conducting these diagnostic evaluations. Nonetheless, no one can deny that an ''Atkins claim'' is the highest of high stakes.

Making a diagnosis of mental retardation in individuals who have severe or profound deficits in intellectual functioning and adaptive behavior is relatively easy. Conversely, it is relatively straightforward to rule-out a diagnosis of mental retardation for someone whose general intellectual functioning and adaptive behavior levels are all measured to be in the low average range. Some of the more challenging conditions for making or ruling-out a diagnosis of mental retardation include individuals whose intellectual functioning and adaptive behavior are at or around the cut-off of two standard deviations below the population mean (Reschly, Myers, & Hartel, 2002; Schalock et al., 2007). It should be noted that the vast majority of individuals with mental retardation (i.e., 85%) are in this range of functioning, at times referred to as ''mild mental retardation'' (American Psychiatric Association, 2000). The vast majority of ''Atkins claims,'' if not all, will likely involve individuals who have intellectual and adaptive functioning levels that are near the diagnostic cut-off range.

ASSESSMENT OF ADAPTIVE BEHAVIOR

Anyone conducting an adaptive behavior assessment is strongly encouraged to consult the chapter by Harrison and Raineri (2008) on the Best Practices in the Assessment of Adaptive Behavior. This chapter reviews the salient assessment issues to consider when assessing adaptive behavior for the purpose of diagnosing or ruling out mental retardation in an individual.

Two of the more challenging aspects of any adaptive behavior assessment of an individual who is incarcerated include: the assessment of the individual's present functioning and the assessment of the individual's typical behavior in meeting community demands and expectations. By definition, the construct of adaptive behavior involves age-indexed skills that are learned and performed to meet the demands and expectations of society and the community across life settings (i.e., home, school, work, community). Thus, assessment of adaptive behavior for the purpose of making a diagnosis of mental retardation involves assessing the individual's present, typical behavior. as well as the individual's functioning as it occurs in the community. It is not a measure of capacity or knowledge, but in fact is a measure of what the individual typically does and what is the degree of independence in performing these skills.

Other important aspects of adaptive behavior assessment that need to be addressed when making or ruling out a diagnosis of mental retardation include:

- assessment the individual's adaptive behavior in relation to his age group and culture
- use of standardized adaptive behavior scale that was normed on the general population
- obtaining corroborating information to support the information obtained on the standardized assessment

Stevens and Price (2006) recommended that future research in the area of adaptive behavior assessment should develop norms on prison populations. This author strongly disagrees with this notion. Norming an adaptive behavior scale on people living in prisons would have as much value as norming a new IQ test on people living in prisons. One would only know if the assessed person is more *intelligent* or more *adapted* than prison inmates. An adaptive behavior instrument normed solely on inmates or another institutional population (e.g., State Mental Retardation Center) would have little relevance when attempting to measure the skills someone has learned and performs to meet societal demands and expectations for someone his or her age and cultural group.

The Adaptive Behavior Scale–Residential and Community Edition (ABS-RC:2; Nihira, Leland, & Lambert, 1993) is normed on individuals with mental retardation (living in the community and in institutional/residential settings). Because of this reason it is an inappropriate instrument to be used in assessing adaptive behavior for the purpose of making or ruling out a diagnosis of mental retardation. However, the ABS-RC:2 has a recognized clinical value when used to assess an individual's adaptive behavior to establish intervention goals and determine the individual's adaptive functioning when compared to other adults with mental retardation.

Fabian (2005) raised the question of the relevance of current adaptive behavior scales since none included individuals on death row in their normative samples. This pushes the aforementioned point one step further. It is important to keep in mind that there are approximately 300 million Americans, of which approximately 3 million have a diagnosis of mental retardation (see Larson et al. 2001). Of that number, a miniscule fraction of all Americans or Americans with mental retardation live on a death row. It is probably safe to say that there will never be a standardized adaptive behavior scale (or a standardized IQ test for that matter) that has any significant representation of individuals living on a death row in its normative sample. The bigger threat to our ability to rely unequivocally on the resulting scores obtained on standardized adaptive behavior scales is more likely to stem from the violations of the instrument's administration procedures. These include: being unable to assess the individual's present adaptive behavior, being unable to assess the adaptive behavior as it

typically occurs in a naturalistic setting such as the community at large, using the instrument to conduct direct testing of an individual's adaptive skills, conducting an adaptive behavior semi-structured interview without having properly established and maintained rapport with the respondent, and relying on protocols in which the respondent provided numerous "guessed" estimates rather than relying on actual observation of the individual's behavior (Harrison & Oakland (2003) cautioned against the results stemming from protocols on which the respondent guessed on more than three items in a skill domain).

USE OF CONVERGENT INFORMATION

There exists no one standardized adaptive behavior scale that captures the entire spectrum of adaptive behavior across all age groups (Luckasson et al., 2002; Thompson, McGrew, & Bruininks, 1999). This does not, however, negate the importance of using such measures when possible. Rather, any comprehensive evaluation of adaptive behavior should seek to corroborate information obtained on standardized measures from sources such as: school records, employment history, social security administration records, medical records, and interviews with respondents who know the individual well but who might not be able to provide comprehensive information sufficient to complete all domains on an adaptive behavior scale. In addition to the use of standardized measures of adaptive behavior, it is crucial to obtaining corroborating information from other sources. For example, the individual's school records can provide a wealth of information regarding conceptual, practical, and social skills. It will be necessary to also consult social security administration records, driving record, employment history, medical records, and social and family history. In addition to interviewing individuals to complete a standardized adaptive behavior scale, it is vital to conduct clinical interviews of relatives, friends, teachers, coaches, employers, roommates, etc. in order to obtain some qualitative information regarding the individual's adaptive behavior. This information can be crucial in providing corroborating information regarding areas of limitations and strengths.

Thus, in addition to an appropriate standardized adaptive behavior scale, any comprehensive assessment of adaptive behavior assessment should include the following information:

- qualitative adaptive behavior interviews with multiple informants who have observed the assessed person in different contexts (e.g., home, school, work, leisure, community)
- review of family history
- review of available school records (e.g., transcripts, psychoeducational evaluations, Individual Education Plans, etc.)
- review of available medical records
- review of all federal and state agency records (e.g., Social Security Administration, Department of Social Services, Department of Motor Vehicles, State Department of Mental Retardation/Developmental Disabilities, Division of Vocational Rehabilitation, prison records, etc.)
- review of employment history/records
- review of all previous psychological/psychiatric/psychosocial evaluations

ADMINISTRATION OF A STANDARDIZED ADAPTIVE BEHAVIOR SCALE

There are at present perhaps four well-known and often-used standardized adaptive behavior scales for the purpose of making or ruling out a diagnosis of mental retardation: Scales of Independent Behavior–Revised (SIB-R; Bruininks, Woodcock, Weatherman, Hill, 1996), Adaptive Behavior Assessment System–2nd Edition (ABAS-2; Harrison & Oakland, 2003). Vineland Adaptive Behavior Scales–2nd Edition (Vineland-II; Sparrow, Cicchetti, & Balla, 2005), and Adaptive Behavior Scale–School Edition (ABS-S:2; Lambert, Nihira, & Leland, 1993). The latter instrument, although used with some frequency in the schools, it is less well known in the forensic setting.

SEMI-STRUCTURED INTERVIEWS VERSUS RATING SCALE ADMINISTRATIONS

All these instruments can be used as a rating scale—that is, given directly to the respondent who reads and responds to the items on their own. It should however be noted that most would agree that there is added value to administering these instruments via a semi-structured interview. For example, the SIB-R provides an easel for interview administration and recommends using the interview format with respondents who do not have prior experience with adaptive behavior assessments (Bruininks, Woodcock, Weatherman, & Hill, 1996). Harrison and Oakland (2003) pointed out that their scale is written at a fifth-grade reading level and some respondents may have difficulty reading and rating the item stems. Perhaps the most comprehensive analysis of the merits of a semi-structured interview administration of an adaptive behavior scale is provided by Sparrow and her colleagues (2005) in the Vineland-II Manual. The Vineland-II Manual recommends use of a semi-structured interview format over the rating scale

format when the adaptive behavior assessment is for purposes of establishing or ruling out a diagnosis of mental retardation. Sparrow et al., stated ''the strength of the semi-structured interview format in eliciting accurate, in-depth descriptions of the individual's functioning make it the preferred method when the results will inform diagnostic decisions.''

Other advantages of administering a standardized adaptive behavior scale via a semi-structured interview, instead of giving the rating scale directly to the respondent to complete on their own, include the following:

- reduces likelihood of reading error on the part of the respondent
- provides an immediate opportunity to address questions about an item stem or provide clarifying information if the respondent appears confused or uncertain regarding the content of the item
- provides the examiner with the opportunity to observe the latency between the reading of the item and the response, which gives an indication of the time taken to think about the item stems before providing a response
- allows the examiner to monitor the respondent's attention and tailor the pace of administration to the respondent's needs
- allows the examiner the opportunity to probe some responses and assess the reliability of the respondent

Selection of Respondents

The ideal respondents are individuals who have the most knowledge of the individual's everyday functioning across settings. Typically, the individual's parents or caregivers are the persons with the most opportunity to observe the assessed individual in his/her everyday functioning. As the individual becomes an adult, this role may shift to a spouse or roommate. Other individuals who may provide valuable adaptive behavior information include: older siblings, grandparents, aunts/uncles, neighbors, teachers, coaches, employers, coworkers, friends, or other adults who may have had multiple opportunities over an extended period of time to observe the individual in his everyday functioning in one or more contexts (e.g., home, leisure, school, work, community).

Correctional Officers as Respondents

Correctional officers and other prison personnel should probably never be sought as respondents to provide information regarding the adaptive behavior of an individual that they've observed in a prison setting. The only extreme circumstance when one might consider interviewing a member of the prison personnel regarding an inmate's adaptive behavior would be if there is absolutely no one alive who can provide any information regarding the individual's functioning prior to incarceration. The main hesitation to involving prison personnel as respondents is related to the nature and contingencies of the prison setting. The prison setting is an artificial environment that offers limited opportunities for many activities and behaviors defining adaptive behavior. In the end, adaptive behavior information obtained from prison personnel should be limited to activities or behaviors that they have had the opportunity to directly observe the individual perform. It should be noted that items cannot be truncated or substituted for setting equivalents. For example, the ABAS-II has an item on the Community Use subscale that assesses the individual's performance regarding mailing a letter in a mailbox or the local post office. This would be an item that is most likely impossible to observe in a prison setting and should not be substituted for anything other than what the stem specifies.

Faking Adaptive Deficits

We usually associate malingering or ''faking bad'' to the feigning of symptoms to appear ill to obtain something desired (e.g. compensation) or to avoid a punishment (e.g., prosecution; American Psychiatric Association, 2000). There is some concern that the individual being assessed for a mental retardation determination might malinger on the IQ test or self-report fewer adaptive skills than he actually possesses in order to meet criteria for a diagnosis of mental retardation. When assessing intellectual functioning, clinicians will generally include one or more measures of effort in an attempt to gauge whether or not the individual is trying his best. Depending on the outcome on these measures, the examiner will generalize that effort to the individual's performance on the test of intelligence.

Malingering may also be a real issue in the case of a self-reported assessment of adaptive behavior. Some adaptive behavior instruments may be more vulnerable than others to a malingered self-report (Doane & Salekin, in press). Relying solely on the individual's self-report is fraught with problems (Patton & Keyes, 2006; Schalock et al., 2007). In fact, as many researchers have documented numerous times, individuals with low IQ may not always be reliable self-reporters. For example, Edgerton (1967) documented that individuals with mild mental retardation often over-estimated their skills and abilities and attempted to conceal their disability to avoid stigmatization. According to Edgerton's groundbreaking research, individuals with mental retardation are perhaps more likely to ''fake good'' on measures of adaptive behavior.

If conducted improperly, adaptive behavior interviews of individuals with mental retardation can yield invalid results. One study comparing self-reported adaptive behavior with respondent ratings showed that individuals with mental retardation showed good agreement with the respondent's ratings of the individual's adaptive behavior (Voelker et al., 1990). Research has also shown that individuals with mental retardation are particularly susceptible to acquiescence and leading questions (Everington & Fulero, 1999; Finlay & Lyons, 2002; Perry, 2004). Individuals with mental retardation often respond in the affirmative to questions they don't fully understand or might not be sure of the correct answer (Finlay & Lyons, 2002). Someone unfamiliar with these characteristics of individuals with mental retardation may misinterpret the individual's actual adaptive behavior. Having reviewed records and interviewed other respondents before conducting the self-report may provide insight into evaluating the reliability of the self-report and provide point upon which to probe the individual to verify the skill. The only standardized adaptive behavior scale that was normed using self-report data is the ABAS-II (Harrison & Oakland, 2003).

It is more common that the respondent is someone other than the assessed individual. The clinician must always assess the respondent's reliability in providing adaptive behavior information. In the capital cases there is a particular worry regarding the bias introduced by family members in reporting on the adaptive behavior of their loved one. This might be interpreted as a form of malingering by proxy, where a parent might want to under-report adaptive skills to intentionally lower their loved one's adaptive behavior performance, in order to increase the likelihood of a diagnosis of mental retardation and result in a reprieve of the death penalty. Again, best practice is to obtain adaptive behavior information from multiple respondents and multiple sources in order to obtain a complete evaluation and identify areas of convergence (Harrison & Oakland, 2003; Schalock et al., 2007).

Retrospective Assessment

A retrospective assessment of adaptive behavior is often considered as the only viable option when the assessed individual is incarcerated. Interviewing a respondent while asking them to recall a time prior to the individual's incarceration is the proposed means of capturing the individual's typical adaptive behavior in the community and establishing a retrospective diagnosis (Schalock et al., 2007). It should be noted that there is no research available examining the reliability or error rate of adaptive behavior assessments obtained retrospectively. At issue is the respondent's ability to correctly recall from memory the assessed individual's actual performance. Memory degradation is a real issue and we do not have any solid research regarding the forgetting curve (Memon & Henderson, 2002) regarding someone's recollection of another person's adaptive behavior.

A retrospective adaptive behavior assessment can be challenging (Everington & Olley, 2008). To assist the clinician with this difficult task, Schalock et al. (2007) recommended specific guidelines to follow when making a retrospective diagnosis of mental retardation, including using multiple respondents and multiple contexts and assessing adaptive functioning within the general community and within the individual's age peers and cultural group. This last point is in reference to the expectations being different in certain cultural groups for specific adaptive behaviors, from mainstream America. For example, using a fork and knife to eat may not be a prerequisite to be adaptive to societal demands in certain cultures (e.g., Asian). To these guidelines, one might add the following instructions when conducting a retrospective adaptive behavior assessment:

- Identify a clear time period during which you want the respondent to focus their report of the individual's adaptive behavior. For example, you might instruct the respondent to recall the assessed individual before he was incarcerated.
- Build rapport with the respondent and ask them to think about where the assessed person was living at that specified time, working, etc. These points of reference will be important to assist the respondent to recall that time period.
- Periodically, remind the respondent that they are assessing the individual's adaptive behavior in that specific time period.

There may be instances when completing a standardized adaptive behavior scale is not possible. It might be that there is no one alive or available to participate as a respondent. Another reason might be that the respondents available are not able to provide a comprehensive picture of the individual's adaptive behavior such that they can complete all the information needed on a standardized scale. It is important for the clinician to use his or her clinical judgment in determining when it is viable to conduct a standardized adaptive behavior scale and when it is not. In the latter case, it is possible to conduct a series of semi-structured interviews with multiple respondents who have reliable information about specific periods of time (e.g., when he was in elementary school) or have knowledge of the individual in one specific context (e.g., when he worked at the local car wash). This information, along with case records, can be helpful in contributing to developing a report regarding the individual's adaptive behavior.

The Role of Clinical Judgment

Professionals should always use clinical judgment throughout the process of making or ruling out a diagnosis of mental retardation. One uses their clinical judgment in selecting an appropriate adaptive behavior assessment instrument, identifying who to interview as a respondent, assessing the respondent's reliability, identifying and reviewing available records, and analyzing and interpreting all the available information to form an opinion. Schalock and Luckasson (2005) defined clinical judgment as being founded upon clinical expertise in a particular area and that clinical judgment is based upon a thorough analysis of extensive data. Equally important, these authors state that ''Clinical judgment should not be thought of as a justification for abbreviated evaluations, a vehicle for stereotypes or prejudices, a substitute for insufficiently explored questions, an excuse for incomplete or missing data, or a way to solve political problems'' (p. 6). Hence, clinical judgment should not be used as a shield when one draws conclusions that are not supported by the assessment results, observations, and/or case records.

DISCUSSION

Making a diagnosis of mental retardation is often challenging and should only be conducted by qualified professionals. Most individuals with mental retardation will have strengths and areas of ability (see Luckasson et al., 2002). These strengths may confound a layperson or a professional with limited clinical experience with individuals who have mild mental retardation. These laypersons may erroneously interpret these pockets of strengths and skills as inconsistent with mental retardation because of their misconceptions regarding what someone with mental retardation can or cannot do. For example, many laypeople believe that individuals with mental retardation cannot read. In fact, it is well established that adults with mild mental retardation can achieve reading and writing commensurate with a grade equivalent of fifth or sixth grade (American Psychiatric Association, 2000; Barclay et al., 1996).

Mental retardation is a clinical diagnosis that should be made or ruled out based on a rigorous and comprehensive professional evaluation of the individual's intellectual functioning and adaptive behavior. If there is a presence of significant deficits, there must be an ascertainment that these deficits were manifest prior to age 18. A person who has been appropriately diagnosed with mental retardation should be identified as having mental retardation regardless of the individual's living arrangement, accommodations, or supports in place that could very well result in better functioning. AAIDD (Luckasson et al., 2002) reminded everyone in their section on the assumptions regarding mental retardation that ''Within an individual, limitations often coexist with strengths,'' and ''With appropriate personalized supports over a sustained period, the life functioning of the person with mental retardation generally will improve'' (Luckasson et al., 2002, p. 1).

Adaptive behavior is best represented by conceptual, practical, and social skills that an individual has learned and typically performs in order to meet societal demands in naturalistic settings (Luckasson et al., 2002). When we assess adaptive behavior for the purpose of making or ruling out a diagnosis of mental retardation, the use of standardized adaptive behavior scales is often central since they provide an objective metric with which to determine whether or not the individual's limitations are significantly below the average of the general population. The information obtained from standardized adaptive behavior scales should be corroborated with information from other sources, such as interviews with other informants and a thorough review of records and previous evaluations.

Assessment of adaptive behavior needs to be conducted using a combination of standardized adaptive behavior scales, adaptive behavior interviews of multiple informants who have observed the individual in different contexts, and a review of all available records. The standardized instrument is not error-free. The results obtained on a standardized adaptive behavior scale must be interpreted in relation to the instrument's reliability and resulting standard error of measurement. Self-ratings on standardized adaptive behavior scales are fraught with potential problems and should be interpreted with caution.

Any breach in administration procedures of a standardization assessment instrument should be clearly documented in the clinician's report, and the results should be interpreted with a certain degree of prudence. Because of the nature of Atkins claims, it is often necessary to conduct a retrospective adaptive behavior assessment. Retrospective adaptive behavior assessments should be well-documented with respect to respondents interviewed, procedure used, assessed time-frame (e.g., when individual was 17 years old), normative group used to interpret results, and source of convergent information that corroborates or contradicts results obtained. As with any type of adaptive behavior assessment, multiple respondents should be used and these respondents should preferably have had the opportunity to observe the assessed individual in different contexts. Results from a retrospective evaluation should be interpreted with caution.

Making a diagnosis of mental retardation is not like baking a cake, where one opens a book, follows the

prescribed instructions, and out comes the certainty of whether or not a diagnosis such as mental retardation exists. Making a diagnosis of mild mental retardation is one of the more challenging diagnoses to make (Schalock et al., 2007). Most forensic psychologists have broad clinical training as well as training and experience to work with the courts and criminal defendants. Mental retardation professionals often have training and experience in working with individuals with and without mental retardation, but lack the training regarding the forensic science. The *Atkins* Supreme Court decision has resulted in the bridging of two fields: forensic psychology and the interdisciplinary field of mental retardation. Perhaps it is time to answer Everington and Olley's (2008) call for forensic and mental retardation professionals to join forces and provide leadership in developing practice guidelines for the diagnosis of mental retardation in the forensic setting. Such proposed practice guidelines should build upon an established national standard for diagnosing mental retardation (such as the AAIDD system), or else we risk creating a clinical diagnosis and a forensic diagnosis of mental retardation.

REFERENCES

American Psychiatric Association (2000). *Diagnostic and statistical manual of mental disorders* (4th ed. text revision; DSM-IV-TR). Washington, DC: Author.

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

Barclay, A. G., Drotar, D. D., Favell, J., Foxx, R. M., Gardner, W. I., Iwata, B. A., Jacobson, J. W., Matson, J. L., Mulick, J. A., Ramey, S. L., Routh. D. K., Schroeder, S. R., Sprague, R. L., Switsky, H. N., & Thompson, T. (1996). Definition of mental retardation. In J. W. Jacobson & J. A. Mulick (Eds.), *Manual of diagnosis and professional practice in mental retardation* (pp. 13–47). Washington, DC: American Psychological Association.

Bruininks, R. H., Woodcock, R., Weatherman, R. F., & Hill, B. K. (1996). *Scales of Independent Behavior Revised (SIB-R)*. Itasca, IL: Riverside Publishing.

Doane, B. M., & Salekin, K. L. (in press). Susceptibility of current adaptive behavior measures to feigned deficits. Law and Human Behavior.

Doll, E. A. (1936). *The Vineland Social Maturity Scale: Revised condensed manual of directions*. Vineland, NJ: The Vineland Training School.

Doll, E. A. (1953). *Measurement of social competence: A manual for the Vineland Social Maturity Scale*. Vineland, NJ: Educational Publishers, Inc.

Duffy, S. A. (2007). Adaptive behavior. In J. W. Jacobson, J. A. Mulick, & J. Rojahn (Eds.), *Handbook on intellectual and developmental disabilities* (pp. 279–291). Washington, DC: Springer.

Edgerton, R. B. (1967). *The cloak of competence: Stigma in the lives of the mentally retarded*. Berkeley: University of California Press, Ltd.

Ellis, J., & Luckasson, R. (1985). Mentally retarded criminal defendants. *George Washington Law Review*, *53*, 414–493.

Everington, A., & Fulero, S. M. (1999). Competence to confess: Measuring understanding and suggestibility of defendant with mental retardation. *Mental Retardation*, *37*, 212–220.

Everington, A., & Olley, J. G. (2008). Implications of *Atkins v. Virginia*: Issues in defining and diagnosing mental retardation. *Journal of Forensic Psychology Practice*, *8*(1), 1–23.

Fabian, J. M. (2005). Life, death, and IQ; it's much more that just a score: The dilemma of the mentally retarded on death row. *Journal of Forensic Psychology Practice*, *5*(4), 1–36.

Finlay, W. M. L., & Lyons, E. (2002). Acquience in interviews with people who have mental retardation. *Mental Retardation*, *40*, 14–29.

Greenspan, S. (2008). Foolish action in adults with intellectual disabilities: The forgotten problem of risk-unawareness. In L. M. Glidden (Ed.), *International review of research in intellectual and developmental disabilities*.

Greenspan, S. (1981). Social competence and handicapped individuals: Practical implications of a proposed model. In B. K. Keough (Ed.), *Advances in special education* (Vol. 3) (pp. 41–82). Greenwich, CT: JAI Press.

Greenspan, S. (2006). Functional concepts in mental retardation: Finding the natural essence of an artificial category. *Exceptionality*, *14*, 205–224.

Greenspan, S., Loughlin, G., & Black, R. (2001). Credulity and gullibility in persons with mental retardation. In L. M. Glidden (Ed.), *International review of research in mental retardation*. New York: Academic Press.

Greenspan, S., & Switzky, H. (2006). In H. N. Switzky & S. Greenspan (Eds.), *What is mental retardation: Ideas for an evolving disability in the 21st century*. Washington, DC: American Association on Mental Retardation.

Harrison, P. L., & Oakland, T. (2003). *Adaptive behavior assessment system, Second edition: Manual*. San Antonio, TX: Harcourt Assessment, Inc.

Harrison, P. L., & Raineri, G. (2008). Best practices in the assessment of adaptive behavior. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology* (5th ed.) (pp. 605–616). Bethesda, MD: NASP Press.

Heal, L. W., & Tassé, M. J. (1999). The culturally sensitive individualized assessment of adaptive behavior. In R. L. Schalock (Ed.), *Adaptive behavior and its measurement: Implications for the field of mental retardation* (pp. 185–208). Washington, DC: American Association on Mental Retardation.

Heber, R. (1959). A manual on terminology and classification in mental retardation: A monograph supplement. *American Journal on Mental Deficiency*, *64* (Monograph Suppl.).

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

Keyes, D., Edwards, W., & Perske, R. (1997). People with mental retardation are dying, legally. *Mental Retardation*, *35*, 59–63.

Lambert, N., Nihira, K., & Leland, H. (1993). *Adaptive Behavior Scale–School edition (ABS-S:2)*. Austin, TX: Pro-Ed Publishing.

Larson, S. A., Lakin, K. C., Anderson, L., Kwak, N., Lee, J. H., & Anderson, D. (2001). Prevalence of mental retardation and developmental disabilities: Estimates from 1994/1995 National Health Survey Disability Supplement. *American Journal on Mental Retardation*, *106*, 231–252.

Luckasson, R., Coulter, D. L., Polloway, E. A., Reiss, S., Schalock, R. L., Snell, M. E., Spitalnik, D. M., & Stark, J. A. (1992). *Mental retardation: Definition, classification, and systems of supports* (9th ed.). Washington, DC: American Association on Mental Retardation.

Luckasson, R., Borthwick-Duffy, S., Buntinx, W. H. E., Coulter, D. L., Craig, E. M., Schalock, R. L., Snell, M. E., Spitalnik, D. M., Spreat, S., & Tassé, M. J. (2002). *Mental retardation: Definition, classification, and systems of supports* (10th ed.). Washington, DC: American Association on Mental Retardation.

Memon, A., & Henderson, S. E. (2002). What can psychologists contribute to the examination of memory and past mental status? In R. I. Simon & D. W. Shauman (Eds.), *Retrospective assessment of mental states in litigation: Predicting the pas*t pp. 307–334). Washington, DC: American Psychiatric Publishing, Inc.

Nihira, K., Leland, H., & Lambert, N. (1993). *Adaptive Behavior Scale–Residential and community edition (ABS-RC: 2)*. Austin, TX: Pro-Ed Publishing.

Patton, J. R., & Keyes, D. W. (2006). Death penalty issues following Atkins. *Exceptionality*, *14*, 237–255.

Perry, J. (2004). Interviewing people with intellectual disabilities. In E. Emerson, C. Halton, T. Thompson, & T. Parmenter (Eds.), *International handbook of applied research on intellectual disabilities* (pp. 115–132). West Sussex, UK: John Wiley & Sons Ltd.

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

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 to mental retardation: definition, classification, and systems of supports: Applications for clinicians, educators, disability program managers, and policy makers (10th ed.)* Washington, DC: American Association on Intellectual and Developmental Disabilities.

Schalock, R. L., & Luckasson, R. (2005). *Clinical judgment*. Washington, DC: American Association on Mental Retardation.

Scheerenberger, R. (1983). *A history of mental retardation: A quartar century of progress*. Baltimore: Bookes.

Sparrow, S. S., Cicchetti, D. V., & Balla, D. A. (2005). *Vineland-II: Vineland Adaptive Behavior Scales*. (2nd ed.). Minneapolis, MN: Pearson Assessments.

Spreat, S. (1999). Psychometric standards of adaptive behavior assessment. In R. L. Schalock (Ed.), *Adaptive behavior and its measurement: Implications for the field of mental retardation* (pp. 103–118). Washington, DC: American Association on Mental Retardation.

Stevens, K. B., & Price, J. R. (2006). Adaptive behavior, mental retardation, and the death penalty. *Journal of Forensic Psychology Practice*, *6*(3), 1–29.

Thompson, J. R., McGrew, K. S., & Bruininks, R. H. (1999). Adaptive and maladaptive behavior: Functional and structural characteristics. In R. L. Schalock (Ed.), *Adaptive behavior and its measurement: Implications for the field of mental retardation* (pp. 15–42). Washington, DC: American Association on Mental Retardation.

Voelker, S. L., Shore, D. L., Brown-More, C., Hill, L.C., Miller, L. T., & Perry, J. (1990). Validity of self-report of adaptive behavior skills by adults with mental retardation. *Mental Retardation*, *28*, 305–309.

Widaman, K. F., & McGrew, K. S. (1996). The structure of adaptive behavior. In J. W. Jacobson & J. A. Mulick (Eds.), *Manual of diagnosis and professional practice in mental retardation* (pp. 97–110). Washington, DC: American Psychological Association.

World Health Organization. (1992). *The international classification of diseases* (10th revision; ICD-10). Geneva, Switzerland: Author.

No. _____

IN THE

Supreme Court of the United States

TAVARES J. WRIGHT,

Petitioner,

v.

SECRETARY, DEPARTMENT OF CORRECTIONS, AND ATTORNEY GENERAL, STATE OF FLORIDA,

Respondents.

ON PETITION FOR A WRIT OF CERTIORARI TO THE UNITED STATES COURT OF APPEALS FOR THE ELEVENTH CIRCUIT

APPENDIX TO THE PETITION FOR A WRIT OF CERTIORARI

DEATH PENALTY CASE

APPENDIX P

Chart of States' Evidentiary Standards for Intellectual Disability

Table of States' Burdens of Proof on Intellectual Disability

State	Burden of Proof	Statute or Case
Alabama	Preponderance of the evidence.	*Smith v. State*, 112 So.3d 1108, 1125 (Ala. Crim. App. 2012); Ala. R. Crim. P. 32.3.
Arizona	Clear and convincing evidence (pretrial). Preponderance of the evidence (sentencing).	*State v. Escalante-Orozco,* 386 P.3d 798, 830-34 (Ariz. 2017); *State v. Grell*, 291 P.3d 350, 357-58 (Ariz. 2013); Ariz. Rev. Stat. Ann. § 13-753.
Arkansas	Preponderance of the evidence.	Ark. Code Ann. § 5-4-618 (2019).
California	Preponderance of the evidence.	Cal. Pen. Code. § 1376(B)(3).
Florida	Clear and convincing evidence.	Fla. Stat. § 921.137 (2013); *Wright v. State*, 256 So. 3d 766, 771 (Fla. 2018).
Georgia	Beyond a reasonable doubt.	Ga. Code Ann. § 17-7-131 (2017).
Idaho	Preponderance of the evidence.	Idaho Code § 19-2515A (2006).
Indiana	Preponderance of the evidence.	*Pruitt v. State*, 834 N.E.2d 90, 103 (Ind. 2005) (preponderance constitutionally required); Ind. Code § 35-36-9-4.
Kansas	None specified.	
Kentucky	Preponderance of the evidence.	*Woodall v. Commonwealth*, 563 S.W.3d 1, 6 n.29 (Ky. 2018); Ky. Rev. Stat. § 532.130.
Louisiana	Preponderance of the evidence.	La. Code Crim. Proc. art. 905.5.1 (2014).
Mississippi	Preponderance of the evidence.	*Chase v. State,* 873 So. 2d 1013, 1029 (Miss. 2004).
Missouri	Preponderance of the evidence.	Mo. Rev. Stat. § 565.030(4)(1) (2016).
Montana	None specified.	
Nebraska	Preponderance of the evidence.	Neb. Rev. Stat. § 28-105.01 (4) (2013).
Nevada	Preponderance of the evidence.	Nev. Rev. Stat. Ann. § 174.098.
North Carolina	Clear and convincing evidence (pretrial). Preponderance of the evidence (sentencing).	N.C. Gen. Stat. § 15A-2005 (2015).
Ohio	Preponderance of the evidence.	*State v. Ford*, 140 N.E. 616 655-56 (Ohio 2019).

Oklahoma	Clear and convincing evidence (pretrial). Preponderance of the evidence (sentencing).	Okla. Stat. tit. 21, § 701.10b (2019).
Oregon	Preponderance of the evidence.	*State v. Agee*, 364 P.3d 971, 983 (Or. 2015) (en banc).
Pennsylvania	Preponderance of the evidence.	*Commonwealth v. Sanchez*, 36 A.3d 24, 63 (Pa. 2011)
South Carolina	Preponderance of the evidence.	*State v. Laney*, 627 S.E.2d 726, 730 (S.C. 2006).
South Dakota	Preponderance of the evidence.	S.D. Codified Laws § 23A-27A-26.3 (2018).
Tennessee	Preponderance of the evidence.	Tenn. Code § 39-13-203 (2021).
Texas	Preponderance of the evidence.	*Ex parte* Van Alstyne, 239 S.W. 3d 815, 823 (Tex. Crim. App. (2007)).
Utah	Preponderance of the evidence.	Utah Code § 77-15a-104 (2018).
Wyoming	None specified.	

No. _____

IN THE

Supreme Court of the United States

TAVARES J. WRIGHT,

Petitioner,

v.

SECRETARY, DEPARTMENT OF CORRECTIONS, AND ATTORNEY GENERAL, STATE OF FLORIDA,

Respondents.

ON PETITION FOR A WRIT OF CERTIORARI TO THE UNITED STATES COURT OF APPEALS FOR THE ELEVENTH CIRCUIT

APPENDIX TO THE PETITION FOR A WRIT OF CERTIORARI

DEATH PENALTY CASE

APPENDIX Q

Excerpts from APA and AAIDD Publications

AMERICAN PSYCHIATRIC ASSOCIATION, DIAGNOSTIC AND STATISTICAL MANUAL OF MENTAL DISORDERS (5th ed. 2013)- Pages: 33-38.

AMERICAN ASSOCIATION ON INTELLECTUAL AND DEVELOPMENTAL DISABILITIES, INTELLECTUAL DISABILITY: DEFINITION, CLASSIFICATION, AND SYSTEMS OF SUPPORTS (11th ed. 2010)- Pages: 5, 31-38, 43-48, 151-153.

AMERICAN ASSOCIATION ON INTELLECTUAL AND DEVELOPMENTAL DISABILITIES, INTELLECTUAL DISABILITY: DEFINITION, CLASSIFICATION, AND SYSTEMS OF SUPPORTS, USER'S GUIDE (11th ed. 2012)- Pages: 22-24.

AMERICAN ASSOCIATION ON INTELLECTUAL AND DEVELOPMENTAL DISABILITIES, INTELLECTUAL DISABILITY: DEFINITION, CLASSIFICATION, AND SYSTEMS OF SUPPORTS (12th ed. 2021)- Pages: 13-15, 39-42.

AMERICAN ASSOCIATION ON INTELLECTUAL AND DEVELOPMENTAL DISABILITIES, THE DEATH PENALTY AND INTELLECTUAL DISABILITY (Edward A. Polloway ed., 2015)- Pages: 21-36, 155-169.

DIAGNOSTIC AND STATISTICAL MANUAL OF MENTAL DISORDERS

FIFTH EDITION

DSM-5™

American Psychiatric Publishing
A Division of American Psychiatric Association

Washington, DC
London, England

Manufactured in the United States of America on acid-free paper.

ISBN 978-0-89042-554-1 (Hardcover) 2nd printing June 2013

ISBN 978-0-89042-555-8 (Paperback) 2nd printing June 2013

American Psychiatric Association
1000 Wilson Boulevard
Arlington, VA 22209-3901
www.psych.org

The correct citation for this book is American Psychiatric Association: Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition. Arlington, VA, American Psychiatric Association, 2013.

Library of Congress Cataloging-in-Publication Data
Diagnostic and statistical manual of mental disorders : DSM-5. — 5th ed.
p. ; cm.
DSM-5
DSM-V
Includes index.
ISBN 978-0-89042-554-1 (hardcover : alk. paper) — ISBN 978-0-89042-555-8 (pbk. : alk. paper)
I. American Psychiatric Association. II. American Psychiatric Association. DSM-5 Task Force.
III. Title: DSM-5. IV. Title: DSM-V.
[DNLM: 1. Diagnostic and statistical manual of mental disorders. 5th ed. 2. Mental Disorders—classification. 3. Mental Disorders—diagnosis. WM 15]
RC455.2.C4
616.89'075—dc23

2013011061

British Library Cataloguing in Publication Data
A CIP record is available from the British Library.

Text Design—Tammy J. Cordova

Manufacturing—R.R. Donnelley

cians an opportunity to document factors that may have played a role in the etiology of the disorder, as well as those that might affect the clinical course. Examples include genetic disorders, such as fragile X syndrome, tuberous sclerosis, and Rett syndrome; medical conditions such as epilepsy; and environmental factors, including very low birth weight and fetal alcohol exposure (even in the absence of stigmata of fetal alcohol syndrome).

Intellectual Disabilities

Intellectual Disability (Intellectual Developmental Disorder)

Diagnostic Criteria

Intellectual disability (intellectual developmental disorder) is a disorder with onset during the developmental period that includes both intellectual and adaptive functioning deficits in conceptual, social, and practical domains. The following three criteria must be met:

A. Deficits in intellectual functions, such as reasoning, problem solving, planning, abstract thinking, judgment, academic learning, and learning from experience, confirmed by both clinical assessment and individualized, standardized intelligence testing.

B. Deficits in adaptive functioning that result in failure to meet developmental and sociocultural standards for personal independence and social responsibility. Without ongoing support, the adaptive deficits limit functioning in one or more activities of daily life, such as communication, social participation, and independent living, across multiple environments, such as home, school, work, and community.

C. Onset of intellectual and adaptive deficits during the developmental period.

Note: The diagnostic term *intellectual disability* is the equivalent term for the ICD-11 diagnosis of *intellectual developmental disorders.* Although the term *intellectual disability* is used throughout this manual, both terms are used in the title to clarify relationships with other classification systems. Moreover, a federal statute in the United States (Public Law 111-256, Rosa's Law) replaces the term *mental retardation* with *intellectual disability,* and research journals use the term *intellectual disability.* Thus, *intellectual disability* is the term in common use by medical, educational, and other professions and by the lay public and advocacy groups.

Specify current severity (see Table 1):

317 (F70) Mild
318.0 (F71) Moderate
318.1 (F72) Severe
318.2 (F73) Profound

Specifiers

The various levels of severity are defined on the basis of adaptive functioning, and not IQ scores, because it is adaptive functioning that determines the level of supports required. Moreover, IQ measures are less valid in the lower end of the IQ range.

TABLE 1 Severity levels for intellectual disability (intellectual developmental disorder)

Severity level	Conceptual domain	Social domain	Practical domain
Mild	For preschool children, there may be no obvious conceptual differences. For school-age children and adults, there are difficulties in learning academic skills involving reading, writing, arithmetic, time, or money, with support needed in one or more areas to meet age-related expectations. In adults, abstract thinking, executive function (i.e., planning, strategizing, priority setting, and cognitive flexibility), and short-term memory, as well as functional use of academic skills (e.g., reading, money management), are impaired. There is a somewhat concrete approach to problems and solutions compared with age-mates.	Compared with typically developing age-mates, the individual is immature in social interactions. For example, there may be difficulty in accurately perceiving peers' social cues. Communication, conversation, and language are more concrete or immature than expected for age. There may be difficulties regulating emotion and behavior in age-appropriate fashion; these difficulties are noticed by peers in social situations. There is limited understanding of risk in social situations; social judgment is immature for age, and the person is at risk of being manipulated by others (gullibility).	The individual may function age-appropriately in personal care. Individuals need some support with complex daily living tasks in comparison to peers. In adulthood, supports typically involve grocery shopping, transportation, home and child-care organizing, nutritious food preparation, and banking and money management. Recreational skills resemble those of age-mates, although judgment related to well-being and organization around recreation requires support. In adulthood, competitive employment is often seen in jobs that do not emphasize conceptual skills. Individuals generally need support to make health care decisions and legal decisions, and to learn to perform a skilled vocation competently. Support is typically needed to raise a family.

TABLE 1 Severity levels for intellectual disability (intellectual developmental disorder) *(continued)*

Severity level	Conceptual domain	Social domain	Practical domain
Moderate	All through development, the individual's conceptual skills lag markedly behind those of peers. For preschoolers, language and pre-academic skills develop slowly. For school-age children, progress in reading, writing, mathematics, and understanding of time and money occurs slowly across the school years and is markedly limited compared with that of peers. For adults, academic skill development is typically at an elementary level, and support is required for all use of academic skills in work and personal life. Ongoing assistance on a daily basis is needed to complete conceptual tasks of day-to-day life, and others may take over these responsibilities fully for the individual.	The individual shows marked differences from peers in social and communicative behavior across development. Spoken language is typically a primary tool for social communication but is much less complex than that of peers. Capacity for relationships is evident in ties to family and friends, and the individual may have successful friendships across life and sometimes romantic relations in adulthood. However, individuals may not perceive or interpret social cues accurately. Social judgment and decision-making abilities are limited, and caretakers must assist the person with life decisions. Friendships with typically developing peers are often affected by communication or social limitations. Significant social and communicative support is needed in work settings for success.	The individual can care for personal needs involving eating, dressing, elimination, and hygiene as an adult, although an extended period of teaching and time is needed for the individual to become independent in these areas, and reminders may be needed. Similarly, participation in all household tasks can be achieved by adulthood, although an extended period of teaching is needed, and ongoing supports will typically occur for adult-level performance. Independent employment in jobs that require limited conceptual and communication skills can be achieved, but considerable support from co-workers, supervisors, and others is needed to manage social expectations, job complexities, and ancillary responsibilities such as scheduling, transportation, health benefits, and money management. A variety of recreational skills can be developed. These typically require additional supports and learning opportunities over an extended period of time. Maladaptive behavior is present in a significant minority and causes social problems.

TABLE 1 Severity levels for intellectual disability (intellectual developmental disorder) *(continued)*

Severity level	Conceptual domain	Social domain	Practical domain
Severe	Attainment of conceptual skills is limited. The individual generally has little understanding of written language or of concepts involving numbers, quantity, time, and money. Caretakers provide extensive supports for problem solving throughout life.	Spoken language is quite limited in terms of vocabulary and grammar. Speech may be single words or phrases and may be supplemented through augmentative means. Speech and communication are focused on the here and now within everyday events. Language is used for social communication more than for explication. Individuals understand simple speech and gestural communication. Relationships with family members and familiar others are a source of pleasure and help.	The individual requires support for all activities of daily living, including meals, dressing, bathing, and elimination. The individual requires supervision at all times. The individual cannot make responsible decisions regarding well-being of self or others. In adulthood, participation in tasks at home, recreation, and work requires ongoing support and assistance. Skill acquisition in all domains involves long-term teaching and ongoing support. Maladaptive behavior, including self-injury, is present in a significant minority.
Profound	Conceptual skills generally involve the physical world rather than symbolic processes. The individual may use objects in goal-directed fashion for self-care, work, and recreation. Certain visuospatial skills, such as matching and sorting based on physical characteristics, may be acquired. However, co-occurring motor and sensory impairments may prevent functional use of objects.	The individual has very limited understanding of symbolic communication in speech or gesture. He or she may understand some simple instructions or gestures. The individual expresses his or her own desires and emotions largely through nonverbal, nonsymbolic communication. The individual enjoys relationships with well-known family members, caretakers, and familiar others, and initiates and responds to social interactions through gestural and emotional cues. Co-occurring sensory and physical impairments may prevent many social activities.	The individual is dependent on others for all aspects of daily physical care, health, and safety, although he or she may be able to participate in some of these activities as well. Individuals without severe physical impairments may assist with some daily work tasks at home, like carrying dishes to the table. Simple actions with objects may be the basis of participation in some vocational activities with high levels of ongoing support. Recreational activities may involve, for example, enjoyment in listening to music, watching movies, going out for walks, or participating in water activities, all with the support of others. Co-occurring physical and sensory impairments are frequent barriers to participation (beyond watching) in home, recreational, and vocational activities. Maladaptive behavior is present in a significant minority.

barriers to participation (beyond watching) in home, recreational, and vocational activities. Maladaptive behavior is present in a significant minority.

vent many social activities.

prevent functional use of objects.

Diagnostic Features

The essential features of intellectual disability (intellectual developmental disorder) are deficits in general mental abilities (Criterion A) and impairment in everyday adaptive functioning, in comparison to an individual's age-, gender-, and socioculturally matched peers (Criterion B). Onset is during the developmental period (Criterion C). The diagnosis of intellectual disability is based on both clinical assessment and standardized testing of intellectual and adaptive functions.

Criterion A refers to intellectual functions that involve reasoning, problem solving, planning, abstract thinking, judgment, learning from instruction and experience, and practical understanding. Critical components include verbal comprehension, working memory, perceptual reasoning, quantitative reasoning, abstract thought, and cognitive efficacy. Intellectual functioning is typically measured with individually administered and psychometrically valid, comprehensive, culturally appropriate, psychometrically sound tests of intelligence. Individuals with intellectual disability have scores of approximately two standard deviations or more below the population mean, including a margin for measurement error (generally +5 points). On tests with a standard deviation of 15 and a mean of 100, this involves a score of 65–75 (70 ± 5). Clinical training and judgment are required to interpret test results and assess intellectual performance.

Factors that may affect test scores include practice effects and the "Flynn effect' (i.e., overly high scores due to out-of-date test norms). Invalid scores may result from the use of brief intelligence screening tests or group tests; highly discrepant individual subtest scores may make an overall IQ score invalid. Instruments must be normed for the individual's sociocultural background and native language. Co-occurring disorders that affect communication, language, and/or motor or sensory function may affect test scores. Individual cognitive profiles based on neuropsychological testing are more useful for understanding intellectual abilities than a single IQ score. Such testing may identify areas of relative strengths and weaknesses, an assessment important for academic and vocational planning.

IQ test scores are approximations of conceptual functioning but may be insufficient to assess reasoning in real-life situations and mastery of practical tasks. For example, a person with an IQ score above 70 may have such severe adaptive behavior problems in social judgment, social understanding, and other areas of adaptive functioning that the person's actual functioning is comparable to that of individuals with a lower IQ score. Thus, clinical judgment is needed in interpreting the results of IQ tests.

Deficits in adaptive functioning (Criterion B) refer to how well a person meets community standards of personal independence and social responsibility, in comparison to others of similar age and sociocultural background. Adaptive functioning involves adaptive reasoning in three domains: conceptual, social, and practical. The *conceptual (academic) domain* involves competence in memory, language, reading, writing, math reasoning, acquisition of practical knowledge, problem solving, and judgment in novel situations, among others. The *social domain* involves awareness of others' thoughts, feelings, and experiences; empathy; interpersonal communication skills; friendship abilities; and social judgment, among others. The *practical domain* involves learning and self-management across life settings, including personal care, job responsibilities, money management, recreation, self-management of behavior, and school and work task organization, among others. Intellectual capacity, education, motivation, socialization, personality features, vocational opportunity, cultural experience, and coexisting general medical conditions or mental disorders influence adaptive functioning.

Adaptive functioning is assessed using both clinical evaluation and individualized, culturally appropriate, psychometrically sound measures. Standardized measures are used with knowledgeable informants (e.g., parent or other family member; teacher; counselor; care provider) and the individual to the extent possible. Additional sources of information include educational, developmental, medical, and mental health evaluations. Scores from standardized measures and interview sources must be interpreted using clinical judgment. When standardized testing is difficult or impossible, because of a variety of

factors (e.g., sensory impairment, severe problem behavior), the individual may be diagnosed with unspecified intellectual disability. Adaptive functioning may be difficult to assess in a controlled setting (e.g., prisons, detention centers); if possible, corroborative information reflecting functioning outside those settings should be obtained.

Criterion B is met when at least one domain of adaptive functioning—conceptual, social, or practical—is sufficiently impaired that ongoing support is needed in order for the person to perform adequately in one or more life settings at school, at work, at home, or in the community. To meet diagnostic criteria for intellectual disability, the deficits in adaptive functioning must be directly related to the intellectual impairments described in Criterion A. Criterion C, onset during the developmental period, refers to recognition that intellectual and adaptive deficits are present during childhood or adolescence.

Associated Features Supporting Diagnosis

Intellectual disability is a heterogeneous condition with multiple causes. There may be associated difficulties with social judgment; assessment of risk; self-management of behavior, emotions, or interpersonal relationships; or motivation in school or work environments. Lack of communication skills may predispose to disruptive and aggressive behaviors. Gullibility is often a feature, involving naiveté in social situations and a tendency for being easily led by others. Gullibility and lack of awareness of risk may result in exploitation by others and possible victimization, fraud, unintentional criminal involvement, false confessions, and risk for physical and sexual abuse. These associated features can be important in criminal cases, including Atkins-type hearings involving the death penalty.

Individuals with a diagnosis of intellectual disability with co-occurring mental disorders are at risk for suicide. They think about suicide, make suicide attempts, and may die from them. Thus, screening for suicidal thoughts is essential in the assessment process. Because of a lack of awareness of risk and danger, accidental injury rates may be increased.

Prevalence

Intellectual disability has an overall general population prevalence of approximately 1%, and prevalence rates vary by age. Prevalence for severe intellectual disability is approximately 6 per 1,000.

Development and Course

Onset of intellectual disability is in the developmental period. The age and characteristic features at onset depend on the etiology and severity of brain dysfunction. Delayed motor, language, and social milestones may be identifiable within the first 2 years of life among those with more severe intellectual disability, while mild levels may not be identifiable until school age when difficulty with academic learning becomes apparent. All criteria (including Criterion C) must be fulfilled by history or current presentation. Some children under age 5 years whose presentation will eventually meet criteria for intellectual disability have deficits that meet criteria for global developmental delay.

When intellectual disability is associated with a genetic syndrome, there may be a characteristic physical appearance (as in, e.g., Down syndrome). Some syndromes have a *behavioral phenotype*, which refers to specific behaviors that are characteristic of particular genetic disorder (e.g., Lesch-Nyhan syndrome). In acquired forms, the onset may be abrupt following an illness such as meningitis or encephalitis or head trauma occurring during the developmental period. When intellectual disability results from a loss of previously acquired cognitive skills, as in severe traumatic brain injury, the diagnoses of intellectual disability and of a neurocognitive disorder may both be assigned.

Although intellectual disability is generally nonprogressive, in certain genetic disorders (e.g., Rett syndrome) there are periods of worsening, followed by stabilization, and in

Intellectual Disability

Definition, Classification, and Systems of Supports

The AAIDD Ad Hoc Committee on Terminology and Classification

11th Edition

aaidd
American Association on Intellectual and Developmental Disabilities

Published by
American Association on Intellectual and Developmental Disabilities
501 3rd Street, NW, Suite 200
Washington, DC 20001-2760

Printed in the United States of America

Library of Congress Cataloging-in-Publication Data

Intellectual disability : definition, classification, and systems of supports / The AAIDD Ad Hoc Committee on Terminology and Classification.—11th ed.
p. cm.
Includes bibliographical references and index.
ISBN 978-1-935304-04-3 (alk. paper)
1. Mental retardation—Classification. I. American Association on Intellectual and Developmental Disabilities.
RC570.C515 2010
616.85'88—dc22 2009040030

CHAPTER 1

DEFINITION OF INTELLECTUAL DISABILITY

Intellectual disability is characterized by significant limitations both in intellectual functioning and in adaptive behavior as expressed in conceptual, social, and practical adaptive skills. This disability originates before age 18.

OVERVIEW

Defining refers to precisely explaining the term and establishing the meaning and boundaries of the term. Significant consequences can result from the way a term is defined. As discussed by Gross and Hahn (2004), Luckasson and Reeve (2001), and Stowe, Turnbull, and Sublet (2006), a definition can make someone eligible or ineligible for services, subjected to something or not subjected to it (e.g., involuntary commitment), exempted from something or not exempted (e.g., from the death penalty), included or not included (as to protections against discrimination and equal opportunity), and/or entitled or not entitled (e.g., as to Social Security benefits or other financial benefits). Our purpose in this chapter is to review briefly the historical approaches to defining *intellectual disability* (ID), present the current definition of ID and the assumptions that are essential to the application of the definition, discuss the historical consistency in regard to the three criteria used to operationally define the construct, and summarize how the boundaries of the construct have been operationalized over the past 50 years.

HISTORICAL APPROACHES TO DEFINING INTELLECTUAL DISABILITY

Historically, four broad approaches (i.e., social, clinical, intellectual, and dual-criterion) have been used to define the construct now referred to as ID. Remnants of these four approaches are still evident in current discussions regarding who is (or should be) diagnosed as an individual with an ID (see, for example, Switzky & Greenspan, 2006a, 2006b).

Social Approach

Historically, persons were defined or identified as having ID because they failed to adapt socially to their environment. Because an emphasis on intelligence and the role of intelligent people in society was to come later, the oldest historical definitional approach

CHAPTER 4

INTELLECTUAL FUNCTIONING AND ITS ASSESSMENT

For purposes of diagnosis, intellectual functioning is currently best conceptualized and captured by a general factor of intelligence. Intelligence is a general mental ability. It includes reasoning, planning, solving problems, thinking abstractly, comprehending complex ideas, learning quickly, and learning from experience. The "significant limitations in intellectual functioning" criterion for a diagnosis of intellectual disability is an IQ score that is approximately two standard deviations below the mean, considering the standard error of measurement for the specific instruments used and the instruments' strengths and limitations.

OVERVIEW

The multidimensional model of human functioning presented in Figure 2.1 includes intellectual abilities as one of the five dimensions of human functioning. Intellectual functioning, which is a broader term than either intellectual abilities or intelligence, reflects the fact that what is considered intelligent behavior is dependent upon other dimensions of human functioning: the adaptive behavior that one exhibits, the person's mental and physical health status, the opportunity to participate in major life activities, and the context within which people live their everyday lives. Thus, as discussed throughout this chapter, commonly used measures/indices of intelligence need to be interpreted within a broader context than a single IQ score.

Although the primary focus in this chapter is on intelligence and its assessment, it is important that readers of this manual should note the following implications of intelligence on the multidimensionality of ID:

- Limitations in intelligence should be considered in light of four other dimensions of human functioning: adaptive behavior, health, participation, and context.
- The measurement of intelligence may have different relevance, depending on whether it is considered for purposes of diagnosis or classification.
- Although far from perfect, intellectual functioning is currently best represented by IQ scores when they are obtained from appropriate, standardized and individually administered assessment instruments.

The assessment of intellectual functioning is essential to making a diagnosis of ID, as virtually all historical definitions of ID (formerly mental retardation) make reference to significantly subaverage intellectual functioning as one of the diagnostic criteria. Our three purposes in this chapter are to present discussions of (a) the definition and nature of intelligence, (b) the operational definition of significant limitations in intellectual functioning, and (c) challenging issues and related guidelines regarding the measurement of intelligence and the interpretation of IQ scores.

DEFINITION AND NATURE OF INTELLIGENCE

Individuals vary in their ability to understand complexities and reason, adapt to the environment, and use thought to solve problems (Neisser et al., 1996). Although reasoning, adaptation, comprehension, and thinking are somewhat descriptive of intelligence, the construct itself has successfully eluded a definition that is acceptable to everyone. Over the past century, three broad conceptual frameworks have been used in an attempt to better define the construct of intelligence: intelligence as a single (i.e., unifactorial) trait; intelligence as a multitrait, hierarchical phenomenon; or intelligence as a multidimensional construct.

Intelligence as a Single Trait

Because so many of the available measures of cognitive ability were highly correlated, Spearman (1927) concluded that the relationship among these various cognitive ability measures could be described as a single factor of general intelligence (i.e., *g*). Most of the more commonly used individual tests of intelligence, such as the Wechsler family of scales and the Stanford-Binet Intelligence Scale, 4th edition (SBIS-4; Thorndike, Hagen, & Sattler, 1986a), provide metrics of this *g* factor. Although Thurstone (1938) was initially unable to replicate the results of Spearman's work, he later acknowledged that there was an error in his factor analytic calculations. When this miscalculation was corrected, Thurstone also obtained Spearman's general factor of intelligence (see Carroll, 1997). In general, this general factor framework is currently the most widely accepted conceptualization of intelligence (Gottfredson, 1997).

Multitrait Hierarchical Phenomenon

Some theorists conceptualize intelligence as a hierarchical structure, with *g* at the apex, supported by various more specialized cognitive abilities. Carroll (1993) reviewed hundreds of intelligence test factor analysis studies published between the 1920s and the 1990s. His analysis yielded a three strata hierarchical model, with the *g* factor at the apex of a pyramidal structure. In Carroll's model, there were approximately 60 discrete narrow abilities at the base of the pyramid. These narrow cognitive abilities were highly correlated and were further factor analyzed into the 10 broader abilities that formed the

second stratum of the hierarchy. Finally, these 10 broader abilities were submitted to factor analysis, which yielded a single factor of *g*.

Multiple Intelligences

Critics (e.g., Ceci, 1990; H. Gardner, 1983; Gould, 1978) of the above two conceptual frameworks noted that the reliance on a single metric of intelligence ignores a number of important areas of mental ability. Gardner argued that most tests of intelligence assess only linguistics, logic, and some aspects of spatial intelligence; other forms and types of intelligence are largely ignored. He went on to note that the paper and pencil format of the typical intelligence test further narrows the focus of intelligence testing to those things that lend themselves to paper and pencil testing.

Recent theories of multiple intelligence have proposed anywhere from two to eight types of intelligence (see Cattell, 1963; Das, Naglieri, & Kirby, 1994; H. Gardner, 1983; Greenspan, 1981). A brief summary of the main theories of multiple intelligences follows.

Cattell (1963) and Horn and Cattell (1966) identified two main factors explaining intellectual ability: *crystallized intelligence* (*gc*) and *fluid intelligence* (*gf*). Crystallized intelligence was defined as those more global activities, such as knowledge and information, that were gained by the individual through life experiences and education. Fluid intelligence was explained in reference to abilities in reasoning and memory. Furthermore, Cattell defined *gc* as a stable trait, whereas *gf* may, in fact, decrease with age.

H. Gardner (1983, 1993) posited a theoretical model of multiple intelligences. Initially, his model consisted of seven different intelligences, each tapping distinctive problem-solving and information-processing capabilities and each with its own distinctive developmental trajectory. The original seven intelligences in Gardner's model were linguistic, logical-mathematical, spatial, musical, bodily kinesthetic, interpersonal, and intrapersonal. In 1998, H. Gardner added an eighth independent ability, naturalistic intelligence, to his model. Of Gardner's eight types of intelligence, he claimed that only three (linguistic, logical-mathematical, and spatial) are assessed by contemporary intelligence tests. Gardner (see Chen & Gardner, 1997) advocated for the use of nonstandardized means of assessing the multiple intelligences; he viewed the process as an ongoing one in which personalized assessments in a variety of contexts should be used. The significant criticism remains valid and pertinent that Gardner's multiple intelligences model lacks an empirical base and psychometric validation.

Das et al. (1994) and Naglieri and Das (1997) proposed a four factor model of cognitive processes that underlie intelligence: planning, attention, simultaneous processing, and successive processing. Referred to as the *PASS model*, its origins may be found in the early work of the Russian neurologist Luria. The *planning* process includes self-regulation, analysis and evaluation of situations, and the use of knowledge to solve problems. The *attentional* process involves the regulation of activity, focusing on specific stimuli while inhibiting responses to other less relevant stimuli. *Simultaneous* processing involves the understanding of groupings of stimuli or the identification of commonalities of a

grouping of stimuli. *Successive* processing involves the process of grouping a number of stimuli into a linear series that makes sense.

Sternberg (1988) and Sternberg and Detterman (1986) proposed a three factor model of intelligence that they called the triarchic theory of human intelligence. According to Sternberg (1988), the three fundamental aspects of intelligence are *analytical, creative,* and *practical.* Analytic abilities involve the capacity to analyze and be critical of ideas. Creativity is defined as a person's ability to generate novel ideas that offer a significant contribution, and practical intelligence is an individual's ability to convert ideas into practical application and to convince others of their utility. This sort of distinction between academic and practical intelligence has been offered by a number of theorists (cf. Neisser, 1976). Sternberg has also faced the challenge of developing a metric with which to assess each of his proposed aspects of intelligence; to date no such instrument exists.

Greenspan's (1981) model of multiple intelligences, which has some overlap with Sternberg's triarchic model as well as the current definition of adaptive behavior presented in chapter 5 of this manual, has evolved over time. The tripartite model of intelligence proposed by Greenspan and his colleagues (Greenspan, 1997, 2006b; Greenspan & Love, 1997; Greenspan, Switzky, & Granfield, 1996) defined *intelligence* as being composed of conceptual, practical, and social intelligence. *Conceptual intelligence* is essentially equivalent to the single factor of *g,* although Greenspan (1996, 1997) vehemently opposed the position of using only *g* or a unitary IQ score as representing an individual's intellectual abilities. *Practical intelligence* involves the performance of everyday skills that are typically measured by adaptive behavior scales, with *social intelligence* being defined as an individual's social and interpersonal abilities (e.g., moral judgment, empathy, social skills). Gullibility and credulity have been added as critical elements of social intelligence (Greenspan & Granfield, 1992; Greenspan, Loughlin, & Black, 2001).

In summary, many of the aforementioned theories of multiple intelligences have not been validated via standardized and quantifiable measures. H. Gardner's multiple intelligences, with the exception of some useful application in educational settings, continues to remain theoretical. Sternberg failed in his attempts to develop a measure capable of reliably measuring his triarchical model of intelligence. The Greenspan and Sternberg models face the common challenge of operationalizing tasks to quantify the constructs of their tripartite models, particularly in the area of social intelligence.

A single dimension of intelligence continues to garner the most support within the scientific community (Carroll, 1997; Gottfredson, 1997; Hernstein & Murray, 1994). Thus, until such measures of multiple intelligences can be assessed reliably and validly, it is the position of AAIDD that intellectual functioning (as defined at the beginning of this chapter) is best conceptualized and captured by a general factor of intelligence (*g*).

SIGNIFICANT LIMITATIONS IN INTELLECTUAL FUNCTIONING: OPERATIONAL DEFINITION

In this *Manual,* and consistent with the 2002 *Manual* (Luckasson et al., 2002), the intellectual functioning criterion for a diagnosis of ID is approximately two standard deviations below the mean, considering the standard error of measurement for the specific assessment instruments used and the strengths and limitations of the instruments. In reference to this operational definition of significant limitations, consider the following guidance:

- The intent of this definition is not to specify a hard and fast cutoff point/score for meeting the significant limitations in intellectual functioning criterion of ID. Rather, one needs to use clinical judgment in interpreting the obtained score in reference to the test's standard error of measurement, the assessment instrument's strengths and limitations, and other factors such as practice effects, fatigue effects, and age of norms used (see following section). In addition, significant limitations in intellectual functioning is only one of the three criteria used to establish a diagnosis of ID.
- The use of "approximately" reflects the role of clinical judgment in weighing the factors that contribute to the validity and precision of a decision. The term also addresses statistical error and uncertainty inherent in any assessment of human behavior. In that regard, the decision-making process cannot be viewed as only a statistical calculation.

CHALLENGING ISSUES AND RELATED GUIDELINES REGARDING THE MEASUREMENT OF INTELLIGENCE AND THE INTERPRETATION OF IQ SCORES

Just as defining intelligence has proven to be a challenging task, measuring or quantifying intelligence is equally difficult. It is important to note that IQ scores derived from an intelligence test are now developed on the basis of a deviation (from the mean) score and not on the older conception of mental age. Thus, in reference to the significant limitations in intellectual functioning criterion for a diagnosis of ID, a valid diagnosis of ID is based on how far the person's score deviates from the mean on the respective standardized assessment instrument and *not* on the ratio of mental age to chronological age.

There are a number of challenges and psychometric issues related to the measurement of intelligence and the interpretation of IQ scores. Although one potentially can take comfort from the fact that intelligence tests generally have good reliability and have demonstrated validity for some purposes, the typical intelligence test is not without psychometric challenges. In that regard, in this section of the chapter, we discuss 10 challenges and related guidelines regarding the measurement of intelligence and the interpretation of IQ scores: measurement error, test fairness, the Flynn Effect, comparability of scores from different tests, practice effect, the utility of scores at the extreme ends of a distribution,

determining a cutoff score, evaluating the role that an IQ score plays in making a diagnosis, assessor credentials, and test selection.

Measurement Error

The results of any psychometric assessment must be evaluated in terms of the accuracy of the instrument used and such is the case with the assessment of intelligence. An IQ score is subject to variability as a function of a number of potential sources of error, including variations in test performance, examiner's behavior, cooperation of test taker, and other personal and environmental factors. Thus, variation in scores may or may not represent the individual's actual or true level of intellectual functioning. The term *standard error of measurement*, which varies by test, subgroup, and age group, is used to quantify this variability and provide a stated statistical confidence interval within which the person's true score falls.

For well-standardized measures of general intellectual functioning, the standard error of measurement is approximately 3 to 5 points. As reported in the respective test's standardization manual, the test's standard error of measurement can be used to establish a statistical confidence interval around the obtained score. From the properties of the normal curve, a range of confidence can be established with parameters of at least one standard error of measurement (i.e., scores of about 66 to 74, 66% probability) or parameters of two standard error of measurement (i.e., scores of about 62 to 78, 95% probability).

Understanding and addressing the test's standard error of measurement is a critical consideration that must be part of any decision concerning a diagnosis of ID that is based, in part, on significant limitations in intellectual functioning. Both AAIDD and the American Psychiatric Association (2000) support the best practice of reporting an IQ score with an associated confidence interval. Both systems rely on the reported standard error of measurement that is derived from the standard deviation of the test and a measure of the test's reliability. Currently, the prevailing best practice standard in test construction, reporting, and interpretation is to use internal consistency measures of reliability (along with the test's standard deviation) to estimate a standard error of measurement. Reporting an IQ score with an associated confidence interval is a critical consideration underlying the appropriate use of intelligence tests and best practices; such reporting must be a part of any decision concerning the diagnosis of ID.

Test Fairness

There are at least two areas in which test fairness may be of particular concern. The first is when tests requiring a verbal response are employed with individuals who have severely limited verbal abilities. In these situations, the test score may underestimate their level of intellectual functioning. The second area involves individuals of diverse ethnicity or culture, who may achieve markedly different results. Readers are referred to chapters 3 and 8 for a discussion of guidelines regarding test selection and test fairness.

The Flynn Effect

Flynn's research (1984, 1987, 2006, 2007) as well as that of others (e.g., Kanaya, Scullin, & Ceci, 2003; Scullin, 2006) found that IQ scores have been increasing from one generation to the next in the United States as well as in all other developed countries for which IQ data are available. This increase in IQ scores over time was called the *Flynn Effect* by Hernstein and Murray (1994). The Flynn Effect refers to the observation (Flynn, 1984) that every restandardization sample for a major intelligence test (e.g., SBIS-4 and Wechsler) from 1932 through 1978 resulted in a mean IQ that tended to increase over time. Flynn (1987) reported that this effect was also observed in samples from other countries. Although the cause of this effect is unknown, Neisser et al. (1996) suggested that potential factors might well be improved nutrition, cultural changes, testing experience, changes in schooling, and changes in child-rearing practices.

The Flynn Effect raises potential challenges for the diagnosis of ID (Kanaya et al., 2003). Because Flynn (1984) reported that mean IQ increases about 0.33 points per year, some investigators (e.g., Flynn, 2006) have suggested that any obtained IQ score should be adjusted 0.33 points for each year the test was administered after the standardization was completed. For example, if the Wechsler Adult Intelligence Scale (WAIS-III; 1997) was used to assess an individual's IQ in July 2005, the population mean on the WAIS-III was set at 100 when it was originally normed in 1995 (published in 1997). However, based on Flynn's data, the population mean on the Full-Scale IQ raises roughly 0.33 points per year; thus the population mean on the WAIS-III Full-Scale IQ corrected for the Flynn Effect would be 103 in 2005 (9 years × 0.33 = 2.9). Hence, using the significant limitations of approximately two standard deviations below the mean, the Full-Scale IQ cutoff would be approximately 73 (plus or minus the standard error of measurement).

There are also data suggesting that the Flynn Effect may not be a purely linear function of time and that the impact of the effect may asymptote or even reverse. Teasdale and Owens (2005), for example, reported on a large sample of Danish males in which the Flynn Effect peaked and subsequently reversed. In a Norwegian sample, Sundet, Barlaug, and Torjussen (2004) reported a slowing and eventual cessation of the Flynn Effect over time. These data would seem to suggest that while the Flynn Effect is evident, how one corrects for it is still a challenging issue.

As discussed in the *User's Guide* (Schalock et al., 2007) that accompanies the 10th edition of this *Manual*, best practices require recognition of a potential Flynn Effect when older editions of an intelligence test (with corresponding older norms) are used in the assessment or interpretation of an IQ score. As suggested in the *User's Guide* (Schalock et al., 2007, pp. 20, 21):

> 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 of tests 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.

Comparability of Scores From Different Tests

Not all scores obtained on intelligence tests given to the same person will be identical. Specifically, IQ scores are not expected to be the same across tests, editions of the same test, or time periods (Evans, 1991). A number of studies have revealed significantly different results from appropriately selected tests. For example, Quereshi and Seitz (1994) reported that the Wechsler Preschool and Primary Scale of Intelligence-Revised (WPPSI), Wechsler Intelligence Scale for Children-Revised (WISC-R), and the Wechsler Preschool and Primary Scale of Intelligence-Revised (WPPSI-R) did not yield the same results when used on young children. Highest IQ scores were obtained on the WPPSI and lowest on the WPPSI-R. The SBIS-4 yielded significantly higher scores (by over 14 points) than did the WISC-R for students with lower IQ scores but yielded significantly lower scores for students with higher IQ scores. The two tests yielded similar scores for students with IQ scores between 70 and 90 (Prewett & Matavich, 1992). Scores on the WISC-III were significantly correlated with scores on the SBIS-4 with a population of students with mild mental retardation, but the average IQ on the WISC-III was 8 points lower (Lukens & Hurrell, 1996). Nelson and Dacey (1999) reported that in a sample of adults who had mild to moderate mental retardation an SBIS will yield a significantly lower score than a Wechsler test. Their results were consistent with earlier work published in the Stanford Binet Technical Manual (Thorndike, Hagen, & Sattler, 1986b).

Users of this *Manual* need to be aware of—and sensitive to—potential differences in scores obtained from two different tests. Sources of variation can result from (a) group versus individually administered tests; (b) the purposes for which the test was administered (e.g., administered initially to measure academic achievement but later used to derive an IQ score); (c) the properties of the test (e.g., using two tests with very disparate standard errors of measurement); (d) nonstandardized administration of the assessment instrument(s); (e) test content across different scales and between different age levels on the same scale; (f) scores obtained on verbal versus nonverbal tests; (g) differences in the standardization samples; (h) changes between different editions of the same scale/test; (i) use of an alternative scale as an individual's chronological age increases; and/or (j) variations in the person's abilities or performance.

Practice Effect

The *practice effect* refers to gains in IQ scores on tests of intelligence that result from a person being retested on the same instrument. Kaufman (1994) noted that practice effect can occur when the same individual is retested on a similar instrument. For example, the WAIS-III *Manual* presents data showing the artificial increase in IQ scores when the same instrument is readministered within a short time interval. The WAIS-III *Manual* also reports the average increase between administrations with intervals of 2 to 12 weeks (Wechsler, 1997). For this reason, established clinical practice is to avoid administering the same intelligence test within the same year to the same individual because it will often lead to an overestimate of the examinee's true intelligence.

CHAPTER 5

ADAPTIVE BEHAVIOR AND ITS ASSESSMENT

Adaptive behavior is the collection of conceptual, social, and practical skills that have been learned and are performed by people in their everyday lives.

For the diagnosis of intellectual disability, significant limitations in adaptive behavior should be established through the use of standardized measures normed on the general population, including people with disabilities and people without disabilities. On these standardized measures, significant limitations in adaptive behavior are operationally defined as performance that is approximately two standard deviations below the mean of either (a) one of the following three types of adaptive behavior: conceptual, social, or practical or (b) an overall score on a standardized measure of conceptual, social, and practical skills. The assessment instrument's standard error of measurement must be considered when interpreting the individual's obtained scores.

OVERVIEW

The inclusion of the concept of adaptive behavior in the diagnosis of persons with *intellectual disability* (ID) has a long history. Nihira (1999), for example, cited early leaders, such as Itard, Seguin, Voison, and Howe, who referred to signs of ID that included the absence of social competency, a need for skill training, an inability to meet social norms, and difficulty with fending for one's self. Although adaptive behavior did not play a formal role in the diagnosis of ID during the first half of the 20th century, the construct's importance to understanding ID was not completely abandoned. Doll, for example, introduced the Vineland Social Maturity Scale in 1936, an instrument that included 117 items focused on practical skills used in everyday situations.

When the intelligence test, resulting in an IQ score, was introduced in the early 1900s, it was embraced as an efficient and objective means to distinguish individuals with ID from the general population (Scheerenberger, 1983). The intelligence test not only produced a highly reliable score, but because it was normed on the general population, it yielded an unambiguous indicator of how much a person deviated from others. However, dissatisfaction with the IQ score as the sole indicator of ID emerged over time. Among the greatest concerns about intelligence testing was that IQ scores only provided a narrow measure of intellectual functioning related to academic tasks (i.e., linguistic, conceptual,

and mathematical abilities and skills), thus ignoring important aspects of intellectual functioning that included social and practical skills. Also, the perception that IQ scores contributed to misdiagnosing children from poor and minority backgrounds shook people's confidence in using the IQ as the sole diagnostic measure (Reschly, Myers, & Hartel, 2002; Scheerenberger, 1983).

As a result of this dissatisfaction, adaptive behavior reemerged in 1959 as one of the three criteria used to diagnose ID. According to Heber in the AAIDD 1959 *Manual on Terminology and Classification*, "measured intelligence cannot be used as the sole criteria of mental retardation [the term in use then] since intelligence test performances do not always correspond to level of deficiency in total adaptation" (pp. 55–56). *Adaptive behavior* was defined by Heber (1959) as

> the effectiveness with which the individual copes with the nature and social demands of his environment. It has two major facets: the degree to which the individual is able to function and maintain himself independently, and the degree to which he meets satisfactorily the culturally-imposed demands of personal and social responsibility. (p. 61)

Grossman (1973, 1983) reaffirmed the importance of adaptive behavior in the diagnosis of ID. Grossman's (1983) definition of adaptive behavior was "the effectiveness or degree with which individuals meet the standards of personal independence and social responsibility expected for his age and cultural group" (p. 1). The importance of adaptive behavior in the diagnosis of ID has been reaffirmed in each of the successive AAIDD *Terminology and Classification Manuals* (Luckasson et al., 1992, 2002).

Both Heber and Grossman recognized the multidimensionality of adaptive behavior and the influence of culture on the assessment of the construct. Heber conceptualized adaptive behavior as consisting of three primary factors: maturation, learning, and social adjustment. These three domains continue to be part of the most current conceptualization of adaptive behavior but are reframed as practical, conceptual, and social skills.

The consensus, based on considerable published research on the factor structure of adaptive behavior (e.g., Harrison & Oakland, 2003; McGrew, Bruininks, & Johnson, 1996; Thompson, McGrew, & Bruininks, 1999), is that adaptive behavior is multidimensional and includes the following:

- *Conceptual skills*: language; reading and writing; and money, time, and number concepts
- *Social skills*: interpersonal skills, social responsibility, self-esteem, gullibility, naïveté (i.e., wariness), follows rules/obeys laws, avoids being victimized, and social problem solving
- *Practical skills*: activities of daily living (personal care), occupational skills, use of money, safety, health care, travel/transportation, schedules/routines, and use of the telephone

In this chapter we discuss the role that adaptive behavior and its assessment plays in the diagnosis of ID. The seven sections of the chapter are (a) key factors to keep in mind

when reading the chapter, (b) the assessment of adaptive behavior, (c) the use of standard error of measurement in score interpretation, (d) adaptive behavior versus problem behavior, (e) special considerations in the assessment of adaptive behavior, (f) guidelines for selecting an adaptive behavior instrument, and (g) future considerations. Throughout the chapter, adaptive behavior is defined as the collection of conceptual, social, and practical skills that have been learned and are performed by people in their everyday lives. Material included in the chapter regarding assessment guidelines and the technical adequacy of adaptive behavior assessment instruments is based on the published work of Finlay and Lyons (2002), Greenspan (1999, 2006a), Harrison and Raineri (2008), and Reschly et al. (2002).

KEY FACTORS TO KEEP IN MIND WHEN READING THIS CHAPTER

In this chapter we discuss in more detail the following 10 key factors about adaptive behavior and its assessment that are relevant to a diagnosis of ID:

1. There are three criteria for a diagnosis of ID: significant limitations in intellectual functioning, significant limitations in adaptive behavior, and age of onset before age 18. Adaptive behavior and intellectual functioning should be given equal consideration.
2. Adaptive behavior is a multidomain construct. The domains that have emerged from a long history of factor-analytic studies are consistent with a conceptual model of adaptive behavior that has three general areas of adaptive skills: conceptual, social, and practical.
3. Adaptive behavior as defined in this *Manual* is the collection of conceptual, social, and practical skills that have been learned and are performed by people in their everyday lives.
4. The concept of adaptive skills implies an array of competencies and provides a foundation for three key points: (a) the assessment of adaptive behavior is based on the person's typical (not maximum) performance, (b) adaptive skill limitations often coexist with strengths, and (c) the person's strengths and limitations in adaptive skills should be documented within the context of community and cultural environments typical of the person's age peers and tied to the person's need for individualized supports.
5. Although no existing measure of adaptive behavior completely measures all adaptive behavior skills, most provide domain scores that represent the three domains used in this *Manual*: conceptual, social, and practical. A comprehensive assessment of adaptive behavior will likely include a systematic review of the individual's family history, medical history, school records, employment records (if an adult), other relevant records and information, as well as clinical interviews with a person or persons who know the individual well.

6. For a person with ID, adaptive behavior limitations are generalized across the domains of conceptual, social, and practical skills. However, because subscale scores on adaptive behavior measures are moderately correlated, a generalized deficit is assumed even if the score on only one domain meets the operational criterion of being approximately two standard deviations below the mean. A total score of two standard deviations below the mean from an instrument that measures conceptual, social, and practical skills will also meet the operational definition of a significant limitation in adaptive behavior.
7. It is important to recognize that personal characteristics and environmental factors can present challenges to the assessment of adaptive behavior. These include (a) personal characteristics, such as concurrent sensory, motor, or mental disabilities; fatigue or illness; high anxiety levels; and the person's motivational history of interaction in assessment situations and (b) environmental factors, such as absence of participation in community settings.
8. Problem or maladaptive behavior is not a characteristic or domain of adaptive behavior, although it often influences the acquisition and performance of adaptive skills. The presence of problem behavior(s) is not considered to be a limitation in adaptive behavior, although it may be important in the interpretation of adaptive behavior scores for diagnosis. The distinction between adaptive behavior and problem behavior is discussed later in this chapter.
9. Adaptive behavior must be examined in the context of developmental periods of infancy and early childhood, childhood and early adolescence, late adolescence, and adulthood. A continuing theme is the importance of the developmental relevance of specific skills within the three adaptive areas.
10. It is sometimes necessary to assess the previous functioning of the individual in those situations where a diagnosis of ID becomes relevant. A retrospective diagnosis may be required, for example, when clinicians are involved in determining eligibility for adult rehabilitation services, evaluating individuals for Social Security disability, or evaluating individuals involved in legal processes, such as guardianship petitions, competence determinations, or sentencing eligibility questions. If adaptive behavior assessments are used and reported in the records reviewed, clinicians should weigh the extent to which (a) multiple informants were used and multiple contexts sampled; (b) that limitations in present functioning were considered within the context of community environments typical of the individual's age peers and culture; (c) important social behavioral skills, such as gullibility and naïveté, were assessed; (d) behaviors that are currently viewed as developmentally and socially relevant were included; and (e) adaptive behavior was assessed in reference to typical and actual functioning in the community. The use of previously administered adaptive behavior scales in a retrospective diagnosis should address these five assessment standards.

Assessment of Adaptive Behavior

Use Standardized Measures

Significant limitations in adaptive behavior are established through the use of standardized measures and, like intellectual functioning, significant *limitations in adaptive behavior* are operationally defined as performance that is approximately two standard deviations below the population average on one of the three adaptive skills domains of conceptual, social, or practical. In evaluating the role that an adaptive behavior score—as assessed on a standardized measure—plays in making a diagnosis of ID, clinicians should (a) determine the standard error of measurement (see following section) for the particular assessment instrument used, realizing that the standard error of measurement is test-specific and is used to establish a statistical confidence interval within which the person's true score falls and (b) assure that within reporting, standard error of measurement is properly addressed.

Focus on Typical Performance

The assessment of adaptive behavior focuses on the individual's typical performance and not their best or assumed ability or maximum performance. Thus, what the person typically does, rather than what the individual can do or could do, is assessed when evaluating the individual's adaptive behavior. This is a critical distinction between the assessment of adaptive behavior and the assessment of intellectual functioning, where best or maximal performance is assessed. Individuals with an ID typically demonstrate both strengths and limitations in adaptive behavior. Thus, in the process of diagnosing ID, significant limitations in conceptual, social, or practical adaptive skills is not outweighed by the potential strengths in some adaptive skills.

Use Knowledgeable Respondents

Using standardized adaptive behavior measures to determine significant limitations in adaptive behavior usually involves obtaining information regarding the individual's adaptive behavior from a person or persons who know the individual well. Generally, individuals who act as respondents should be very familiar with the person and have known him/her for some time and have had the opportunity to observe the person function across community settings and times. Very often, these respondents are parents, older siblings, other family members, teachers, employers, and friends. Parents are often the best respondents available because they have known the individual the longest and observed attainment of developmental milestones, maturation, and the achievement of adaptive behavior skills. Because adaptive behavior assessment relies on third party respondents, it is important for clinicians to assess the reliability of any respondent providing adaptive behavior information. Obtaining information from multiple respondents and other relevant sources (e.g., school records, employment history, previous evaluations) is essential to providing corroborating information that provides a comprehensive picture of the individual's functioning.

When Standardized Assessments Cannot Be Used

If a standardized assessment measure cannot be used (e.g., if the assessment cannot be reliably administered per the test's recommended administrative procedures or if there are no reliable respondents to provide adaptive behavior information regarding the assessed person), other sources of adaptive behavior information can be used. In these infrequent cases, other information-gathering methods can be employed, such as direct observation (see chapter 8 for guidelines); review of school records, medical records, and previous psychological evaluations; or interviews with individuals who know the person and have had the opportunity to observe the person in the community but may not be able to provide a comprehensive report regarding the individual's adaptive behavior in order to complete a standardized adaptive behavior scale. In reference to any method used, when a standardized adaptive behavior assessment instrument cannot be used, the following guidelines should be followed:

- Use multiple types and sources of information to obtain convergence of information regarding the individual's limitations in comparison to same-age peers.
- Use reasonable caution when weighing qualitative information obtained from respondents, especially in the presence of conflicting information.
- Interpret results obtained from direct observations of adaptive skills with caution because these may not be reflective of the individual's typical behavior and may be a narrow measure of actual adaptive behavior. For example, having the person screw in a light bulb does not fully capture all aspects of the adaptive behavior of identifying when it is time to change a burnt light bulb, what wattage is needed for the replacement bulb, knowing how to get a replacement bulb, and safely accessing an electrical outlet and replacing the light bulb.
- Use clinical judgment (see chapter 8) to guide the evaluation of the reliability of information provided by respondents as well as possible sources of bias (positive or negative).
- Analyze critically all types of information for accuracy and pertinence. One should also consider the comparison group when determining significant limitations. For example, in some special education programs, a grade of *C* denotes something very different in achievement level than a *C* grade given in a general education classroom.

USE OF STANDARD ERROR OF MEASUREMENT IN SCORE INTERPRETATION

The established procedure in psychological measurement, in which standardized measures are used, is to report results using a statistical confidence interval around the obtained score(s). As discussed in chapter 4, the standard error of measurement, which varies by test, subgroup, and age group, is used to estimate this statistical confidence interval.

CHAPTER 12

SUPPORT NEEDS OF PERSONS WITH INTELLECTUAL DISABILITY WHO HAVE HIGHER IQ SCORES

Individuals with intellectual disability who have higher IQ scores face significant challenges in society across all areas of adult life, and many individuals who may not receive formal diagnoses of ID or who fall slightly above the upper ceiling for a diagnosis of ID share this vulnerability. Only through an increased understanding of the ongoing strengths and limitations of each individual with ID can we achieve better clinical judgment and identify appropriate supports and, with the provision of individualized supports, accomplish fairness in society.

OVERVIEW

Our purpose in this chapter is to (a) describe the support needs of individuals with *intellectual disability* (ID) who have higher IQ scores, (b) discuss how intellectual limitation exists along a continuum that reveals many similarities in human functioning limitations between individuals on either side of the definitional dividing line, and (c) reiterate the critical importance of creating accessible, individualized supports for these individuals. Those with ID who have higher IQ scores struggle in society (for more detailed analysis and references, see Snell & Luckasson, 2009). This is true despite the fact that all individuals with ID typically demonstrate strengths in functioning alongside relative limitations. Those with ID who have higher IQ scores comprise about 80 to 90% of all individuals diagnosed with ID. Frequently, they have no identifiable cause for the disability, they are physically indistinguishable from the general population, they have no definite behavioral features, and their personalities vary widely, as is true of all people. Although many of these individuals will need supports, some may be able to live independently, at least for part of the time. Documented successful outcomes of individuals with appropriate supports contrast sharply with incorrect stereotypes that these individuals never have friends, jobs, spouses, or children or are good citizens.

People in this group primarily are identified when they are in school, because school demands place their intellectual and adaptive behavior limitations in clear relief and because schools have a legal obligation to identify disabilities in all children. Beyond school age, however, when activities may be less "intellectual," bureaucracies do not routinely identify people because of intellectual limitations, and needed services and supports are unavailable or rejected. As a result, these people continue to experience significant difficulties achieving success or even a healthy existence in adulthood.

Frequently, the gap between their capabilities and the demands from their environments grows as they leave school, as society becomes more complex, and as the standards for successful adulthood climb. Well-designed individualized supports can help bridge the gap between capabilities and demands, but the reality is that many of these individuals do not have access to needed supports. Thus, life's demands frequently impose overwhelming challenges to those who live with significantly limited intellectual ability and adaptive behavior.

CLASSIFICATION SYSTEMS AND INTELLECTUAL DISABILITY

All people with ID, including those with higher IQ scores, belong to a single disability group (people with ID). However, the application of various classification systems to subdivide the group leads to somewhat different ways of understanding these individuals and their needs. As discussed in chapter 7, classification systems based on relevant criteria should be selected by clinicians and others for explicit professional purposes that benefit the individuals who are classified. For example, service providers may choose classification systems that subdivide the group of people with ID into smaller groups based on support needs, such as using the Supports Intensity Scale assessment to classify individuals by the intensity of their support needs (Thompson et al., 2004a).

The variety of classification systems based on different criteria may partially account for why this group historically has had so many different names. Earlier names, most of which now are highly stigmatizing (e.g., feebleminded, moron, moral idiots [Trent, 1994, p. 20]) were followed by new names taken from then current definitions or classification systems: educable mental retardation and mild mental retardation or names reflecting time periods challenging particular characterizations of this group or an expansion of this group: the "six-hour retarded child" (President's Committee on Mental Retardation, 1969), students with general learning disability (MacMillan, Siperstein, & Gresham, 1996), and the forgotten generation (the combined group of people with ID with higher IQ scores and people without ID but with lower IQ scores, whose IQ scores are just beyond the ID range; Tymchuk, Lakin, & Luckasson, 2001). Generally, the names have followed from the classification system or purpose for classifying.

Similarities to the Borderline Classification

Whatever classification system is used, however, it is critical to point out that the challenges faced by individuals with ID who have higher IQ scores are significant. Thus,

references to "mild" are misleading. Moreover, individuals with ID with higher IQ scores (slightly below the ceiling of approximately 70–75) share much in common with individuals without a diagnosis of ID whose functioning is sometimes referred to as *borderline* (individuals who do not technically have ID but who have low IQ scores, above the ceiling of approximately 70–75). Edgerton wrote that "perhaps the most sobering realization is that the majority of these individuals [former 'six-hour retarded children'] are not cited in the research literature nor are they known to the mental retardation/developmental disabilities service delivery system" (Edgerton, 2001, p. 3).

Mild Intellectual Disability Is Misleading

In some ways, it may seem counterintuitive to consider the challenges of individuals with ID with higher IQ scores as being equal to or sometimes greater than those with ID at lower IQ scores. Several factors, however, aggravate their challenges: expectations for performance are higher for people with ID with higher IQ scores than for those with lower IQ scores; the tasks given to them are more demanding because of the higher expectations; and a failure to meet those expectations is frequently met by others blaming the individual or the individual blaming him or herself. Moreover, many individuals with ID who have higher IQ scores attempt to hide their disability or attempt to pass as "normal" or try to appear intellectually capable and thus miss out on or even reject accommodations that might have been available to them if their disability had been declared or identified. In addition, the impact of their ID may be increased by the lack of access to needed mental health care, medical care, dental care, nutrition, and relationship and parenting assistance. Society's increasing lack of neighborly care for one another may hit people with ID in poorer neighborhoods especially hard.

To further describe the challenges faced by many individuals with ID who have higher IQ scores, in this chapter we address areas in which societal threats are especially marked (e.g., education, socioeconomic status, employment, and housing), and the often inadequate response systems regarding individuals with intellectual limitations that increase their vulnerability in everyday life. The chapter concludes with a discussion of the need for a supports framework that spans IQ limitations.

EVERYDAY LIVES OF PEOPLE WITH INTELLECTUAL DISABILITY WHO HAVE HIGHER IQ SCORES

The lifelong experience of having reduced intellectual and adaptive abilities creates a vulnerability that is shared among members of this group. As adults, these people have limited academic skills, are often poor, are underemployed or unemployed, and tend not to live independently. These societal issues impacting their everyday lives are summarized in Table 12.1 and discussed more fully on subsequent pages.

User's Guide

To Accompany the 11th Edition of Intellectual Disability: Definition, Classification, and Systems of Supports

Applications for Clinicians, Educators, Organizations Providing Supports, Policymakers, Family Members and Advocates, and Health Care Professisonals

Developed by the AAIDD User's Guide Work Group

Robert L. Schalock, Ruth Luckasson, Val Bradley, Wil Buntinx, Yves Lachapelle, Karrie A. Shogren, Martha E. Snell, James R. Thompson, Marc Tassé, Miguel A. Verdugo-Alonso, and Michael L. Wehmeyer

aaidd
American Association on Intellectual and Developmental Disabilities

Printed in the United States of America

FOSTERING JUSTICE WHEN DEALING WITH FORENSIC ISSUES

Clinicians in the field of ID may be involved in forensic issues that arise when persons with ID are involved with the civil or criminal justice system. The more common of these forensic issues center around personal competence, guardianship, property and financial management, victimization in crime, or accusations of committing a crime. This section of the *User's Guide* discusses best practices and clinical judgment guidelines that address how clinicians can foster justice when dealing with these forensic issues. These practices and guidelines relate to: (1) interpreting assessment information, (2) understanding foundational aspects of ID that are critically important in fostering justice for people with ID, and (3) overcoming common stereotypes.

Interpreting Assessment Information

There are five critical areas involving the valid interpretation of assessment information that have emerged from clinical experiences dealing with forensic issues. These five areas involve understanding the following: (1) the concept of a confidence interval (CI), (2) the concept of a cutoff score, (3) that corrections need to be made in an obtained IQ score if the score was based on aging norms (i.e., the Flynn effect; Flynn, 2006), (4) the influence of practice effects on test results, and (5) the potential effect on test results attributable to faking.

Confidence interval (CI). A score obtained on a standardized psychometric instrument that assesses intellectual functioning or adaptive behavior is not absolute because of variability in the obtained score because of factors such as limitations of the instrument used, examiner's behavior and expertise, personal factors (e.g., health status of the person), or environmental factors (e.g., testing environment or testing location). Thus, an obtained score may or may not represent the individual's actual or true level of intellectual functioning or adaptive behavior because of these aforementioned factors. *Standard error of measurement (SEM),* which varies by test, subgroup, and age group, is used to quantify the variability that is attributable to the test itself and *provides the basis for establishing a statistical CI within which the person's true score is likely to fall.*

- For well-standardized measures of general intellectual functioning, the SEM is approximately 3 to 5 points. As reported in the respective test's standardization manual, the test's SEM can be used to establish a *statistical confidence interval (CI) around the obtained score.* From the properties of the normal curve, a range of confidence can be established with parameters of at least one standard error of measurement (i.e. scores of about 66 to 74, 66% probability) or parameters of two standard error of measurement (i.e. scores of about 62 to 78, 95% confidence).
- For well-standardized measures of adaptive behavior the SEM for obtained scores is comparable to that of standardized tests of intelligence. Thus, the use of plus/minus one standard error of measurement yields a statistical confidence interval (around the obtained score) within which the person's true score will fall 66% of the time; the use of plus/minus two standard error of measurement yields a sta-

tistical confidence (around the obtained score) in which the person's true score will fall 95% of the time. Thus, an obtained score on an adaptive behavior scale should be considered as an approximation that has either a 66% or 95% likelihood of accuracy, depending on the confidence interval used. There is no evidence suggesting that the population mean on standardized tests of adaptive behavior is increasing at a rate comparable to that observed on standardized tests of intelligence (i.e., Flynn effect). Because of the differences in test construction and administration between intellectual functioning and adaptive behavior, practice effect is not an issue with standardized adaptive behavior scales. One source of measurement error may be specific to measures of adaptive behavior and that is the concern that individuals may exaggerate their adaptive skills when asked to self-report their adaptive behavior. For this reason, numerous sources (e.g. Edgerton, 1967; Finlay & Lyons, 2002; Greenspan & Switzky, 2006; Schalock et al., 2010) have recommended against relying on self-reported measures of adaptive behavior when ruling-in or -out a diagnosis of ID.

Cutoff score. A cutoff score is the score(s) that determines the boundaries of the "significant limitations in intellectual functioning and adaptive criteria" for a diagnosis of ID.

- For both criteria, the cutoff score is approximately 2 standard deviations (SD) below the mean of the respective instrument, considering the SEM (see *Confidence interval*) for the specific instrument used, and the strengths and limitations of the instrument.
- A fixed point cutoff for ID is not psychometrically justifiable. The diagnosis of ID is intended to reflect a clinical judgment rather than an actuarial determination.

Flynn Effect. The *Flynn Effect* refers to the increase in IQ scores over time (i.e., about 0.30 points per year). The Flynn Effect effects any interpretation of IQ scores based on outdated norms. Both the 11th edition of the manual and this *User's Guide* recommend that in cases in which a test with aging norms is used as part of a diagnosis of ID, a corrected Full Scale IQ upward of 3 points per decade for age of the norms is warranted (Fletcher et al., 2010; Gresham & Reschly, 2011; Kaufman, 2010; Reynolds et al., 2010; Schalock et al., 2010). For example, if the Wechsler Adult Intelligence Scale (WAIS-III; 1997) was used to assess an individual's IQ in July, 2005, the population mean on the WAIS-III was set at 100 when it was originally normed in 1995 (published in 1997). However, on the basis of Flynn's data (2006), the population mean on the WAIS-III Full-Scale IQ corrected for the Flynn Effect would be 103 in 2005 (9 years $\times$ 0.30 = 2.7). Hence, using the significant limitations of approximately 2 SDs below the mean, the Full-Scale IQ cutoff would be approximately 73 and not approximately 70 (plus or minus the SEM).

Practice effect. The practice effect refers to gains in IQ scores on tests of intelligence that result from a person being tested on the same instrument. The established clinical best practice is to avoid administering the same intelligence test within a year to the same individual because it will often lead to an overestimation of the examinee's true intelligence.

Claims of faking. Sometimes in a contested legal case an allegation of intentional "faking bad" is made, asserting that the individual is attempting to gain a benefit by deliberately faking a disability. Such claims of faking, when they are made, are usually in cases involving mental disorders because mental illness can have a later-life onset, subjective symptoms, and waxing and waning symptoms.

Allegations that an individual is intentionally faking bad, by faking ID, occur in some legal cases. The cases in which such allegations occur are cases in which rights such as eligibility for financial supports or exemption from the death penalty would come into play if the individual has an ID (Keyes, 2004). The term *malingering* is often used to refer to "faking bad." The *DSM-IV-TR* (APA, 2000) defined malingering as intentionally and purposefully feigning an illness to achieve some recognizable goal or tangible benefit (e.g., feigning ID to be spared the death penalty). Such allegations that a person is faking ID must be analyzed cautiously, however, for several reasons. First, the elements required for a diagnosis of ID must have been present from an early age (ID must originate before the age of 18), so there is almost always a documented lifetime history, usually beginning at birth or early childhood and extending through the school years, of significant limitations in intellectual functioning and adaptive behavior. Second, in cases in which an earlier diagnosis of ID cannot be documented because the individual grew up in another country and/or there are no assessment records, a clinician may conduct or access a current assessment of intellectual functioning and adaptive behavior, including a history, to determine current functioning, and together with clinical judgment make a retrospective diagnosis if indicated. Third, the more common faking direction when an individual with ID attempts to fake is to "fake good" so as to hide their ID and try to convince others that he or she is more competent (Edgerton, 1967).

Claims of faking ID in an individual should be addressed by a clinician in ID conducting a thorough evaluation for ID using the diagnostic and clinical strategies outlined in the 11th edition of the AAIDD manual and in this *User's Guide.* The authors of this *User's Guide* are aware of the concern that some (e.g., Doane & Salekin, 2008) have expressed about the potential to feign deficits on currently used adaptive behavior scales. Clinicians need to be aware of this potential and ensure that they interview multiple individuals who know the person well and who have had the opportunity to directly observe the person engaging in his or her typical behaviors across multiple contexts (i.e., home, community, school, and work).

Clinicians who similarly attempt to use specific "malingering" tests in individuals with ID must use considerable caution because of two factors: (1) the lack of a research base supporting the accuracy of such tests for persons with ID (Hayes et al., 1997; Hurley & Deal, 2006); and (2) the documented misuse of common malingering tests even when the test manual explicitly precludes use with individuals with ID (Keyes, 2004). Standardized assessment instruments used to inform the clinician whether the person is putting forth his or her best effort (i.e., malingering) have not, for the most part, been normed for persons with ID (MacVaugh & Cunningham, 2009). In addition, recent studies have documented unacceptable error rates (i.e., false positive for malingering) when used with persons with IQ scores from 50 to 78 (Dean et al., 2008; Hurley & Deal, 2006). Thus, the assessment of "faking bad" with individuals with low IQs (i.e., below 80) should be conducted with great prudence when relying on standardized measures that are not strictly normed or validated with persons being assessed for ID.

INTELLECTUAL DISABILITY

Definition, Diagnosis, Classification, and Systems of Supports

Robert L. Schalock
Ruth Luckasson
Marc J. Tassé

12TH EDITION

American Association on Intellectual and Developmental Disabilities

aaidd

The suggested citation for the AAIDD Manual is as follows:

Schalock, R. L., Luckasson, R., & Tassé, M. J. (2021). *Intellectual disability: Definition, diagnosis, classification, and systems of supports* (12th ed.). American Association on Intellectual and Developmental Disabilities.

Published by
American Association on Intellectual and Developmental Disabilities
8403 Colesville Road, Suite 900
Silver Spring, MD 20910
www.aaidd.org

To order
AAIDD Order Fulfillment
8403 Colesville Road, Suite 900
Silver Spring, MD 20910
Phone: 202-387-1968 x216
Email: books @aaidd.org
Online: http://aaidd.org/publications/bookstore-home

Product No. 4174
ISBN 978-0-9983983-6-5

2

Definition of Intellectual Disability

The Definition of Intellectual Disability

ID is characterized by significant limitations both in intellectual functioning and in adaptive behavior as expressed in conceptual, social, and practical adaptive skills. This disability originates during the developmental period, which is defined operationally as before the individual attains age 22.

Users of this chapter will find:

- An operational definition of ID
- The assumptions of the definition of ID.
- The purposes of the definition of ID.
- Historical definitions of ID formulated by AAIDD and by APA.
- Alignment of definitions of ID among AAIDD, APA, and WHO.
- What has changed and what has remained consistent in the definition of ID over time.
- Practice guidelines regarding defining ID and implementing the definition.

The purposes of a definition are to explain precisely a term (in this case ID), establish the meaning and boundaries of the term, and separate who is included within the term from those who are outside the term. Significant consequences can result from the way a term is defined. A definition can make someone eligible or ineligible for services, subjected to something or not subjected to it (e.g., involuntary commitment), exempted from something or not exempted (e.g., from the death penalty), included or not included (as to protections against discrimination and equal opportunity), and/or entitled or not entitled (e.g., certain Social Security benefits or other financial benefits).

The authoritative definition of ID is that of the AAIDD. The definition of ID found in the 12th edition of the AAIDD Manual is the same as that found in the 11th edition, except for the age of onset criterion. In the 11th edition the age of onset criterion was "originates before age 18" (Schalock, Borthwick-Duffy et al., 2010, p. 1). In the 12th edition, the age of onset criterion is stated as "originating during the developmental period, which is defined operationally as before the individual attains age 22." Readers are referred to the section in Chapter 3 entitled, "Age of Onset" for the rationale and justification for this change.

Assumptions Regarding Implementation of the Definition

Assumptions are an essential part of the definition of ID because they clarify the context from which the definition arises and indicate how the definition should be applied. Thus, the definition of ID cannot stand alone. The following assumptions are essential to the definition's implementation:

1. Limitations in present functioning must be considered within the context of community environments typical of the individual's age peers and culture.
2. Valid assessment considers cultural and linguistic factors, as well as differences in communication, sensory, motor, and behavioral factors.

3. Within an individual, limitations often coexist with strengths.
4. An important purpose of describing limitations is to develop a profile of needed supports.
5. With appropriate personalized supports over a sustained period, the life functioning of the person with ID generally will improve.

These five assumptions reflect the distinction between the diagnosis of ID (which involves significant limitations in both intellectual functioning and adaptive behavior and age of onset during the developmental period) and the expression of ID, which involves the reciprocal engagement among human functioning dimensions, systems of supports, and human functioning outcomes. This reciprocal engagement is depicted in the integrated model of human functioning that was discussed in Chapter 1 and shown graphically in Figure 1.1.

Definitional Consistency

Although the term or name has changed over time, the three essential elements of ID—limitations in intellectual functioning, behavioral limitations in adapting to environmental demands, and early age of onset—have not changed significantly over the last 60 years (Schalock, Borthwick-Duffy et al., 2010; Tassé et al., 2016). This historical consistency regarding the AAIDD definitions is shown in Table 2.1.

Alignment of Definitions Among AAIDD, American Psychiatric Association, and World Health Organization

Alignment Between AAIDD and American Psychiatric Association

The definition of ID promulgated by AAIDD in the 12th edition is that ID is characterized by significant limitations both in intellectual functioning and in adaptive behavior as expressed in conceptual, social, and practical adaptive skills, and that this disability originates during

information needed to complete a standardized adaptive behavior scale, an alternative assessment should be used with caution and adaptive behavior information should be obtained from: (a) interviewing multiple respondents (e.g., family members, teachers, neighbors, job supervisors, etc.) who may have more discrete but overlapping information about the person's typical performance across all three domains of adaptive behavior (conceptual, social, and practical); and (b) reviewing thoroughly all available records, including educational, social, and medical, that might contain collateral information regarding the person's adaptive behavior.

- Does not administer the same intelligence test within the same year to the same individual, because frequent re-administrations may lead to overestimating the examinee's true intelligence (i.e., practice effects).
- Conducts the evaluation in a comfortable environment free from extraneous noise, distractions, and interruptions.
- Uses assessment strategies that are appropriate to the individual's cultural and linguistic background.
- Reviews social, educational, and medical records/histories.
- Synthesizes information from multiple sources, and gives equal weight and joint consideration to intellectual functioning and adaptive behavior information in a diagnosis of ID.

Avoiding False Positives and False Negatives

Because of the high stakes involved in a correct diagnosis, it is essential to avoid making an incorrect diagnosis of either a false positive (the person is incorrectly/falsely diagnosed as an individual with ID when in fact the person does not have ID) or a false negative (the diagnosis of ID is not made when the person does in fact have ID). The following strategies can assist in avoiding these potential errors:

- Recognizing that false positives may occur when a test is used whose norms and language are culturally or linguistically inappropriate for the individual assessed.

- Giving equal consideration to intellectual functioning and adaptive behavior scores in the diagnosis of ID.
- Recognizing that all people with ID have strengths, but that the diagnosis of ID focuses on their significant limitations.
- Being aware of how one's diagnostic accuracy is influenced by an assessment instrument's sensitivity and specificity. "Sensitivity" refers to the proportion of cases in which there are significant deficits in intellectual functioning or adaptive behavior, and for which a diagnosis of ID has been made. In distinction, "specificity" refers to the proportion of cases in which the test's standard scores exclude individuals who do have a diagnosis of ID. Matthey and Petrovski (2002) and Balboni et al. (2014) suggest that sensitivity coefficients of >.70 and specificity coefficients of >.80 are considered appropriate benchmarks to be attained by diagnostic tests.
- Using the 95% confidence interval to establish the interval within which the individual's true score falls.
- Synthesizing and corroborating information from multiple sources, including a thorough social history and medical and educational records.

Resolving Claims of Faking

Sometimes in a contested legal case an allegation of intentional faking, poor effort, or malingering may be made, claiming that the individual is attempting to gain or benefit by deliberately faking a disability. These cases typically involve a secondary gain associated with the evaluation outcome, such as eligibility for financial supports, or mitigation or exemption from a criminal penalty. Resolving allegations of the individual faking to gain a benefit is facilitated when clinical judgment involves:

- Verifying that the intellectual functioning, adaptive behavior, and age of onset criteria for a diagnosis of ID are met.
- Conducting or procuring a current assessment of intellectual functioning and adaptive behavior in cases in which an earlier diagnosis

of ID cannot be documented because the individual grew up in another country or there are no assessment records.

- Synthesizing information from multiple sources, including a thorough social, medical, and educational history.
- Not using self-report for the assessment of adaptive behavior. Self-report may be susceptible to biased responding.
- Realizing that most instruments used to detect malingering have not been normed for individuals with ID (Dean et al., 2008; MacVaugh & Cunningham, 2009).
- Exercising clinical judgment in interpreting all information.

Making a Retrospective Diagnosis

It is possible to make a retrospective diagnosis of ID after the individual attains age 22. To do so, the clinician must establish that the significant deficits in both intellectual functioning and adaptive behavior were present during the period of the individual's development. In this situation, when the person does not have a diagnosis of ID established during the developmental period, it is necessary for clinicians to assess the past functioning of the individual to determine whether a valid diagnosis of ID applies to the person. Such a retrospective diagnosis may become relevant in determining eligibility for adult rehabilitation services, evaluating individuals for Social Security disability, or evaluating individuals involved in legal processes such as guardianship petitions, competency determinations, or sentencing eligibility questions. In these situations, using clinical judgment to enhance the accuracy of a retrospective diagnosis of ID involves:

- Using a thorough social, medical, and educational history.
- Basing the diagnosis on multiple valid data points.
- Interpreting previously administered adaptive behavior assessments in terms of the extent to which the assessments: (a) included direct observation of the person engaging in his or her typical behaviors in the home, community, school, and/or work; (b) used multiple

informants and multiple contexts; (c) measured limitations in important social behavior skills such as gullibility and naiveté; (d) used an adaptive behavior assessment instrument that included the behaviors that are viewed as developmentally and socially relevant; (e) recognized that adaptive behavior refers to typical functioning and not to capacity or maximum functioning; and (f) recognized the limitations in present functioning are considered within the context of community environments typical of the individual's age peers and culture.

- Interpreting previously administered intellectual functioning assessments in terms of the extent to which the assessment: (a) used a standardized and individually administered comprehensive intelligence test; (b) was the [then] most recent version of the standardized test used, including the most recent norms; (c) took into consideration the confidence interval within which the person' true score fell; and (d) was corrected for the age of the norms employed. Current best practice guidelines recommend that in cases in which an IQ test with aged norms is used as part of a diagnosis of ID, a correction of the full-scale IQ score of 0.3 points per year since the test e-norms were collected is warranted (Fletcher et al., 2010; Gresham & Reschly, 2011; Kaufman, 2010; Reynolds et al., 2010).

Practice Guidelines Regarding Assessing Intellectual Functioning and Adaptive Behavior and Making a Diagnosis of Intellectual Disability

A major emphasis in the 12th edition is to provide best practice guidelines regarding the diagnosis of ID. Practice guidelines regarding the assessment of intellectual functioning are found in Table 3.5; for the assessment of adaptive behavior in Table 3.6; and for making a diagnosis of intellectual disability in Table 3.7.

The DEATH PENALTY and INTELLECTUAL DISABILITY

Edward A. Polloway, Editor

American Association on Intellectual and Developmental Disabilities
Washington, D.C.

Published by
American Association on Intellectual and Developmental Disabilities
501 3rd Street, NW, Suite 200
Washington, D.C. 20001
www.aaidd.org

To order
AAIDD Order Fulfillment
501 3rd Street, NW, Suite 200
Washington, D.C. 20001
phone: 202-387-1968 x 216
email: books@aaidd.org
online: http://aaidd.org/publications/bookstore-home

Product No. 4134

Printed in the United States of America

3 Intellectual Disability

Gary N. Siperstein
Melissa A. Collins

Within the diagnosis of intellectual disability (ID), there is immense variation in both cognitive functioning and adaptive behaviors, with the majority of individuals with ID functioning at the upper end of the disability range. In 1992, the American Association on Mental Retardation (AAMR) estimated that 89% of people with ID fell within the mild category (Petersilia, 2000; more recent estimates are not available due to the elimination of the severity categories in official American Association on Intellectual and Developmental Disabilities (AAIDD) definitions and the extreme difficulty in estimating prevalence.). It is these individuals—at the upper end of the ID spectrum—who are the focus of this chapter, because in a categorical sense they are the most difficult to diagnose and the least immediately recognizable as having ID. As MacMillan, Siperstein, and Leffert (2006) put it, in contrast with others with more significant disabilities, "individuals with mild intellectual disability represent 100% of the cases in which the answer to the question 'Does this individual have an intellectual disability?' is actually in doubt," and, consequently, "professionals must depend upon a definition and classification system for help in resolving uncertainty" (p. 197). The history of this group of persons is complex and controversial. Due to issues related to classification, increased vulnerability to negative outcomes, and inaccurate public perceptions, this group requires careful attention, particularly within the context of the American judicial system.

History of Categorization Within ID

It is important to note that the differentiation of intellectual disability into discrete categories has a long history. In fact, the subcategorization of ID based on functional level has been an issue of contention for hundreds of years, and the classification of the

group of persons with mild ID specifically has long been a source of debate. Interestingly, one of the first attempts to classify ID began with the differentiation of ID from mental illness, a distinction first made in the sixteenth and seventeenth centuries by philosophers, including Fitzherbert in 1534 and John Locke in 1690 (Braddock & Parish, 2002; MacMillan & Reschly, 1997). Long after that initial and important distinction was made, practitioners recognized that there was a need for further classification, with subcategorization within ID beginning in the mid-nineteenth century (MacMillan & Reschly, 1997). Because it is a disability characterized by impairments in cognitive functioning, it is not surprising that ID was first differentiated by degree of impairment. For example, though offensive terms by today's standards, "idiot" and "imbecile" were used to distinguish individuals by perceived level of functioning as early as the 1830s, as was "moron" in the early 20th century. However, although level of impairment was one of the earliest categorization dimensions, other factors such as etiology were also considered along the way.

Beginning in the late nineteenth century, individuals were often grouped together based on the believed sources of their disability. For example, William Ireland based his 1877 10-category classification system largely on etiology (e.g., "genetous idiocy," "inflammatory idiocy") and included "idiocy by deprivation" for those without evident physical causes (Scheerenberger, 1983). Etiology continued to be used as a major factor in classification through the mid-twentieth century, with Heber's categories including ID due to disease or infection in 1959 and cerebral palsy and convulsive disorders in 1961. Though classification by etiology eventually fell out of practice, some continued to argue even as late as the 1980s that behavioral differences existed for individuals with ID of different etiologies (e.g., Burack, Hodapp, & Zigler, 1988).

Another framework for classification emerged in the mid-twentieth century when service providers began to classify individuals with intellectual disability based on their prognosis and malleability. In this period, perceived educability became the main dimension of classification, and individuals were diagnosed as "educable" or "trainable." Individuals considered educable were believed to be higher functioning and capable of learning some academic subjects, while those considered trainable were believed to be incapable of academic learning but to be capable of learning certain life skills if given appropriate support (Weber, 1962). Thus, by the 1960s, classification had shifted from focusing on the etiology of disability to focusing on the potential to learn if given appropriate supports.

Irrespective of these different classification factors, arguably no factor has been more integral to the history of classification of intellectual disability than the intelligence quotient (IQ). With the emergence of IQ testing in 1905, the identification of individuals with intellectual disability was standardized for the first time (Cardona, 1994). The former subjective categories of "idiot" and "imbecile" that had been used since the 1830s were standardized such that individuals with an IQ test score in the range of 50–75 were considered morons, those with IQ test scores between 25–50 were imbeciles, and those

with IQ scores less than 25 were idiots. Over time, the terminology for different categories shifted away from these pejorative terms, but the categories of intellectual disability continued to be based largely on IQ (MacMillan & Reschly, 1997).

Although IQ did allow for more standardized diagnosis within ID, it has not been without criticism and debate. The underlying theme of this controversy is directly relevant to the group of persons with mild levels of ID, and it relates to how inclusive or exclusive the diagnosis and subcategories of ID should be. First, there was significant debate within the field regarding where to set the upper IQ score cutoff. This boundary determines who has a disability and who does not, and, consequently, determines who is eligible to receive services and supports. The most common cutoff has typically been two standard deviations below the mean (100), or an IQ test score of 70–75, but there was a period in the 1960s where the cutoff was recommended to be as high as one standard deviation (–1 *SD*) below the mean, or an IQ test score of 85 (Heber, 1961). This change led to an instantaneous potential increase in prevalence of ID from 2% to 16% of the population (Zigler, Balla, & Hodapp, 1984). Until Grossman's 1973 definition, those individuals who were between one and two standard deviations below the mean (IQ 85–70) were labeled "borderline MR," a distinct category within ID that was subsequently eliminated when a new definition was released that year. Some have argued that the elimination of this category had negative impacts for those in the borderline group, as it left them without any means for government support (Zetlin & Murtaugh, 1990). At the same time, however, the positive outcome from this change was to avoid association of people with higher functioning with people classified as having intellectual disability.

The other source of debate over IQ relates to the level of flexibility allowed and/or needed in interpreting test results. While some argue for flexibility in interpreting IQ test scores based on the standard errors of the tests (Baroff, 2006), others have countered that too much flexibility can lead to subjective evaluation and, consequentially, perceptions of unfairness in diagnosis (MacMillan, Siperstein, & Leffert, 2006; MacMillan & Siperstein, 2002). Again, this debate is especially relevant to the group of persons with mild levels of ID. Clearly, whether the cutoff is rigid or flexible is most consequential for people with ID at the upper end of the spectrum or on the borderline, as too rigid a cutoff may lead to false negatives and denial of services for those who, for all intents and purposes, actually have ID, while too flexible a cutoff may lead to ambiguity and indecision.

Notwithstanding these debates, IQ did more than provide much-needed standardization in the diagnosis of ID. The emergence of IQ also presented a means of estimating expected distributions of intelligence within the population. IQ is a standardized measurement of intelligence that is calibrated to be normally distributed (i.e., essentially a model to reflect the fact that most individuals have IQs that are within one standard deviation above or below the mean and relatively few have IQs that are further above or below) around the population mean of 100. However, it was discovered early

on that there are many more in the low range than expected based on the normal distribution (i.e., Gaussian curve). This statistical aberration led to the hypothesis that there are two groups that have been labeled in a number of different ways over time. In 1933, Lewis made one of the earliest attempts to differentiate these groups by dividing ID into subcultural, or "extreme variety of normal variations of cognitive capacities," and pathological, or "mental defectiveness . . . associated with and in most cases a result of recognized organic insult" (as cited in Burack, 1990, p. 31). Other labels have included "physiological" and "pathological" (Dingman & Tarjan, 1960) and "organic" and "non-organic" or "cultural-familial" (Burack, 1990, pp. 30-31).

Regardless of the specific wording applied, these terms reflect a two-group approach, which is "based on the theoretical premise that the majority of [individuals with intellectual disability] do not differ qualitatively from the normal [*sic*] population" (Burack, 1990, pp. 30–31). According to this model, there are individuals "who deviate *statistically* from the norms for average functioning even though they may not differ *qualitatively*" (pp. 30–31). In other words, only a fraction of individuals with ID are qualitatively different from the normal population in that their impaired cognitive functioning is due to biological factors such as chromosomal irregularities and physical trauma. Conversely, the majority of individuals with ID reflect expected variability in intelligence based on the Gaussian curve and statistical chance, even in the absence of physical causes. For some time, these individuals were considered to have familial intellectual disability, which has no known etiology, is more common in groups in lower socioeconomic positions, and typically have IQ test scores in the upper range of ID.

In sum, ever since the advent of compulsory education in the United States brought their existence to light and no matter the particular label applied to the group (e.g., cultural-familial, familial, educable mentally retarded, intergenerational), individuals with levels of ID at the upper end of the spectrum have long been recognized as possessing different abilities and needs than those with more significant impairments (MacMillan et al., 2006). Moreover, there has been a clear trend over the last few centuries of confusion and disagreement regarding those near the higher end of the spectrum who have less severe levels of impairment. Indeed, throughout the many fluctuations in terminology and classification across the history of intellectual disability, those individuals who are near the top of the spectrum and whose impairments are less extreme and often are context-specific have consistently presented the greatest challenge for classification.

Current Status of Categorization of Intellectual Disability

Given the history of classification within the field of ID, it is perhaps not surprising that some disagreement remains regarding how to most appropriately classify individuals with these disabilities. AAMR removed subcategories from its 1992 definition of mental retardation in order to focus less on deficits and more on supports needed (Luckasson et al., 1992). Rather than having classifications of mild, moderate, severe,

and profound, the 1992 definition divided individuals based on a continuum of supports needed and included the designations of intermittent, limited, extensive, and pervasive levels of supports (Greenspan & Switzky, 2006; Luckasson et al., 1992). The 2002 AAMR (Luckasson et al., 2002) and 2010 AAIDD (Schalock et al., 2010) definitions continue the practice from the 1992 manual with regard to the elimination of severity levels and the omission of any reference to the former subcategories. Following the recommendations of these manuals, the use of the term "mild intellectual disability" is no longer affirmed, although clearly there is a substantial number of individuals characterized as having ID whose level of functioning is at the higher end of the IQ test score range for the ID spectrum. As noted earlier, it is this population of individuals that is the focus of consideration in *Atkins* cases.

Ultimately, perspectives on the most appropriate way to differentiate groups within the ID spectrum depends on the individual's or group's goals of classification. An effective classification system allows for a systematic way to allocate resources and distribute services. However, the purpose behind classification can determine how specific or inclusive categories should be. As MacMillan and Reschly (1997) contended, "persons with different perspectives (legislators who must appropriate funds vs. advocates who wish to serve all deserving persons) often attempt to promote less or more inclusive interpretations of existing diagnostic constructs" (p. 48). For the purposes of the courts, having clearly defined classification levels could facilitate diagnosis of individuals on the borderline and could increase awareness of the different functioning levels of individuals with ID. Furthermore, classification levels also could clarify expectations for abilities and challenges for individuals of different functioning levels. In this way, classification levels are more than just diagnostic terminology, but rather provide information for people interacting with, providing services to, or making decisions about persons with ID. Therefore, since the dissolution of "mild ID"' as an official category in the AAIDD definition, some have advocated for a return to severity level classification (e.g., MacMillan et al., 2006), and many continue to use the category designation in research and practice (e.g., Larkin, Jahoda, & MacMahon, 2012). Nevertheless, it is important for individuals in the court system to examine the diagnosis of ID within the context of scientific shifts and to ensure that treatment of intellectual disability in the court system reflects the current standards of the field.

What Is the Upper End of the Intellectual Disability Spectrum Today?

Individuals at the upper level of the ID spectrum differ in the presentation of their disability when compared to those with more significant levels of impairment. MacMillan and colleagues (2006) contended that individuals at the upper end of the ID spectrum have "distinctive characteristics" that distinguish them from those with more severe levels of impairment, and that the differences among severity levels are "qualitative rather than merely a matter of degree" (p. 198). To be sure, many of the impairments or

difficulties shown in moderate or severe levels of impairment are not present in those at the higher end of the ID spectrum. For individuals at the middle and lower levels of ID, in addition to academic and cognitive difficulties, daily living skills are impaired, independent living is often not possible, and it is generally easy to detect that a disability is present. They may have difficulty in basic adaptive behaviors, such as toileting and dressing, and struggle with even low-level cognitive skills, such as memorizing their phone numbers (MacMillan et al., 2006).

Comparatively, the limitations in individuals with ID at the upper end of the spectrum are more subtle, more difficult to detect, and often context-specific. Most individuals with ID at the upper end of the spectrum do not experience problems in the practical skills measured by adaptive behavior scales, such as dressing oneself or using the telephone. However, they typically display significant deficits in adaptive skills in the social and conceptual domains. Family members, employers, friends, and others who interact closely with an individual with ID at the upper end of the spectrum typically observe qualitative differences in their behavior in comparison to others in the environment. While they generally do not recognize the problem as ID, they frequently describe the individual as displaying characteristics of ID such as being "slow," having difficulties with memory and directions, or understanding social pragmatics.

Indeed, rather than displaying significant general dysfunction, individuals with ID at the upper end of the spectrum struggle more with abstract thinking (MacMillan, Siperstein, & Gresham, 1996), and they generally show deficits in planning, problem solving, and decision making. They may also have difficulty in social perception, understanding, and judgment (Leffert & Siperstein, 2002; Leffert, Siperstein, & Widaman, 2010; MacMillan et al., 2006). Additionally, some individuals with ID at the upper end of the spectrum are vulnerable to experiencing comorbidity, such as attention deficit hyperactivity disorder (Rose, Bramham, Young, Paliokostas, & Xenitidis, 2008), autism (Matson & Shoemaker, 2009), communication disorders (Pinborough-Zimmerman, Satterfield, & Miller, 2007), and psychiatric conditions, such as schizophrenia (Lehotkay, Varisco, Deriaz, Douibi, & Carminati, 2009). Overall, compared to the typically developing population, they are more likely to live in poverty (Emerson, 2007); be socially isolated (Hemphill & Siperstein, 1990; Lippold & Burns, 2009); and be more suggestible, gullible, and credulous (Baroff, 2006), putting them at risk for engaging in criminal or antisocial behaviors. Considering the cumulative impact of these challenges, the group of persons with ID at the upper end of the spectrum seems at times to be at risk for unsuccessful integration into society.

Despite these risk factors, if given the opportunity, individuals with ID at the upper end of the spectrum can participate in their communities in ways that far exceed public expectations. After graduating high school, some individuals with ID at the upper end of the spectrum take advantage of the limited but increasing opportunities to attend postsecondary education, such as vocational programs at community colleges (Papay & Bambara, 2011) and, more recently, participating in 4-year colleges. Furthermore,

although limitations in reading and similar academic skills may hinder their chances of successful employment in certain areas (Baroff, 2006), people with ID at the upper end of the spectrum can sustain gainful employment in appropriate settings (e.g., Jahoda et al., 2009). Indeed, the abilities of people with ID at the upper end of the spectrum are evident throughout their daily activities. People with mild levels of ID can participate in a variety of community and leisure activities (Dusseljee, Rijken, Cardol, Curfs, & Groenewegen, 2011), such as competing in sporting events (Harada, Siperstein, Parker, & Lenox, 2011), attending religious services (Shogren & Rye, 2005), volunteering in the community (Trembath, Balandin, Stancliffe, & Togher, 2010), driving cars (Dixon & Reddacliff, 2001), and having long-term relationships (Siebelink, de Jong, Taal, & Roelvink, 2006). Additionally, research has demonstrated their ability to master independent living skills, such as using ATMs (Davies, Stock, & Wehmeyer, 2003), cooking (Taber-Doughty et al., 2011), and making financial decisions (Suto, Clare, Holland, & Watson, 2005). Many can use computers, the Internet, and other technologies (Wehmeyer et al., 2006), and navigate urban settings (Wright & Wolery, 2011) or ride public transportation (Davies, Stock, Holloway, & Wehmeyer, 2010). This range of abilities and activities corresponds with the fact that many are able to live independently, with varying levels of support (Bond & Hurst, 2009).

The range of abilities and activities of people at the upper end of the ID spectrum and the varying presentation of their disabilities make identification a challenge. In other words, individuals functioning in this range can be difficult to identify as a result of their higher levels of adaptive behaviors. Consequently, many may go undiagnosed or misdiagnosed because they do not demonstrate obvious impairments in these skills and behaviors.

In fact, whether or not an individual functioning in this range is actually diagnosed is largely a factor of context and definition. Both the 2002 AAMR (Luckasson et al., 2002) and 2010 AAIDD (Schalock et al., 2010) definitions of ID included requirements that "limitations in present functioning must be considered within the context of community environments typical of the individual's age peers and culture" (Luckasson et al., 2002). Any consideration of context in diagnosing ID highlights the important role of current societal demands. As Connolly and Bruner (1974) stated, "in any given society there are sets of skills which are essential for coping with existing realities," and the extent to which any individual functions in society depends on the acquisition and application of these skills (p. 4). Therefore, Leland (1969) contended that

> we must remember that as society becomes more complex and as the intellectual requirements placed on the individual become more demanding, many behaviors which in a previous period were acceptable as representing an average level, no longer may be considered to do so. (p. 534).

Taking a historical perspective shows us that higher-functioning individuals, who we now know constitute the majority of those with mild levels of ID, were not even

diagnosed for hundreds of years. Most people in the general population were illiterate and manual jobs were not cognitively demanding; consequently, intellectual disability was not manifested (Mesibov, 1974).

Thus, given that context is such a critical factor in diagnosis, ID "can be understood only in terms of the transaction between the individual's cognitive inefficiencies and the environmental demands for problem-solving" (MacMillan, Siperstein, & Gresham, 1996, p. 356). Consequently, the same individual could potentially be seen as competent in one environment and incompetent in another. Indeed, deficits associated with the upper end of ID generally present themselves only within specific contexts, such as schools (MacMillan et al., 2006). The importance of context for diagnosis is reflected by the fact that it is much more common in low SES, minority families (Browman, Nichols, Shaughnessy, & Kennedy, 1987) and may be related to geographic locale (Reschly & Jipson, 1976). Therefore, researchers have advocated for sensitivity to differences in language and culture when evaluating cognitive abilities in our more diversified society (Greenspan & Switzky, 2006).

Public Perceptions and Misconceptions

The conundrum, both for the courts and for society at large, is that the public may not *perceive* these individuals to have disabilities. Indeed, even considering their impaired cognitive and social functioning, the greatest challenge that individuals with mild levels of ID face is their own invisibility. Individuals with mild levels of ID are in a precarious position—they possess a number of abilities that distinguish them from others with greater levels of impairment, yet they are still vulnerable to a host of challenges as compared to the typically developing population. The influential role of context creates much ambiguity in the diagnosis of ID, and the manifestations of mild levels of ID do not align with societal expectations. Notwithstanding the conceptual and definitional approach to categorizing ID and defining and understanding mild levels of ID, what is the societal understanding of the disability?

Research overall has shown vast misconceptions regarding ID in the general public, and even among professionals. For example, although pediatricians have been shown to recognize different abilities for children with ID at the upper end of the spectrum, as compared with those at the middle and lower ranges of the ID spectrum (Wolraich, Siperstein, & O'Keefe, 1987), disagreement exists among professionals from medicine, education, and social work regarding the abilities and disabilities of persons across the spectrum of the disability (Wolraich & Siperstein, 1986). While professionals may disagree about the extent of variability in functioning within ID, the public does not even recognize that such variability exists. According to public opinion, the average person with ID is believed to have a significant impairment (Siperstein, Norins, Corbin, & Shriver, 2003).

For example, a national study of youth attitudes found that students tend to perceive peers with ID as being moderately, rather than mildly, impaired (Siperstein, Parker,

Norins Bardon, & Widaman, 2007). Earlier research showed that, when asked to imagine a person with ID, the public tended to picture a person with Down syndrome (Gottlieb & Siperstein, 1976; Siperstein & Gottlieb, 1977). The public also tends to view ID as due to physical causes, permanent in nature (Goodman, 1989; Gottlieb, 1975; Jones et al., 1984), and physically evident (Gottlieb, 1975), which are all common indicators of more significant levels of impairment. These perceptions are evident as young as third grade (Goodman, 1989) and lead to public reactions of hopelessness and rejection (Jones et al., 1984). In addition, multinational research has found that these misperceptions are present to varying degrees all over the world (Siperstein et al., 2003).

Despite the wide spectrum within ID, individuals with more significant levels of impairment have become the default prototype for ID in the public's eye. Consequently, the public tends to underestimate the abilities of many people with ID. When asked about the perceived abilities of an individual with ID, 83% reported that someone with ID could wash and dress themselves, but only half thought they could prepare their own food or handle money (Siperstein et al., 2003). These misconceptions have serious consequences. Despite the importance of early intervention, there is often reluctance to diagnose a child with ID, as parents do not perceive their child's impairment to be significant enough to warrant diagnosis. Considering a hypothetical 4-year-old child, just 51% of parents responded that they would refer the child for special education services if he or she demonstrated a mild level of impairment, compared with 91% for moderately impaired children (Goodman, 1987). When one considers the fact that the large majority of individuals with ID actually function at the upper end of the spectrum, the disconnect between what the public believes and what in reality is true is stark. Consequently, those with ID at the upper end of the spectrum are most likely to be misdiagnosed or not diagnosed at all because others don't *perceive* them as having a disability or because of the stigma historically associated with the label "mental retardation." As a result, these individuals may be frequently not served or underserved—particularly within the context of public education—placing them at risk for a number of negative outcomes in adulthood (Zetlin & Murtaugh, 1990).

Challenges in the American Judicial System

For all of the challenges in classification and diagnosis, widespread misconceptions of capabilities, and general low understanding of functioning variability within ID, those individuals at the upper end of the ID spectrum are the most challenging for the courts. As Baroff (2006) stated, "for judges who must follow state legal statutes that set IQ [at] 70, or 69, as the boundary for (intellectual disability previously known as) mental retardation, the distinction, in capital cases, may truly be one of life or death" (p. 34). In general, individuals with ID are viewed as having a lesser level of culpability than persons without ID (Baroff, 2006). In fact, multinational research has shown that the majority of the public believes that individuals with ID should receive special dispensation in a court of law (Siperstein et al., 2003). However, this can be a difficult issue when one

considers the discrepancy between what the public perceives ID to be and the actual characteristics of a person with ID at the upper end of the spectrum. How might public opinion change in the face of a person who does not fit the perceived schema for ID?

The gravity of the situation is aggravated by the prevalence of offenders with ID. In accordance with the traditional idea of a "six-hour retarded child" (The President's Committee on Mental Retardation, 1969), adults who formerly received special education services frequently "disappear into society in their adult years" (Larson et al., 2001, p. 232). However, the same cognitive impairments remain and, because these individuals may no longer be part of a system of supports, these impairments may resurface through criminal behavior. Subsequently, when suspects with ID are arrested, they are uniquely vulnerable while navigating the court system because they may lack understanding of their own legal rights and the judicial process (e.g., Applebaum 1994); may have difficulty processing instructions, commands, and questions; or may have difficulty remembering the details or sequences of events of a case (Davis, 2009). Thus, identification and proper procedures in the court system are not only extremely complex, due to the ambiguity in borderline cases, but also extremely consequential.

In conclusion, it is quite clear that people with ID represent a diverse group, with those individuals who are at the borderline of being diagnosed or not diagnosed being the most prevalent and being both quantitatively and qualitatively different from those with more significant levels of impairment. History underscores this and points to factors such as etiology and context as driving who is diagnosed as having ID and who is not—not to mention who is misdiagnosed. All of this is compounded by the fact that the public (e.g., employees, service professionals, and even jurors, who represent a cross-section of the public) understand ID to be singular. This "schema" for a person with ID is far from a representation of a person who, albeit has difficulty in abstract thinking and complex problem solving, is capable of being employed, living independently (with supports), and engaging in the community. The public's image on the contrary can be found in an eleventh-grade boy's characterization of a person with ID:

> [People with ID] are not able to comprehend what life really is. They are unable to function as normal (*sic*) people because of brain disease or damage. I know this from viewing them doing their menial tasks and from books I have read. They got that way because of a lack of air during birth, thus their brain damage, or because of freak mutations like too many chromosomes—just one extra will do it. They are outwardly obvious, that is, they have cloudy haircuts, outdated clothes, and cheap eyeglasses. They feel nothing. They haven't the capabilities to understand what they are. (as quoted in Siperstein & Bak, 1980, p. 207)

Beyond empirical support for this overall lack of understanding of the capabilities of people with ID, this perception is pervasive throughout television and film. Judges and jurors potentially, without an understanding and appreciation of the wide spectrum of challenges and capabilities of individuals with ID, may find it difficult to juxtapose and

reconcile their beliefs and perceptions with the reality of an individual with ID at the upper end of the spectrum. *Atkins v. Virginia* (2002) put the history of subgrouping individuals with ID and the conceptual and practical/programmatic issues of defining and identifying individuals with ID squarely in the courtroom. The other chapters in this book address the identification process of ID, such as measurement and assessment considerations and issues related to intellectual functioning and adaptive behavior. While considering all these matters, however, it is critical that the population of individuals at the upper end of the ID spectrum not be viewed as a special or unusual circumstance; rather, it must be remembered that they make up the majority of the population with ID, and, yet, at the same time, are uniquely different from people in the middle and lower ranges of the ID spectrum and from public expectations.

References

American Association on Intellectual and Developmental Disabilities. (2010). FAQ on the AAIDD definition on intellectual disability. Retrieved from http://www.aaidd.org/IntellectualDisabilityBook/content_7473.cfm?navID=366

Applebaum, K. L. (1994). Assessment of criminal-justice-related competences in defendants with mental retardation. *The Journal of Psychiatry and Law, 22*(3), 311–327.

Atkins v. Virginia, 536 U.S. 304, 310 (2002).

Baroff, G. S. (2006). On the 2002 AAMR definition of mental retardation. In H. N. Switzky & S. Greenspan (Eds.), *What is mental retardation? Ideas for an evolving disability in the 21st century* (pp. 29–38). Washington, DC: American Association on Intellectual and Developmental Disabilities.

Bond, R. J., & Hurst, J. (2009). How adults with learning disabilities view living independently. *British Journal of Learning Disabilities, 38*(4), 286–292. doi: 10.1111/j.1468-3156.2009.00604.x

Braddock, D., & Parish, S. L. (2002). An institutional history of disability. In Albrecht, G. L., Seelman, K. D., & Bury, M. (Eds.), *Handbook of disability studies* (pp. 11–68). Washington, DC: American Association on Mental Retardation.

Browman, S., Nichols, P. L., Shaughnessy, P., & Kennedy, W. (1987). *Retardation in young children: A developmental study of cognitive deficit.* Hillsdale, NJ: Erlbaum.

Burack, J. A. (1990). Differentiating mental retardation: The two-group approach and beyond. In R. M. Hodapp, J. A. Burack, & E. Zigler (Eds.), *Issues in the developmental approach to mental retardation* (pp. 27–48). New York, NY: Cambridge University Press.

Burack, J. A., Hodapp, R. M., & Zigler, E. (1988). Issues in the classification of mental retardation: Differentiating among organic etiologies. *Journal of Child Psychology and Psychiatry, 29*(6), 765–779.

Cardona, F. A. (1994). Milestones in the history of mental retardation. *The Journal of the South Carolina Medical Association, 90*(6), 285–288.

Connolly, K. J., & Bruner, J. S. (1974). Competence: Its nature and nurture. In K. J. Connolly & J. S. Bruner (Eds.), *The growth of competence* (pp. 3–7). London, England: Academic Press.

Davis, L. A. (2009). People with intellectual disabilities in the criminal justice system: Victims and suspects. *The Arc Q & A.* Retrieved from http://www.thearc.org/page.aspx?pid=2458

Davies, D. K., Stock, S. E., Holloway, S., & Wehmeyer, M. L. (2010). Evaluating a GPS-based transportation device to support independent bus travel by people with intellectual disability. *Intellectual and Developmental Disabilities, 48*(6), 454–463. doi: http://dx.doi.org/10.1352/1934-9556-48.6.454

Davies, D. K., Stock, S. E., & Wehmeyer, M. L. (2003). Application of computer simulation to teach ATM access to individuals with intellectual disabilities. *Education and Training in Developmental Disabilities, 38*(4), 451–456.

Dingman, H. F., & Tarjan, G. (1960). Mental retardation and the normal distribution curve. *American Journal of Mental Deficiency, 64*, 991–994.

Dixon, R. M., & Reddacliff, C. A. (2001). Family contribution to the vocational lives of vocationally competent young adults with intellectual disabilities. *International Journal of Disability, Development, and Education, 48*(2), 193–206. doi: http://dx.doi.org/10.1080/10349120120053667

Dusseljee, J., Rijken, P., Cardol, M., Curfs, L., & Groenewegen, P. (2011). Participation in daytime activities among people with mild or moderate intellectual disability. *Journal of Intellectual Disability Research, 55*(1), 4–18. doi: 10.1111/j.1365-2788.2010.01342.x

Emerson, E. (2007). Poverty and people with intellectual disabilities. *Mental Retardation & Developmental Disabilities Research Reviews, 13*(2), 107–113. doi: 0.1002/mrdd.20144

Goodman, J. F. (1987). Reluctance to refer the mildly retarded child: Implications for labeling. *Early Child Development and Care, 29*, 331–341. doi: 10.1002/mrdd.20144

Goodman, J. F. (1989). Does retardation mean dumb? Children's perceptions of the nature, cause, and course of mental retardation. *Journal of Special Education, 23*(3), 313–329.

Gottlieb, J. (1975). Public, peer, and professional attitudes toward mentally retarded persons. In M. J. Begab & S. A. Richardson (Eds.), *The mentally retarded and society: A social science perspective* (pp. 99–125). Baltimore: University Park Press.

Gottlieb, J., & Siperstein, G. (1976). Attitudes toward mentally retarded persons: Effects of attitude referent specificity. *American Journal of Mental Deficiency, 80*, 376–381.

Greenspan, S., & Granfield, J. M. (1992). Reconsidering the construct of mental retardation: Implications of a model of social competence. *American Journal of Mental Retardation, 96*(4), 442–453.

Greenspan, S., & Switzky, H. N. (2006). Forty-four years of AAMR manuals. In H. N. Switzky & S. Greenspan (Eds.), *What is mental retardation? Ideas for an evolving disability in the 21st century* (pp. 3–28). Washington, DC: American Association on Intellectual and Developmental Disabilities.

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

Harada, C. M., Siperstein, G. N., Parker, R. C., & Lenox, D. (2011). Promoting social inclusion for people with intellectual disabilities through sport: Special Olympics International, global sport initiatives and strategies. *Sport in Society, 14*(9), 1131–1138. doi: 10.1080/17430437.2011.614770

Heber, R. (1959). A manual on terminology & classification in mental retardation [Monograph Supplement]. *American Journal of Mental Deficiency, 64*(2).

Heber, R. (1961). Modifications in the manual on terminology and classification in mental retardation. *American Journal of Mental Deficiency, 65*(4), 499–500.

Hemphill, L., & Siperstein, G. (1990). Conversational competence and peer response to mildly retarded children. *Journal of Educational Psychology, 82*(1), 128–134. doi: 10.1037/0022-0663.82.1.128

Jahoda, A., Banks, P., Dagnan, D., Kemp, J., Kerr, W., & Williams, V. (2009). Starting a new job: The social and emotional experience of people with intellectual disabilities. *Journal of Applied Research in Intellectual Disabilities, 22*(5), 421–425. doi: 10.1111/j.1468-3148.2009.00497.x

Jones, E. E., Farina, A., Hastorf., A., Marjus, H., Miller, O., & Scott, R. (1984). *Social stigma: The psychology of marked relationships*. New York, NY: W. H. Freeman.

Larkin, P., Jahoda, A., & MacMahon, K. (2012). Interpersonal sources of conflict in young people with and without mild to moderate intellectual disabilities at transition from adolescence to adulthood. *Journal of Applied Research in Intellectual Disabilities, 25*(1), 29–38. doi: 10.1111/j.1468-3148.2011.00652.x

Larson, S. A., Lakini, K. C., Anderson, L., Kwak, N., Lee, J. H., & Anderson, D. (2001). Prevalence of mental retardation and developmental disabilities: Estimates from the 1994/1995 National Health Interview Survey Disability Supplements. *American Journal on Mental Retardation, 106*, 231–252.

Leffert, J. S. & Siperstein, G. N. (2002). Social cognition: A key to understanding adaptive behavior in individuals with mental retardation. In L. M. Glidden (Vol. Ed.), *International review of research in mental retardation* (vol. 25, pp. 135–181). San Diego, CA: Academic Press.

Leffert, J. S., Siperstein, G. N., & Widaman, K. F. (2010). Social perception in children with intellectual disabilities: The interpretation of benign and hostile intentions. *Journal of Intellectual Disability Research, 54*(2), 168–180. doi:10.1111/j.1365-2788.2009.01240.x

Lehotkay, R., Varisco, S., Deriaz, N., Douibi, A., & Carminati, G. G. (2009). Intellectual disability and psychiatric disorder: More than a dual diagnosis. *Schweizer Archiv für Neurologie und Psychiatrie, 160*(3), 105–115.

Leland, H. (1969). The relationship between "intelligence" and mental retardation. *American Journal of Mental Deficiency, 73*, 533–535.

Lippold, T., & Burns, J. (2009). Social support and intellectual disabilities: A comparison between social networks of adults with intellectual disability and those with physical disability. *Journal of Intellectual Disability Research, 53*(5), 463–473. doi: 10.1111/j.1365-2788.2009.01170.x

Luckasson, R., Borthwick-Duffy, S., Buntinx, W. H. E., Coulter, D. L., Craig, E. M., Reeve, A., Schalock, R. L., Snell, M. E., Spitalnik, D. M., Spreat, S., & Tassé, M. J. (2002). *Mental retardation: Definition, classification, and systems of supports* (10th ed.). Washington, DC: American Association on Mental Retardation.

Luckasson, R., Coulter, D. L., Polloway, E. A., Reiss, S., Schalock, R. L., Snell, M. E., . . . Stark, J. A. (1992). Mental retardation: Definition, classification, and systems of supports (9th ed.). Washington, DC: American Association on Mental Retardation.

MacMillan, D. L., Gresham, F. M., & Siperstein, G. S. (1993). Conceptual and psychometric concerns about the 1992 AAMR definition of mental retardation. *American Journal on Mental Retardation, 98*, 325–355.

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

MacMillan, D. L., & Reschly, D. J. (1997). Issues of definition and classification. In W. E. MacLean, Jr. (Ed.), *Ellis' handbook of mental deficiency, psychological theory, and research* (3rd ed., pp. 47–74). Mahwah, NJ: Erlbaum.

MacMillan, D. L., & Siperstein, G. N. (2002). Learning disabilities as operationally defined by schools. In R. Bradley, L. Danielson & D. Hallahan (Eds.), *Identification of learning disabilities: Research to practice*. Mahwah, NJ: Lawrence Erlbaum Assoc.

MacMillan, D. L., Siperstein, G. N., & Gresham, F. M. (1996). A challenge to the viability of mild mental retardation as a diagnostic category. *Exceptional Children, 62*, 356–371.

MacMillan, D. L., Siperstein, G. N., & Leffert, J. S. (2006). Children with mild mental retardation: A challenge for classification practices–revised. In H. N. Switzky & S. Greenspan (Eds.), *What is mental retardation? Ideas for an evolving disability in the 21st century*. (pp. 197–220). Washington, DC: American Association on Mental Retardation.

Matson, J. L., & Shoemaker, M. (2009). Intellectual disability and its relationship to autism spectrum disorders. *Research in Developmental Disabilities: A Multidisciplinary Journal, 30*(6), 1107–1114. doi: 10.1016/j.ridd.2009.06.003

Mesibov, G. B. (1974). *Attributions of responsibility: A cognitive interpretation.* Brandeis University, Waltham: MA.

Papay, C. K., & Bambara, L. M. (2011). Postsecondary education for transition-age students with intellectual and other developmental disabilities: A national survey. *Education and Training in Autism and Developmental Disabilities, 46*(1), 78–93.

Petersilia, J. (2000). *Doing justice? Criminal offenders with developmental disabilities.* Berkeley, CA: California Policy Resource Center, University of California Berkeley.

Pinborough-Zimmerman, J., Satterfield, R., & Miller, J. (2007). Communication disorders: Prevalence and comorbid intellectual disability, autism, and emotional/behavioral disorders. *American Journal of Speech-Language Pathology, 16*(4), 359–367. doi:10.1044/1058-0360(2007/039)

President's Committee on Mental Retardation. (1969). *The black six-hour retarded child: A report on a conference on problems of education of children in the inner city (Warrenton, Virginia, August 10–12, 1969).* Washington, DC: Author.

Reschly, D. J., & Jipson, F. J. (1976). Ethnicity, geographic locale, age, sex, and urban-rural residence as variables in mild retardation. *American Journal of Mental Deficiency, 81*(2), 154–161.

Rose, E., Bramham, J., Young, S., Paliokostas, E., & Xenitidis, K. (2008). Neuropsychological characteristics of adults with comorbid ADHD and borderline/mild intellectual disability. *Research in Developmental Disabilities, 30*(3), 496–502. doi: 10.1016/j.ridd.2008.07.009

Schalock, R. L., Borthwick-Duffy, S. A., Bradley, V. J., Buntinx, W. H. E., Coulter, D. L., Craig, E. M., Gomez, S. C., Lachapelle, Y., Luckasson, R., Reeve, A., Shogren, K. A., Snell, M. E., Spreat, S., Tassé, M. J., Thompson, J. R., Verdugo-Alonso, M. A., Wehmeyer, M. L., & Yeager, M. H. (2010). *Intellectual disability: Definition, classification, and systems of supports (11th ed.).* Washington, DC: American Association on Intellectual and Developmental Disabilities.

Scheerenberger, R. C. (1983). *A history of mental retardation.* Baltimore, MD: Brookes.

Shogren, K. A., & Rye, M. S. (2005). Religion and individuals with intellectual disabilities: An exploratory study of self-reported perspectives. *Journal of Religion, Disability, & Health, 9*(1), 29–53. doi:10.1300/J095v09n01_03

Siebelink, E. M., de Jong, M. D. T., Taal, E., & Roelvink, L. (2006). Sexuality and people with intellectual disabilities: Assessment of knowledge, attitudes, experiences, and needs. *Mental Retardation, 44*(4), 283–294. doi: 10.1352/0047-6765(2006)44[283:SAPWID]2.0.CO;2

Siperstein, G., & Bak, J. J. (1980). Students' and teachers' perceptions of the mentally retarded child. In J. Gottlieb (Ed.), *Educating mentally retarded persons in the mainstream* (pp. 207–230). Baltimore, MD: University Park.

Siperstein, G., & Gottlieb, J. (1977). Physical stigma and academic performance as factors affecting children's first impressions of handicapped peers. *American Journal of Mental Deficiency, 81*, 455–462.

Siperstein, G. S., Norins, J., Corbin, S. B., & Shriver, T. (2003). *Multinational study of attitudes toward individuals with intellectual disabilities* [Special report]. Washington, DC: Special Olympics.

Siperstein, G., Parker, R., Norins Bardon, J., & Widaman, K. (2007). A national study of youth attitudes toward the inclusion of students with intellectual disabilities. *Exceptional Children, 73*(4), 435–455. doi: 10.1177/001440290707300403

Suto, W., Clare, I., Holland, A., & Watson, P. (2005). Capacity to make financial decisions among people with mild intellectual disabilities. *Journal of Intellectual Disability Research, 49*(3), 199–209. doi: 10.1111/j.1365-2788.2005.00635.x

Taber-Doughty, T., Bouck, E. C., Tom, K., Jasper, A. D., Flanagan, S. M., & Bassette, L. (2011). Video modeling and prompting: A comparison of two strategies for teaching cooking skills to students with mild intellectual disabilities. *Education and Training in Autism and Developmental Disabilities, 46*(4), 499–513.

Trembath, D., Balandin, S., Stancliffe, R. J., & Togher, L. (2010). Employment and volunteering for adults with intellectual disability. *Journal of Policy and Practice in Intellectual Disabilities, 7*(4), 235–238. doi: 10.1111/j.1741-1130.2010.00271.x

Weber, E. W. (1962). *Educable and trainable mentally retarded children.* Springfield, IL: Thomas.

Wehmeyer, M. L., Palmer S. B., Smith, S. J., Parent, W., Davies, D. K., & Stock, S. (2006). Technology use by people with intellectual and developmental disabilities to support employment activities: A single-subject design meta-analysis. *Journal of Vocational Rehabilitation, 24*, 81–86.

Wolraich, M. L., & Siperstein, G. N. (1986).Physicians' and other professionals' expectations and prognoses for mentally retarded individuals. *American Journal of Mental Deficiency, 91*(3), 244–249.

Wolraich, M. L., Siperstein, G. N., & O'Keefe, P. (1987). Pediatricians' perceptions of mentally retarded individuals. *Pediatrics, 80*(5), 643–649.

Wright, T., & Wolery, M. (2011). The effects of instructional interventions related to street crossing and individuals with disabilities. *Research in Developmental Disabilities, 32*(5), 1455–1463.

Zetlin,A., & Murtaugh, M. (1990). Whatever happened to those with borderline IQs? *American Journal of Mental Retardation, 94*(5), 463–469.

Zigler, E., Balla, D., & Hodapp, R. (1984). On the definition and classification of mental retardation. *American Journal of Mental Deficiency, 89*(3), 215–230.

10 Norm Obsolescence: The Flynn Effect

Kevin S. McGrew

Nature of the Problem

A person's IQ test score is based on the comparison of the person's tested performance to an age-appropriate norm reference group. The *norms* for an IQ test are developed to represent the snapshot of the general U.S. population (at each age level the test covers) at the time the norm or standardization data are collected (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, NCME], 1999). (VandenBos, 2007, defines a norm as "a standard or range of values that represents the typical performance of a group or of an individual [of a certain age, for example] against which comparisons can be made" [p. 631]). The person's test performance is compared to this standard reference group. For example, the WISC-R IQ test was published in 1974 and the WISC-R norm data was gathered on children ages 6 through 16 from 1971 through 1973 (Wechsler, 1974). (1972 is thus considered the official date of the WISC-R norm/standardization sample.) Thus, a child who is 7 years, 2 months old who was administered the WISC-R in 1974 would have the calculation of his or her IQ test score based on a comparison to the performance of children from ages 7 years, 0 months through 7 years, 3 months in the year 1972. (The WISC-R norm tables are provided in 3 month intervals within each year of age.) If the WISC-R was administered to a child of the same age (7 years, 2 months) in 1984, rather than being compared to other children of the same age in 1984, this child's performance would still be evaluated against similarly aged children from 1972. This second comparison results in a test-date/test-norm *mismatch* of 12 years (1984 – 1972 = 12). As explained next, comparing an individual's performance on an IQ test with outdated test norms results in *a comparison to a historical reference group from the past—not the person's contemporary peers.* This *norm obsolescence* problem is more commonly referred to as the *Flynn effect* (Flynn, 1984,

1985, 2000, 2006, 2007a, 2009). The Flynn effect produces inflated and inaccurate IQ test scores.

In simple terms, psychologists and psychological measurement experts typically describe the Flynn effect as the result of a "softening" of IQ tests norms with the passage of time. That is, individuals tested today on an IQ test normed many years earlier will obtain artificially inflated IQ test scores, because the older test norms reflect a level of overall performance that is lower than that of individuals in contemporary society. This is one of the primary reasons why authors and publishers of IQ tests make every effort to periodically provide "freshened" norms by collecting new nationally representative sample data for IQ test batteries. The professional consensus among test developers is that the "shelf life" of an IQ test's norms is approximately 10 years. According to Weiss (2010), Vice President of Pearson Clinical Assessment, the company and division that develops and publishes the various Wechsler IQ batteries, "there is no definition of when a test becomes obsolete. When asked privately, most Flynn effect researchers have 10 years in mind" (p. 492). If new norms are not provided, individuals tested using IQ tests with outdated norms will typically obtain inflated and inaccurate IQ test scores.

The Flynn effect recognizes that the normal curve distribution of intelligence shifts upward over time. Thus, the same raw score performance on an IQ test, when compared to outdated norms, will produce a markedly different IQ score when it is compared to updated norms based on a contemporary sample of abilities for a person of the same age. The person's tested performance (i.e., the number of correct responses across all parts of the IQ test) does *not* change, but the person's relative standing in the distribution of IQ scores across the population *does* change as a function of which norm reference group his or her performance is compared against. The same performance that is considered average in the contemporary norm sample, yielding an IQ test score of 100 in the distribution, will result in a higher IQ test score when using older norms (Schalock, 2012).

As a result of the Flynn effect, it is possible that one or more IQ test scores reported for an individual being considered for a diagnosis of intellectual disability (ID) may be inaccurate and inflated estimates. Given the high-stakes nature of *Atkins*, ID cases and their tendency to artificially focus on specific "bright line" cutoff scores, the impact of the Flynn effect must be recognized and an adjustment to the inflated scores is recommended.

Summary of Related Research

Origins of the Flynn Effect

Probably the first widely recognized scholarly report of IQ norm obsolescence was published by Lynn in 1983. Reflecting Lynn's early writings, some intelligence scholars refer to IQ norm obsolescence as the *Lynn-Flynn effect* (Woodley, 2012a). Recently, Lynn

(2013) provided evidence that 24 studies, the first being Runquist (1936), reported on the phenomenon of norm obsolescence before the "effect was rediscovered by Flynn" (1984). Lynn (2013) argued that the proper designation of IQ test norm obsolescence should be the "Runquist effect." Although Lynn (2013) provided a compelling argument (based on the customary practices in the history of science for naming phenomena), the term *Flynn effect* is used here given its prominent and frequent use in intelligence research and *Atkins* court cases.

Seventeen years prior to the 2002 *Atkins* decision, Flynn (1985) published an article in the *American Journal on Mental Deficiency* (now the *American Journal on Intellectual and Developmental Disabilities*). This article, titled "Wechsler Intelligence Tests: Do We Really Have Criterion of Mental Retardation?" first raised the issue of a possible "adjustment" in the context of an ID diagnosis. In hindsight, Flynn's 1985 article was the "canary in the coal mine" in that it first demonstrated that the Flynn effect may have a significant impact on the proportion of the population of individuals that would be identified as ID. At that time, Flynn proposed a form of adjusting for the softening of tests norms, although it was in a slightly different form than the current recommended Flynn effect adjustment procedure.

Flynn (1985) proposed that to account for the softening of test norms, an IQ test score of 70 on a "reference" IQ test (i.e., WAIS-R) would be set in as the *absolute criterion for mental retardation* (that is, on the intellectual functioning prong of the definition). Then, to account for norm obsolescence, each time a new IQ test was published there would be a lowering of the MR cutting line. Flynn's 1985 idea was that whenever a new IQ test was published, it would be given together with the established reference IQ test (e.g., WAIS-R) and the average mean IQ test score difference between the new test and the reference test would be used to "derive a new score equivalent to the old cutting line" (p. 243). Although different from what is now considered the standard Flynn effect adjustment approach (i.e., subtracting 3 IQ test score points from an individual's total IQ test score for every 10 years for which the test was administered to a person who was normed prior to the date of individual's testing), conceptually Flynn's 1985 proposal accomplished the same goal as the currently employed Flynn effect adjustment procedure.

Fifteen years later, and still 2 years prior to the *Atkins* decision, Flynn (2000) again sounded the alarm regarding the implication of norm obsolescence related to the diagnosis and classification of mental retardation:

> It is certain that over the past 50 years, literally millions of Americans evaded the label of mentally retarded designed for them by the test manuals. Whether this was good or bad depends on what one thinks of the label. Some will say millions avoided stigma. Others will say that millions missed out on needed assistance and classroom teachers were left unaided to cope with pupils for whom aid was needed (p. 197).

The potential impact of the Flynn effect on other diagnoses was also reported in 2001 and 2003. Truscott and colleagues (Sanborn, Truscott, Phelps, & McDougal, 2003; Truscott & Frank, 2001) reported on the impact of the Flynn effect on learning disability (LD) identification, not identification of individuals with ID. Although these authors did not offer or endorse any IQ test score adjustment procedure, these researchers concluded that

> A critical finding of this study is that the FE probably contributes to misdiagnosis of LD. If this research is combined with previous reports that academic achievement may be unaffected by the FE (Neisser, 1998) it strongly suggests that, over the life of a test version, IQ-achievement discrepancies, the most salient LD criterion, are exaggerated. One potential result of such an exaggeration of IQ-achievement discrepancies would be that, as test norms aged, fewer students would score in the mentally retarded range (Flynn, 2000) and more students would qualify for LD based on inflated severe discrepancies (p. 300).

In conclusion, the recognition of the impact of *norm obsolescence* (i.e., the Flynn effect) on IQ test scores, and more importantly, the potential for misdiagnosis of ID and other conditions (e.g., LD), has been recognized and documented as early as the 1980s. It continued to be discussed prior to and after the 2002 ID-related *Atkins* decision by researchers and professionals who did not anticipate nor were influenced by the 2002 *Atkins* decision. For obvious reasons (i.e., the life-or-death implications of the *Atkins* decision), there has been increased interest in the Flynn effect adjustment procedure since the *Atkins* decision. The facts indicate that the recognition of the impact of norm obsolescence on IQ test scores (and the idea of a norm obsolescence IQ test score adjustment) was established prior to the *Atkins v Virginia* (2002) U.S. Supreme Court decision.

Scientific Basis of the Flynn Effect

There is a scientific and professional consensus that the Flynn effect is a scientific fact. A complete reading of the extant Flynn effect research literature leads to the conclusion that, despite debates regarding the causes of the Flynn effect, differences in the rate of Flynn effect change in different countries. Whether the Flynn effect has started to plateau in Scandinavian countries or whether the Flynn effect differs by different levels of intelligence and different methodological issues in various studies, *the consensus of the relevant scientific community is that the Flynn effect is real* (Cunningham & Tassé, 2010; Fletcher, Stuebing & Hughes, 2010; Flynn, 2009; Greenspan, 2006, 2007; Gresham & Reschly, 2011; Kaufman, 2010a, 2010b; McGrew, 2010; Rodgers, 1999; Trahan, Stuebing, Fletcher, & Hiscock 2014; Weiss, 2010; Zhou, Zhu, & Weiss, 2010). The robustness of this conclusion may best be represented by Rogers' (1999) statement where, after raising valid methodological issues regarding various statistical analysis and conclusions across Flynn effect studies, that even with a "healthy dose of

skepticism, the effect rises above purely methodological interpretation, and appears to have substantive import" (p. 354).

The research literature regarding the Flynn effect is extensive. Trahan et al. (2014) found over 4,000 articles in their comprehensive literature review. (Most all norm obsolescence references and articles can be found at the regularly updated *Flynn Effect Archive Project* [http://www.atkinsmrdeathpenalty.com/2010/01/atkins-mrid-capital-punishment-flynn_11.html]. As of 2014, this archive includes approximately 190 publications.) A thorough treatment of all this research is beyond the scope of the current chapter. Fortunately, key contemporary Flynn effect issues bearing on an ID diagnosis in the *Atkins* context were covered in a special 2010 issue of the *Journal of Psychoeducational Assessment* (*JPA*). A variety of invited scholars confirmed the scientific consensus regarding the validity of the Flynn effect. For example, Dr. Alan Kaufman (2010a), arguably the most prominent scholar on intelligence testing and interpretation of the various Wechsler IQ tests, stated that

> The Flynn effect (FE) is well known: Children and adults score higher on IQ tests now than they did in previous generations (Flynn, 1984, 2007, 2009b). The rate of increase in the United States has apparently remained a fairly constant 3 points per decade since the 1930s (p. 382).

The consensus of almost all authors who contributed to the *JPA* Flynn effect issue (Fletcher et al., 2010; Flynn, 2010; Hagan, Drogin, & Guilmette, 2010a; Kaufman, 2010a, 2010b; Kaufman & Weiss, 2010; McGrew, 2010; Reynolds, Niland, Wright, & Rosenn, 2010; Sternberg, 2010; Weiss, 2010; Zhou et al. 2010) was that IQ test norm obsolescence (i.e., the Flynn effect) is an established scientific fact. The following select quotes from recent peer-reviewed articles capture the essence of the convergence of opinion regarding the validity of the Flynn effect.

> The Flynn effect (FE) is real. The FE has been shown to be nearly 3 points per decade on average across a large number of studies, countries, and tests (Weiss, 2010, p. 491).

> The point is that a person tested on an outdated test will earn spuriously high scores as each year goes by, and the amount of the spuriousness amounts to about 3 points per decade for Americans (Kaufman, 2010b, p. 503).

> The FE, whatever its cause, is as real as virtually any effect can be in the social sciences. Studies have observed an increase of 0.3 points per year in average IQs; thus, for a test score to reflect accurately the examinee's intelligence, 0.3 points must be subtracted for each year since the test was standardized (Reynolds et al., 2010, p. 478).

> The Flynn effect is a well-established psychometric fact documenting substantial increases in measured intelligence test performance over time (Gresham & Reschly, 2011, p. 131).

Since the publication of the 2010 special *JPA* Flynn effect issue, many additional Flynn effect research and commentary articles have appeared (e.g., Battarjee, Khaleefa, Ali, & Lynn, 2013; Baxendale, 2010; Cunningham & Tassé, 2010; Hagan, Drogin, & Guilmette, 2010b; Kanaya & Ceci, 2011, 2012; Lynn, 2013; Nijenhuis, 2013; Nijenhuis, Cho, Murphy, & Lee, 2012; Nijenhuis, Murphy, & van Eeden, 2011; Nijenhuis & van der Flier, 2013; Pietschnig, Voracek, & Formann, 2011; Nijman, Scheirs, Prinsen, Abbink, & Blok, 2010; Rindermann, Schott, & Baumeister, 2013; Rönnlund, Carlstedt, Blomstedt, Nilsson, & Weinehall, 2013; Skirbekk, Stonawski, Bonsang, & Staudinger, 2013; Trahan et al., 2014; Wai & Putallaz, 2011; Woodley, 2011, 2012a, 2012b; Young, 2012). The continued flow of the Flynn effect related to peer-reviewed articles confirms the consensus that the Flynn effect is a scientifically important and studied phenomenon among intelligence scholars.

Adjusting IQ Test Scores for the Flynn Effect in Atkins *Cases Is Best Practice*

Not only is there a scientific consensus that the Flynn effect is a valid and real phenomenon, there is also a consensus that individually obtained IQ test scores derived from tests with outdated norms must be adjusted to account for the Flynn effect, particularly in *Atkins* cases. (The use of a Flynn effect correction in clinical settings is less of an issue given that psychologists in such settings typically have more leeway to interpret scores as ranges, invoke clinical judgment, and incorporate information regarding measurement error in interpretation of the scores when making a diagnosis. In contrast, certain high stakes settings [e.g., *Atkins* cases; eligibility for Social Security Disability benefits] may have strict point-specific cut-scores [i.e., "bright line" criteria] where examiners, or the recipients of the scores [e.g., the courts], do not allow for such clinical interpretation. Thus, the Flynn effect adjustment is more relevant, appropriate, and primarily discussed in literature and law dealing with this type of high stakes IQ testing.) The most prominent and relevant professional consensus-based guidelines for ID diagnosis (Schalock et al., 2007, 2010, and 2012) support a Flynn effect adjustment for scores based on obsolete IQ test norms. *Intellectual Disability: Definition, Classification, and Systems of Supports* (11th ed.; Schalock et al., 2010), based on an expert-consensus process, provides a written guideline that endorses the appropriateness of the Flynn effect adjustment in the diagnosis of ID. (The 11th edition was created using a group-based consensus process conducted by the AAIDD Ad Hoc Committee on Terminology and Classification [Schalock et al., 2010]). AAIDD recommends that psychologists use the most recent versions of IQ tests and, if scores are reported from an IQ test with outdated norms, a correction for the age of norms is warranted (Schalock et al., 2007). The 11th edition states

> As discussed in the *User's Guide* (Schalock et al., 2007) that accompanies the 10th edition of this *Manual,* best practices require recognition of a potential Flynn effect when older editions of an intelligence test (with corresponding older norms) are used in the assessment or interpretation of an IQ score. (p. 37)

As suggested in the *User's Guide to Mental Retardation: Definition, Classification, and Systems of Supports* (Schalock, 2007, pp. 20–21),

> 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 of tests 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. (p. 37)

The AAIDD's more recent *User's Guide to Intellectual Disability: Definition, Classification, and Systems of Supports* (Schalock et al., 2012) states

> The *Flynn effect* refers to the increase in IQ scores over time (i.e., about 0.30 points per year). The Flynn effect affects any interpretation of IQ scores based on outdated norms. Both the 11th edition of the manual and this *User's Guide* recommend that in cases in which a test with aging norms is used as part of a diagnosis of ID, a corrected Full Scale IQ upward of 3 points per decade for age of norms is warranted. (p. 23)

A consensus among the professional and scientific community of intelligence and ID scholars has emerged. This consensus is that given the high-stakes nature of *Atkins* ID cases and their tendency to artificially focus on specific "bright line" cutoff scores, *a Flynn effect correction to a person's scores in this setting is now considered best or standard practice.* This conclusion is supported by a significant number of scholars and researchers in the areas of intelligence and ID (Cunningham & Tassé, 2010; Fletcher et al., 2010; Flynn, 2006, 2007b; Flynn & Widaman, 2008; Greenspan, 2006, 2007; Gresham & Reschly, 2011; Kaufman, 2010b; McVaugh & Cunningham, 2009; Reynolds et al., 2010; Schalock , 2007; Schalock, 2012). One example of this support is the statement of Reynolds et al. (2010) that "as a generally accepted scientific theory that could potentially make the difference between a constitutional and unconstitutional execution, the Flynn effect must be applied in the legal context" (p. 480). Reynolds et al. (2010) go as far as to state that "the failure to apply the Flynn correction as we have described it is tantamount to malpractice. No one's life should depend on when an IQ test was normed" (p. 480).

A minority of scholars have offered a different approach to the issue of correcting IQ test scores due to the Flynn effect. Weiss (2010), while acknowledging the scientific validity of the Flynn effect, advocates that experts should simply inform the fact finder of what the research shows and the trier-of-fact should evaluate and decide if and how to apply it when interpreting individual scores. Hagan et al. (2010b) also agree with the need to consider the Flynn effect in capital cases but their disagreement "lies in how psychologists should convey IQ scores in light of the observation that mean scores drift over time" (p. 420). It is important to note that the more conservative positions of Weiss (2010) and Hagan et al. (2010a, 2010b) represent a minority position in the professional literature. More importantly, they do not argue against the scientific validity of the Flynn

effect or even the need to consider the effect in *Atkins* cases. Rather, their difference of opinion with the majority is only as to whether a specified score adjustment should be made to the original score or whether testifying experts should instead address the Flynn effect in narrative form.

Recently, legal scholars have also supported the application of the Flynn effect correction in *Atkins* cases. Young's (2012) recent law review article ("A More Intelligent and Just *Atkins:* Adjusting for the Flynn Effect in Capital Determinations of Mental Retardation or Intellectual Disability") concluded that

> adjusting for the Flynn effect reflects a practice consistent with both *Atkins* and the known world of IQ measurements. While a freakish strike of lightning is difficult to avoid, the potentially deadly and unconstitutional consequences of refusing to account for the Flynn effect are wholly preventable. Thus, for the intelligent and just enforcement of *Atkins*, courts and juries should adjust IQ score from outdated tests for the Flynn effect. (p. 663)

What Is the Correct Flynn Effect Adjustment for Norm Obsolescence?

The AAIDDs' *User's Guide* (Schalock, 2012) recommends a Flynn effect correction of 3 points per decade (0.3 points per year). The 3 points per decade rule-of-thumb is consistent with the previously cited comments of Kaufman (2010a, 2010b) and Weiss (2010), and is also consistent with the recommendation of most scholars in the areas of intelligence and ID (e.g., Fletcher et al., 2010; Gresham & Reschly, 2011; Trahan, et al., 2014; Widaman, 2007).

The 3 points per decade rule-of-thumb is based primarily on Flynn's (2009) seminal article where he synthesized the results of 14 estimates of IQ test score gains over time. Flynn reported an average IQ test score change, across the 14 studies, of 0.311 points per year. An average mean score of 0.299 points was reported for the Wechsler comparisons only. Flynn concluded that "the evidence suggests that a rate of 0.30 is about right, and varying it from case to case lacks any rationale" (p. 104).

More recently Fletcher et al. (2010) applied more precise quantitative meta-analytic procedures to Flynn's (2009) data and reported a weighted mean of 2.80 points per decade. After removing two outlier studies, the weighted mean per decade was 2.96. These researchers concluded that "the level of precision we reported of a mean of about 3 and a *standard error of the mean* (SEM) of about 1 supports the correction and is consistent with the Flynn correction of 3 points per decade" (p. 472). In the most comprehensive meta-analysis research synthesis of 285 studies, Trahan et al. (2014) found that for modern intelligence tests the Flynn effect size was a similar 2.93 points per decade. These researchers concluded that their "findings are consistent with previous research and with the argument that it is feasible and advisable to correct IQ scores for the Flynn effect in high-stakes decisions" (p. 22).

The best available research syntheses consistently converge on a Flynn effect rule-of-thumb of 3 IQ test score points per decade (of IQ test norm obsolescence). Although

scientific journals may report Flynn effect results to the second decimal place (e.g., 3.11 per decade or 0.311 per year), the psychometrics of IQ testing and research cannot partition human behavior with such precision. As noted by Widaman (2007), much of the variation between scores from different Flynn effect studies is due to sampling and measurement error. Using Flynn effect adjustment formulae that use numbers to the second decimal place would be akin to slicing butter with a laser beam. Consequently, the current best estimate of IQ norm obsolescence, and the recommended Flynn effect adjustment, is 3 IQ points per decade, or 0.3 points per year.

Researching the Flynn Effect "Black Box": Implications for Practice

Recently a significant portion of Flynn effect research has shifted from a focus on the secular changes in the global IQ test scores over time to changes on more specific intellectual abilities, possible differential effects by level of intelligence, and a search for the cause of the Flynn effect (Kaufman, 2010a). Zhou et al. (2010) characterized this shift to a focus on the "black box" of the Flynn effect.

The cause of the Flynn effect. In the context of the special articles in the 2010 *JPA* Flynn effect issue, Weiss (2010) stated that "Except for Flynn, there is general agreement . . . that we know precious little about the causes of the effect" (p. 487). Explanations and theories have touched on such causative variables as genetics, environmental factors (e.g., nutrition, education, improved public health, increased use of computer games), ethnicity, and different societal risks and benefits associated with different generations (Kaufman & Weiss, 2010; Weiss, 2010). Flynn (2007a), in his book *What Is Intelligence? Beyond the Flynn Effect*, suggests that the effect that bears his name is due to systematic shift in societies from concrete to abstract scientific thinking. Confounding the search for the cause(s) of the Flynn effect has been idiosyncratic and armchair-based speculations (Weiss, 2010).

In the current context, knowing that the Flynn effect exists trumps a lack of consensus regarding causation. The impact of norm obsolescence on IQ test scores is real and the professional consensus is that it should be accounted for in *Atkins* ID determination. Understanding the "why" of the Flynn effect is beyond the scope of the current chapter and is not necessary for recognizing the scientifically and professionally based consensus that IQ test scores suffering from norm obsolescence need to be adjusted in *Atkins* cases. As stated by Kaufman (2010b), "The Flynn effect is a fact, even if its cause is elusive, and it must be considered carefully when making high stakes decisions such as the death penalty" (p. 503).

Differential Flynn effects by specific intellectual abilities. The foundation of Flynn's (2007a) theoretical explanation of the Flynn effect is based primarily on the interpretation of differential rates of score changes as a function of different specific intellectual abilities (e.g., smaller gains on verbal and crystallized ability tasks and larger changes on visual-spatial and abstract fluid reasoning tasks—not a singular focus on the global IQ test score). If differential specific ability Flynn effects are eventually found to be valid, the potential implication is that different Flynn effect adjustments

may be recommended for different composite or cluster "part" scores in IQ tests, and not just the global IQ score. This would introduce a new layer of complexity in the interpretation of IQ test scores (and part scores) in *Atkins* cases.

Although the recent methodologically sophisticated attempt by Zhou et al. (2010) to examine differential ability Flynn effects within the Wechsler tests represents an important step forward in this area of inquiry, their research produced inconsistent and contradictory findings. Although differential specific ability Flynn effect findings may eventually be identified, currently the supporting research results are sparse, mixed in results, and suffer from significant measurement and methodological flaws (McGrew, 2010). The foundation of Flynn effect causal theory, which hinges on the presence of differential specific ability Flynn effects, has been questioned on logical, theoretical, measurement and methodological grounds (Kaufman, 2010a, 2010b; McGrew, 2010; Weiss, 2010). Currently the extant research is not mature enough to support differential specific-ability Flynn effect adjustments in clinical or forensic contexts.

Differential Flynn effects by level of intelligence. The use of the 3 IQ test score points per decade Flynn effect adjustment rule-of-thumb has been questioned by research suggesting that the Flynn effect may not be uniform across all levels of general intelligence (Kanaya & Ceci, 2007; Kanaya, Ceci, & Scullin, 2003; Sanborn et al. 2003; Zhou et al., 2010). More important has been the suggestion that the Flynn effect may be larger at the IQ score range at the threshold for ID diagnosis. Cunningham and Tassé (2010) have referred to this research as the investigation of the Flynn effect in the "zone of ambiguity" (IQ test scores from 71–80). Studies reviewed by Cunningham and Tassé (2010) report IQ per decade changes ranging from roughly 4 to 5 points in the zone of ambiguity. Zhou et al. (2010) also reported differential Flynn effects by level of intelligence, but the results were inconsistent in the directions of the variation and may differ for different tests or age groups.

Similar to the differential Flynn effect by specific ability research, the ability-specific research has not been fully vetted through a sufficiently large number of studies and has been questioned on methodological grounds (McGrew, 2010; Widaman, 2007; Zhou et al., 2010). As summarized by Weiss (2010), "a small number of studies have suggested differential Flynn effect by ability level, but not enough is known about this at present" (p. 492). Reynolds et al. (2010) reinforce this conclusion, when after commenting on the Zhou et al. (2010 differential Flynn effects by levels of intelligence findings, that the results were inconsistent and "for now, best practice is the application of the Flynn correction as a constant by year across the distribution" (p. 480). Until more studies replicate the possibility of larger Flynn effects near the ID diagnostic threshold, the 3 points per decade Flynn effect rule-of-thumb should be employed across all levels of general intelligence.

Implications for Practice

The following implications are based on the integration of the content of the current chapter as well as the recommendations from the *User's Guide to the 10th edition,* the *11th edition,* and the *User's Guide to the 11th edition* (Schalock et al., 2007, 2010, 2012):

First, the potential problem of norm obsolescence can be minimized, but not always eliminated, by assessment professionals using IQ tests with the most up-to-date norms. When a new version of an IQ battery is published (e.g., WAIS-IV replaces WAS-III), assessment professionals should use the newest version (WAIS-IV) in *Atkins* cases. Assessment professionals have an ethical responsibility to stay abreast with the publication of new versions of IQ batteries and when the option exists to select among different IQ tests to administer to an individual. The relative degree of norm obsolescence of each possible IQ test should be one important factor incorporated into the IQ test selection decision.

Second, in cases where current or historical IQ test scores are impacted by norm obsolescence (i.e., Flynn effect), and the scores are to be used as part of the diagnosis of ID in *Atkins* or other high stakes decisions, the global scores impacted by outdated norms should be adjusted downward by 3 points per decade (0.3 points per year) of norm obsolescence.

Third, the recommended formula for the Flynn effect adjustment is: *FE adjustment = (Date test administered – date test was normed) × 0.3.* Stated simply, subtract the date the IQ test was normed (see point seven below) from the date the test was administered to the individual, multiply the obtained difference by 0.3. The obtained Flynn effect adjustment value should then be subtracted from the inflated obtained IQ score. The final Flynn effect adjustment value should be an integer value. Thus, the treatment of decimals in the final value should adhere to standard mathematical rules of "rounding to the nearest integer." The rationale for the particular rounding strategy employed should be described in the report. Current research does not support the application of different Flynn effect adjustment values for different part scores on IQ tests or at different levels of general intelligence. The best scientific evidence and professional consensus is that until sufficient research evidence produces evidence to the contrary, the 3 points per decade (0.3 points per year) adjustment rule-of-thumb should be used only on the global IQ test score and should be employed uniformly across all levels of general intelligence.

Fourth, both the original obtained (unadjusted) and Flynn effect adjusted scores should be included in all reports or court related statements or declarations provided by assessment professionals.

Fifth, the rationale for employing a Flynn effect correction should be described with supporting references. This chapter is intended to serve this function and can be cited as an authoritative source for the use of the Flynn effect adjustment in reports.

Sixth, when writing and discussing the Flynn effect, such as in psychological reports, legal declarations, or expert testimony, professionals should make frequent use of the term *norm obsolescence* when explaining the Flynn effect. Norm obsolescence is a much more descriptive and understandable means for conveying the essence of the Flynn effect.

Seventh, the calculation of the years of norm obsolescence should be based on the difference between the year the test was administered to an individual and the best

estimate of the year the IQ test was *normed* (see also Chapters 7 and 8). The data of publication of an IQ test does not accurately capture the time period when the test norm data were gathered. For example, the WISC-R IQ test was published in 1974 and the WISC-R norm data was gathered on children from 6 through 16 years of age from 1971 through 1973 (Wechsler, 1974). Thus, the middle most year of the actual norm data collection period is 1972. For the WISC-R, the year 1972 should be subtracted from the date of testing to determine the number of years of norm obsolescence. The test norm years reported for the different IQ tests by Flynn (2009) are recommended for uniformity purposes. For tests not reported in Flynn (2009), professionals need to consult the technical manuals for the IQ test in question and establish the best year estimate that is at the middle of the norm data collection period. If not readily available, professionals should seek the expertise of the test authors, publisher, or other intelligence test experts who may possess this information.

This chapter concludes with an example from an *Atkins* case. In 1998 an individual was administered the WAIS-R and obtained a Full Scale IQ of 80. Despite knowing that the WAIS-R had been revised and published as the WAIS-III in 1997, the psychologist administered the WAIS-R despite 20 years of norm obsolescence. The WAIS-R was published in 1981 and the best estimate of the date the actual test norms were gathered, as per the recommended procedures above, is 1978. Thus, the difference between the date of WAIS-R testing (1998) and date of test norming (1978) was 20 years, Using the 0.3/year Flynn effect adjustment, the best estimate of the magnitude of IQ test score inflation due to norm obsolescence is 6 IQ test score points ($0.3 \times 20 = 6.0$). Thus, this individual's Flynn effect adjusted WAIS-R score is 74 ($80 - 6 = 74$). This example represents one of the most dramatic instances of norm obsolescence (20 years) and also reflects the fact that the examiner did not engage in proper practice by administering the WAIS-III which was available at the time the individual was assessed.

References

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

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

Batterjee, A. A., Khaleefa, O., Ali, K., & Lynn, R. (2013). An increase in intelligence in Saudi Arabia, 1977–2010. *Intelligence, 41*(2), 91–93. doi: 10.1016/j.intell.2012.10.011

Baxendale, S. (2010). The Flynn effect and memory function. *Journal of Clinical and Experimental Neuropsychology, 32*(7), 699–703. doi: 10.1080/13803390903493515

Cunningham, M. D., & Tasse, M. J. (2010). Looking to science rather than convention in adjusting IQ scores when death is at issue. *Professional Psychology: Research and Practice, 45*(5), 413–419. doi: 10.1037/a0020226

Fletcher, J., Stuebing, K., & Hughes, L. (2010). IQ scores should be corrected for the Flynn effect in high stakes decisions. *Journal of Psychoeducational Assessment, 28*(5), 469–473.

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

Flynn, J. R. (1985). Wechsler Intelligence Tests: Do we really have a criterion of mental retardation? *American Journal of Mental Deficiency, 90*(3), 236–244.

Flynn, J. R. (2000). The hidden history of IQ and special education: Can the problems be solved? *Psychology Public Policy and Law, 6*(1), 191–198.

Flynn, J. R. (2006). Tethering the elephant: Capital cases, IQ, and the Flynn effect. *Psychology, Public Policy, and Law, 12,* 170–189. doi:10.1037/1076-8971.12.2.170

Flynn, J. R. (2007a). *What is intelligence? Beyond the Flynn effect.* New York: Cambridge University Press.

Flynn, J. R. (2007b). Capital offenders and the death sentence: A scandal that must be addressed. *Psychology in Mental Retardation and Developmental Disabilities, 32(3),* 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,* 98–104. doi: 10.1080/09084280902864360

Flynn, J. R., & Widaman, K. F. (2008). The Flynn effect and the shadow of the past: Mental retardation and the indefensible and indispensible role of IQ. In L. M. Glidden (Ed.), *International review of mental retardation* (Vol. 35, pp. 121–149). Boston, MA: Elsevier.

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), 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*(3), 7–8.

Gresham, F., & Reschly, D. J. (2011). Standard of practice and Flynn effect testimony in death penalty cases. *Intellectual and Developmental Disabilities, 49*(3), 131–140. doi: 10.1352/1934-9556-49.3.131

Hagan, L. D., Drogin, E. Y., & Guilmette, T. J. (2010a). IQ Scores should not be adjusted for the Flynn effect in capital punishment cases. *Journal of Psychoeducational Assessment, 28*(5), 474–476. doi: 10.1177/0734282910373343

Hagan, L. D., Drogin, E. Y., & Guilmette, T. J. (2010b). Science rather than advocacy when reporting IQ scores. *Professional Psychology Research and Practice, 41*(5), 420–423.

Kanaya, T., & Ceci, S. J (2007). Mental retardation diagnosis and the Flynn effect: General intelligence, adaptive behavior, and context. *Child Development Perspectives, 1*(1), 62–63. doi: 10.1111/j.1750-8606.2007.00013.x

Kanaya, T., & Ceci, S. J (2011). The Flynn effect in the WISC subtests among school children tested for special education services. *Journal of Psychoeducational Assessment, 29*(2), 125–136. doi:10.1177/0734282910370139

Kanaya, T., & Ceci, S. (2012). The impact of the Flynn effect on LD diagnoses in special education. *Journal of Learning Disabilities, 45*(4), 319–326. doi: 10.1177/0022219410392044

Kanaya, T., Ceci, S. J., & Scullin, M. H. (2003). The rise and fall of IQ in special ed: Historical trends and their implications. *Journal of School Psychology, 41*(6), 453–465. doi:10.1016/j.jsp.2003.08.003

Kaufman, A. (2010a). "In what way are apples and oranges alike?": A Critique of Flynn's interpretation of the Flynn effect. *Journal of Psychoeducational Assessment, 28*(5), 382–398. doi: 10.1177/0734282910373346

Kaufman, A. (2010b). Looking through Flynn's rose-coloured scientific spectacles. *Journal of Psychoeducational Assessment, 28*(5), 494–505. doi: 10.1177/0734282910373573

Kaufman, K., & Weiss, L. (2010). Guest editor's Introduction to the special issue of *JPA* on the Flynn effect. *Journal of Psychoeducational Assessment, 28*(5), 379–381. doi:10.1177/0734282910373344

Lynn, R. (1983). IQ in Japan and the United States shows a growing disparity. *Nature, 306*, 291–292.

Lynn, R. (2013). Who discovered the Flynn effect? A review of early studies of the secular increase in intelligence. *Intelligence,41*(6), 765–769. doi: 10.1016/j.intell.2013.03.004

McGrew, K. (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*(5), 448–468. doi:10.1177/0734282910373347

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

Nijenhuis, J. T. (2013). The Flynn effect, group differences, and *g* loadings. *Personality and Individual Differences*, 55, 224–228. doi:10.1016/j.paid.2011.12.023

Nijenhuis, J. T., Cho, S. H., Murphy, R., & Lee., K. H. (2012). The Flynn effect in Korea: Large gains. *Personality and Individual Differences, 53*(2), 147–151. doi: 10.1016/j.paid.2011.03.022

Nijenhuis, J. T., Murphy, R., & van Eeden, R. (2011). The Flynn effect in South Africa. *Intelligence, 39*(6), 456–467.

Nijenhuis, J. T., & van der Flier, H. (2013). Is the Flynn effect on g? A meta-analysis. *Intelligence*, 41, 802–807.

Nijman, E. E., Scheirs, J. G., Prinsen, M. J., Abbink, C. D., & Blok, J. B. (2010). Exploring the Flynn effect in mentally retarded adults using a nonverbal intelligence test for children. *Research in Developmental Disabilities, 31*, 1404–1411. doi: 10.1016/j.ridd.2010.06.018

Reynolds, C., Niland, J., Wright, J., & Rosenn, M. (2010). Failure to apply the Flynn correction in death penalty litigation: Standard practice of today maybe, but certainly malpractice of tomorrow. *Journal of Psychoeducational Assessment, 28*(5), 477–481.

Rindermann, H., Schott, T., & Baumeister, A, (2013). Flynn effect in Turkey: A comment on Kagitcibasi and Biricik (2011). *Intelligence, 41*, 178–180. doi: 10.1016/j.intell.2013.02.003

Rodgers, J. L. (1999). A critique of the Flynn effect: Massive IQ gains, methodological artifacts, or both? *Intelligence, 26*(4), 337–356.

Rönnlund, M., Carlstedt, B., Blomstedt, Y., Nilsson, L. G., & Weinehall, L. (2013). Secular trends in cognitive test performance: Swedish conscript data 1970–1993. *Intelligence, 41*(1), 19–24.

Runquist, E. A. (1936). Intelligence test scores and school marks in 1928 and 1933. *School and Society, 43*, 301–304.

Sanborn, K. J., Truscott, S. D., Phelps, L., & McDougal, J. L. (2003). Does the Flynn effect differ by IQ level in samples of students classified as learning disabled? *Journal of Psychoeducational Assessment, 21*(2), 145–159.

Schalock, R. L., Buntinx, W. H. E., Borthwick-Duffy, S. A., Luckasson, R., Snell, M. E., Tassé, M. J., & Wehmeyer, M. L. (2007). *User's guide to mental retardation: Definition, classification, and systems of supports (10th ed.).* Washington, DC: American Association on Intellectual and Developmental Disabilities.

Schalock, R. L., Borthwick-Duffy, S. A., Bradley, V. J., Buntinx, W. H. E., Coulter, D. L., Craig, E. M., Gomez, S. C., Lachapelle, Y., Luckasson, R., Reeve, A., Shogren, K. A., Snell, M. E., Spreat, S., Tassé, M. J., Thompson, J. R., Verdugo-Alonso, M. A., Wehmeyer, M. L., & Yeager, M. H. (2010). *Intellectual disability: Definition, classification, and systems of supports (11th ed.).* Washington, DC: American Association on Intellectual and Developmental Disabilities.

Schalock, R. L., Luckasson, R., Bradley, V., Buntinx, W. H. E., Lachapelle, Y., Shogren, K. A., Snell, M. E., Thompson, J. R., Tassé, M. J., Verdugo-Alonso, M. A., and Wehmeyer, M. L. (2012). *User's guide to intellectual disability: Definition, classification, and systems of supports.* Washington, DC: American Association on Intellectual and Developmental Disabilities.

Skirbekk, V., Stonawski, Bonsang, E., & Staudinger, U. M. (2013). The Flynn effect and population aging, *Intelligence, 41*(3), 169–177. doi: 10.1016/j.intell.2013.02.001

Sternberg, R. (2010). The Flynn effect: So what? *Journal of Psychoeducational Assessment, 28*(5), 434–440. doi: 10.1177/0734282910373349

Trahan, L. H., Stuebing, K. K., Fletcher, J. M., & Hiscock, M. (2014, June 30). The Flynn effect: A meta-analysis. *Psychological Bulletin*. Advance online publication. http://dx.doi.org/10.1037/a0037173

Truscott, S. D., & Frank, A. J. (2001). Does the Flynn effect affect IQ scores of students classified as LD? *Journal of School Psychology, 39*(4), 319–334.

Wai, J., & Putallaz, M. (2011). The Flynn effect puzzle: A 30-year examination from the right tail of the ability distribution provides some missing pieces. *Intelligence, 39*(6), 443–455. doi:10.1016/j.intell.2011.07.006

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

Weiss, L. G. (2010). Considerations on the Flynn effect. *Journal of Psychoeducational Assessment, 28*(5), 482–493. doi: 10.1177/0734282910373572

Widaman, K. (2007). Stalking the roving IQ score cutoff: A commentary on Kanaya and Ceci (2007). *Child Development Perspectives,1(1)*,57–59. doi: 10.1111/j.1750-8606.2007.00011..x

Woodley, M. A. (2011). Heterosis doesn't cause the Flynn effect: A critical examination of Mingroni (2007). *Psychological Review, 118*(4), 689–693. doi: 10.1037/a0024759

Woodley, M. A. (2012a). A life history model of the Lynn-Flynn effect. *Personality and Individual Differences, 53*(2), 152–156. doi: 10.1016/j.paid.2011.03.028

Woodley, M. A. (2012b). The social and scientific temporal correlates of genotypic intelligence and the Flynn effect. *Intelligence, 40*(2), 189–204. doi:10.1016/j.intell.2011.12.002

Young, G. W. (2012). A more intelligent and just *Atkins:* Adjusting for the Flynn effect in capital determinations of mental retardation or intellectual disability. *Vanderbilt Law Review*, 615–755.

Zhou, X., Zhu, J., & Weiss, L. (2010). Peeking inside the "blackbox" of the Flynn effect: Evidence from three Wechsler instruments. *Journal of Psychoeducational Assessment, 28*(5), 399–411.

Vandenbos, G. (2007). *APA dictionary of psychology.* Washington, DC: American Psychological Association.

No. _____

IN THE

Supreme Court of the United States

TAVARES J. WRIGHT,

Petitioner,

v.

SECRETARY, DEPARTMENT OF CORRECTIONS, AND ATTORNEY GENERAL, STATE OF FLORIDA,

Respondents.

ON PETITION FOR A WRIT OF CERTIORARI TO THE UNITED STATES COURT OF APPEALS FOR THE ELEVENTH CIRCUIT

APPENDIX TO THE PETITION FOR A WRIT OF CERTIORARI

DEATH PENALTY CASE

APPENDIX R

Report by Dr. Alan Waldman, M.D., dated October 9, 2002

Alan J. Waldman, M.D.
Forensic Psychiatry

802 South Main Street Suite G
Gainesville, Florida 32601

Phone: 352-377-3771 FAX: 352-377-3717

October 9, 2002

Mr. David R. Carmichael
Franklin Law Firm, PA
310 East Main Street
Bartow, FL 33810

Dear Mr. Carmichael:

As you are aware, I evaluated Tavares Wright at the Polk County Jail on Monday, October 7, 2000. Mr. Wright seemed to have some difficulty in processing the full nature and scope of the evaluation but was for the most part a willing understanding participant. I was able to accomplish a full psychiatric and neuropsychiatric evaluation, which included a higher cortical function exam and a physical exam with a complete neurological assessment.

On inspection, I found his face to be consistent with fetal alcohol syndrome. Fetal alcohol syndrome is consistent with his developmental history as his mother is an alcoholic and a crack abuser and abused these substances during Tavares or T.J.'s gestation. He appeared to have a flat faces with slightly abnormally wide set eyes and a minimal or absence of a philtrum (sub nasal folds). Fetal alcohol syndrome is also well documented to be associated with a variety of abnormal adult behaviors.

The psychiatric portion of the examination revealed a very immature, somewhat silly at times, black male who was either acting younger than his chronological age with one might call an "Opie Taylor" type of demeanor or he was attempting to justify being tough without any rationale, primitive or otherwise. The basis for his tough like behavior, except that he has seen others do it, is that he feels that regardless of circumstance all individuals are his equals in all modes. He is unable to transpose individuals in different roles or circumstances having different responsibilities or social strata. He was unable to process the difference between me, the highly trained physician in my role and himself in his. He lacked the ability to process any hierarchical roles that children learn at an early age (e.g. teacher, principle, police officer etc.), therefore sees no need to modulate behaviors to others.

This ties into what appears to be evidence of hypofrontality or a dysfunction in the use of his frontal lobes. TJ essentially lacks the ability to plan behaviors and certainly lacks the ability to weigh consequences or alternative behaviors when placed in a situation requiring choices. He readily parrots those around him but cannot formulate any of his own coherent courses of action.

The higher cortical function exam was significant for an abnormal frontal lobe essentially leaving him as an individual who lacks the neurologic capacity to do anything except from go from thought to action.

The physical exam was essentially normal though was significant for frontal lobe deficits as well. On the occiput (the very back) of Tavares's skull is a 1 cm x 1 cm bony protrusion consistent with a bony growth following a head impact, which would have caused a contracoup type of injury to the frontal lobe. Tavares Wright has had multiple impacts to the head during his life.

IMPRESSIONS

I opine with reasonable medical certainty that Tavares Wright suffers from fetal alcohol syndrome as well as from a neurologic syndrome resulting in an impaired frontal lobe. His history points to him being in essence a Farrell child being raised by whichever relative has a place for him to stay with no guidance, impartation of values or consistency.

Children such as TJ are set up for gang activity as it is the only outlet for a sense of belonging that they have ever experienced in their life. It is not surprising that his gang activity increased following the death of his only stable caregiver, his grandmother and his rejection from joining the Navy, his fantasy of becoming part of normal society.

RECOMMENDATIONS

I have very serious doubts whether Tavares Wright has the neurologic capacity to have done any more than followed in any of these criminal behaviors, as opposed to formulating the plans or even forming intent. It is my suggestion that he receive a full neurologic workup with neuropsychological testing, an electroencephalogram with temporal lobe leads and an MRI to rule out any structural lesions that might explain Tavares abnormal neurologic findings. Regardless, this is an individual who is neurologically abnormal, most probably has fetal alcohol syndrome, lacks frontal lobe capabilities and appears behaviorally void the neurologic ability to plan anything but simple 1-2 step behaviors thoughroughly inconsistent with the three-day period of the crime spree of which he was involved.

Sincerely,

Alan J. Waldman, M.D

Diplomate American Board of Psychiatry and Neurology
General and Forensic Psychiatry

Clinical Assistant Professor, Department of Psychiatry
University of Florida, College of Medicine

No. _____

IN THE

Supreme Court of the United States

TAVARES J. WRIGHT,

Petitioner,

v.

SECRETARY, DEPARTMENT OF CORRECTIONS, AND ATTORNEY GENERAL, STATE OF FLORIDA,

Respondents.

ON PETITION FOR A WRIT OF CERTIORARI TO THE UNITED STATES COURT OF APPEALS FOR THE ELEVENTH CIRCUIT

APPENDIX TO THE PETITION FOR A WRIT OF CERTIORARI

DEATH PENALTY CASE

APPENDIX S

Report by Dr. Joel Freid, dated August 25, 1997

JOEL B. FREID, Ph.D., P.A.

CLINICAL PSYCHOLOGY

ADULT CLINICAL PSYCHOLOGY
CHILD CLINICAL PSYCHOLOGY
FORENSIC PSYCHOLOGY
FLORIDA LICENSE #PY0002213

4440 FLORIDA NATIONAL DRIVE
LAKELAND, FLORIDA 33813
Telephone (863) 644-0586
Fax (863) 644-7522

August 25, 1997

RESUME OF PSYCHOLOGICAL EVALUATION

TO: Lillie McQueen, Medical Disability Adjudicator
Office of Disability Determinations
P. O. Box 5340
Tallahassee, Florida 32314-9944

RE: TAVARES WRIGHT
3235 Skyview Drive
Lakeland, Florida 33801

Social Security #:	594-03-3546
Date of Evaluation:	8/12/97
Date of Birth:	2/7/81
Chronological Age:	16 Years, 6 Months
Education:	9th (Special Education)
Marital Status:	Single

EVALUATION PROCEDURES:

Wechsler Adult Intelligence Scale-Revised; Reynolds Adolescent Depression Scale; Review of Records; Mental Status Examination; Clinical Interview and Observations.

REFERRAL AND PRESENTING PROBLEM:

Tavares Wright is a 16 year, 6 month old Special Education student who was referred to the examiner by the Office of Disability Determinations for the purpose of a General Intellectual Evaluation and a General Clinical Evaluation with Mental Status Examination. Tavares was accompanied to the evaluation session on August 12, 1997, by his mother, Patricia Anderson. When questioned about Tavares' disability, his mother stated, "He's an ESE student - classified as slow learning disability - emotionally handicapped. He is not in regular classes." Tavares is in the "Adjudicate Program" under Bill Duncan School. Tavares is in that program because of trouble he has been in - "car theft, fighting in school."

ADDITIONAL HISTORY AND BACKGROUND INFORMATION:

Tavares was born on February 7, 1981, and he is now 16 years, 6 months of age.

Tavares is a ninth grade student. Apparently, he has received Special Education services through the SLD Program. However, presently, he is in the Special Education Program for the Emotionally Handicapped. According to the records, Tavares received speech therapy in school.

Tavares was previously evaluated in 1991 by Kevin Kendalin, Ph.D. At that time Tavares was evaluated to be functioning within the Borderline Range of General Intellectual Ability as measured on the WISC-R. He obtained a Verbal Scale I.Q. of 84, a Performance Scale I.Q. of 72, and a Full Scale I.Q. of 76.

Tavares is presently living with his mother and one sister and one brother. Mrs. Anderson is 33 years of age and is employed by a dental company. Tavares' biological parents were never married. Tavares thinks that his father is in Lakeland. Tavares sees his father now and then.

There is no prior history of abuse.

Tavares is reported to be in good physical health. He was involved in an automobile accident in 1994 and he injured his leg (he has a pin in it) and his back and neck as a result of that accident.

There is no prior history of head injury.

Tavares is not taking any medications at the present time.

There is no prior history of mental health treatment other than counseling through his probation.

Tavares used to smoke marijuana - he no longer smokes. He denies the use of alcohol and cigarettes.

Tavares has been arrested numerous times. He has been in the juvenile detention center approximately three times. He has been arrested for such things as Grand Theft Auto, Assault, Trespassing, etc. Tavares is no longer on probation.

When he is not in school, Tavares stays home. He enjoys playing football and baseball and basketball. He does not drive an automobile - he has no license. He helps with chores occasionally, but often doesn't listen to his mother's request to do some chores - "most of the time he doesn't," according to his mother.

Tavares performs all of his self-care skills independently.

MENTAL STATUS EXAMINATION:

Tavares is a tall, thin, handsome young man who presented an acceptable personal appearance.

Tavares was alert but lethargic appearing. His concentration is decreased, questions have to be repeated to him. Eye contact is poor.

Tavares was generally cooperative with the examiner. There was no verbal spontaneity. He responded to questions relevantly and appropriately. His speech articulation was clear and there was no evidence of any confusion. He speaks in a very low voice.

Tavares' affect was appropriate. His mood was mildly depressed and he seems to have a rather dour expression on his face. According to his mother, he bites his nails and twists his hair. Hair twisting was observed by the examiner during the interview. According to Tavares, however, he does not feel sad and he doesn't cry - "Tears may run out of my eyes sometimes." Tavares thinks about his grandmother a lot who died a little over one year ago. Tavares was very close to her. At this point, both Tavares and his mother began to cry as they talked about his maternal grandmother. Tavares then stated that he wanted to stop this interview and go home. He repeated this several times, however, he finally agreed to continue with encouragement and support from his mother and the examiner. Tavares' performance on the Reynolds Adolescent Depression Scale would indicate that he is denying any significant symptomatology associated with a depressive disorder.

WRIGHT, Taveres

Tavares is a behavior problem at school. He doesn't like his teacher and thought that she was against him. Tavares has been to the point where he has wanted to quit school in the past. At times, he gets into fights. At home, he gets mad at his mother and takes it out on his siblings. He curses. He feels that his mother doesn't care about him at times.

Tavares was well oriented to person, place, and time.

Tavares' general fund of information, immediate recall for auditory, and vocabulary and verbal fluency were all assessed to be very low for his age. His arithmetic reasoning ability, social insight and judgement, and verbal abstract reasoning abilities were all assessed to be low.

There was no evidence of any auditory and/or visual hallucinatory phenomena.

During the formal testing phase of the evaluation, Tavares was cooperative, however, he needed a great deal of encouragement to complete certain tasks before giving up rather easily.

EVALUATION RESULTS:

INTELLECTUAL FACTORS:

Wechsler Adult Intelligence Scale - Revised:

Areas Measured	Scaled Score	Rating	Approx. Percentile:
Verbal Scale			
Information	3	Very Low	1
Digit Span	4	Very Low	2
Vocabulary	4	Very Low	2
Arithmetic	6	Low	9
Comprehension	5	Low	5
Similarities	5	Low	5
	Verbal Ability: Level - Borderline Approx. Percentile - 4		
Performance Scale			
Picture Completion	6	Low	9
Picture Arrangement	7	Low	16
Block Design	9	Average	37
Object Assembly	7	Low	16
Digit Symbol	5	Low	5

620

Nonverbal Cognitive Ability:
Level - Low Average (Lower Limits)
Approx. Percentile - 9

General Intellectual Ability:
Level - Borderline
Approx. Percentile - 5

Tavares' overall performance on the Wechsler Adult Intelligence Scale-Revised fell within the Borderline Range of General Intellectual Ability at the 5th Percentile. On the WAIS-R, he obtained a Verbal Scale I.Q. of 73, a Performance Scale I.Q. of 80, and a Full Scale I.Q. of 75.

CONCLUSIONS AND RECOMMENDATIONS:

Tavares Wright is a 16 year, 6 month old Special Education student who was evaluated to be functioning within the Borderline Range of General Intellectual Ability as measured on the WAIS-R. He obtained a Verbal Scale I.Q. of 73, a Performance Scale I.Q. of 80, and a Full Scale I.Q. of 75.

CLINICAL IMPRESSIONS:

(1) Conduct Disorder with Antisocial Personality Disorder Developing,

(2) Adjustment Disorder with Mixed Emotional Features,

(3) Borderline Intellectual Functioning.

I would certainly recommend that Tavares continue to participate in his Special Education in his public school program and not drop out.

I also feel that Tavares is in need of mental health counseling and should avail himself of such on an outpatient basis, as needed.

Joel B. Freid, Ph.D.
Clinical Psychologist

JBF/sp