Science Rather Than Advocacy When Reporting IQ Scores

Leigh D. Hagan
Virginia Commonwealth University, and Independent Practice, Chesterfield, VA

Eric Y. Drogin
Harvard Medical School

Thomas J. Guilmette
Providence College

The existence of shifts in mean IQ scores over time is well established. However, on a case-by-case basis, such shifts vary unreliable, rendering specific adjustments to a given individual’s IQ score inacalculable. Based upon data presented previously (Hagan, Drogin, & Guilmette, 2008) as well as a review of more recent studies that have further detailed the wide variability of mean score shifts, any proposal to “correct” IQ scores in forensic evaluations due to the “Flynn effect” (FE) is unjustifiable. To offer the court an unreliable new IQ score in place of an allegedly unreliable old one—and to do so specifically in capital murder cases as opposed to any other context—appears far more reflective of result-focused advocacy than objective scientific practice. Forensic psychologists are explicitly encouraged to address likely ranges of IQ score variability and to discuss in relevant detail the strengths and weaknesses of the specific studies—however much at odds these may be—that attempt to define and quantify mean score shifts.

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

Our significant overall concurrence with Cunningham and Tassé (this issue, pp. 413–419) should be considered in conjunction with their few but emphatic points of disagreement with our original article and its underlying research (Hagan, Drogin, & Guilmette, 2008). We agree with Cunningham and Tassé that psychologists should report obtained IQ scores and should explain factors that might impact the reliability and validity of these data. We agree that experts’ work product and testimony should arise from the best scientific methods and procedures. We agree that mean IQ scores shift over time, that psychologists should use the most current appropriate IQ tests, and that test developers should update norms.

We also agree with McGrew (2010) that “the consensus of most (but not all) intelligence scholars is that the Flynn Effect (FE), at the level of the global IQ score, is real” (p. 449). We did not contest that view in 2008 and do not argue against it today. We do assert—as forensic psychologists employed in criminal matters by defense attorneys and prosecutors alike—that it is important to recognize the full spectrum of available science unfettered by a priori positions or by the consequences of having cast one’s professional lot with advocacy for a particular legal result.

Our disagreement lies in how psychologists should convey IQ scores in light of the observation that mean scores drift over time. Cunningham and Tassé (this issue) maintain that psychologists should report the obtained scores, describe the FE, and “report the corrected IQ scores” (p. 418). Our earlier contention that “adjusting the numerical scores” is inappropriate (Hagan et al., 2008, p. 623) remains unchanged (Hagan, Drogin, & Guilmette, 2010), and we appreciate this opportunity to state plainly that since research scientists are unable to agree on how much the FE might have affected an examinee’s obtained score, it is at best—to adopt our colleagues’ turn of phrase—“disingenuous or uninformed” (Cunningham & Tassé, p. 417) to cloak in science and offer up to the court an unreliable new number in place of an allegedly unreliable old one.

This is not to suggest that we somehow seek to place obtained IQ scores upon a pedestal, such that their potential unreliability should never be addressed in a legal context. On the contrary, we explicitly encourage forensic psychologists who testify in death penalty cases at the request of the defense or prosecution to address for the court’s edification the likely ranges of IQ score

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 his J.D. 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 Board Certified in Clinical Neuropsychology and is a Professor of Psychology in the Department of Psychology at Providence College in Providence, Rhode Island. His areas of professional interest include neuropsychology and psychology and law. The authors derive income from evaluations and testimony in all manner of criminal forensic contexts, employed in such matters by defense attorneys and prosecutors alike.

CORRESPONDENCE CONCERNING THIS ARTICLE should be addressed to Leigh D. Hagan, P.O. Box 350, Chesterfield, VA 23832. E-mail: lhagan@leighhagan.com
variability and to discuss in relevant detail the strengths and weaknesses of the specific studies—however much at odds these may be—that attempt to define and quantify mean score shifts. However, we do not encourage experts to “reset” the IQ score based on these score shifts.

Irrespective of which party calls them to the stand, testifying experts should add to their current trial preparation checklists a reminder that to “discuss in relevant detail” means just that. Experts are present to provide information that will “assist the trier of fact” (Fed. R. Evid. 702). We would be sorely disappointed should our own advice be misconstrued, for example, as an invitation to disgorge a torrent of obscure research citations for the purpose of discrediting an opposing expert or confusing a jury.

Such maneuvering— which was in no way suggested by Cunningham and Tassé—is the very antithesis of assistance. The goal of acknowledging differences of opinion in the scientific literature is fairness, not filibustering. Testimonial candor is a professional obligation, not a tactical ploy.

In support of their perspective on how to convey IQ scores, Cunningham and Tassé (this issue) emphasize that to bind the profession of psychology for all time to a chronologically fixed “standard of practice” would ultimately stifle needed evolution in assessment procedures. Not only do we agree, but also, of course, we never suggested otherwise. We undertook to research the current standard of practice not in an attempt to stop the clock on progress in psychological testing but rather to investigate how the courts are likely to treat activities that we discovered are no more grounded in the established procedures of practitioners and the experts who train them than they are in agreed-upon scientific principles. If one is distressed or offended by the prevailing standard of practice, then one should seek to transform it via research, scholarly debate, and professional example, rather than to assert—to the bewilderment of the courts, who will always need to know what that “standard” is—that identifying its current status is irrelevant or counterproductive.

For Cunningham and Tassé (this issue) to criticize the definition of a “standard” on the basis of it having been “taken from a legal dictionary” (p. 414) is ironic in light of their adoption of the overtly legal notion—cited not to any scientific articles but rather to four different legal decisions—that “death is different” (p. 414).

We agree that “death is different” . . . . for lawyers. For those administering the IQ test and reporting the scores, however, “death” should most decidedly not be treated as “different.” The ethical obligation to conduct assessments competently and to interpret outcomes accurately attaches irrespective of the context in which these activities are conducted. Adherence to professional standards should not be parcelled out based upon which legal undertakings psychologists deem the most worthy or consequential.

Mean IQ Scores Over Time: Further Considerations

While the general proposition that mean scores change over time appears well settled, the term “the Flynn effect” is a bit misleading. When judges or juries hear about “correcting for the Flynn effect,” there is the risk that they will assume that the testimony describes a singular, unidirectional, and linear statistical phenomenon, which is relevant, reliable, has been tested, and has controls for its operation (Daubert v. Merrell Dow Pharmaceuticals, Inc., 1993) or at least is generally accepted (Frye v. United States, 1923) in the field of psychology. These assumptions simply are not established.

The drift in mean scores has not been shown to be reliable in degree or direction. For almost a quarter century, studies have repeatedly demonstrated a wide range of score shifts. The data for at least eight IQ tests across 13 countries showed per-year shifts from 1.250 to $-0.133$ (Flynn, 1987), a range that is four-fold the adjustment proposed by the American Association on Intellectual and Developmental Disabilities (Schalock et al., 2007). Recent data continue to substantiate a wide range of per-year shifts from 0.917 to $-0.117$ (Flynn, 2009a), which is triple the proposed adjustment, even when the field is limited to U.S. scores on the Wechsler and Stanford-Binet tests.

Several specific factors impact score shifts, including where the score falls in the distribution. Zhou, Zhu, and Weiss (in press) found that with PIQ scores the degree of shift was a function of ability: “This finding is consistent for all Wechsler intelligence tests studied spanning the range from preschool to older adult. Overall, our findings suggest that the average IQ gain Flynn initially described may only be valid as an aggregated phenomenon” (p. 408).

Age makes a difference. In revisiting his earlier research, Kaufman (2010) reported WISC/WISC-R FSIQ shifts of .237 to .094 for different age ranges. Kaufman (2010) also pointed out the significant differences between iterations of the tests (“apples and oranges”), including changes in the subtests, order of subtest administration, rules for querying, scoring rubrics and other modifications that make it difficult to attribute generational shifts in mean scores to any one source, particularly when one subtest alone (Similarities) changed by an average .565 points per year. Kaufman (2010) concluded, “Flynn’s mistake was in assuming that the subtests did not change in meaningful ways over time and that is wrong. It is also, therefore, wrong to build a theory based on the unique abilities supposedly measured by specific subtests” (p. 393).

Although we present the rate of shifts as “per-year” throughout this paper, it is important to note that there exists little or no research to support the assumption that shifts occur in equal annual increments or even in the same direction each year post-norming. Identifying the direction in which scores drift over time is also problematic. Over the last 22 years, Flynn’s data (1987, 2009a) demonstrated some reversals, an observation documented by other research cited in our original paper and reconfirmed more recently with respect to advanced Western countries (Russell, 2010).

The shift discrepancies appear to be wide-ranging, making it difficult to justify calculating a simple average rate of change as best-science. As Zhou et al. (2010) concluded, “the Flynn effect is a much more complicated phenomenon than a simple overall increase of IQs” (p. 409). The proponents of score adjustments calculate an average among a wide range of score shifts between and within the same and different IQ tests without due consideration to the impacts of age, ability, and subtest loading and then call it a “standard.” Equally problematic is the proposal from Flynn (2006) and Fletcher, Stuebing, and Hughes (2010) to discard outlier data in order to fit a model, an unscientific endeavor Weiss (2008) roundly rejected.

Cunningham and Tassé (this issue) refer to “progressively obsolete norms” (p. 416) as if they have no value, or worse, contrib-
ute more error than accuracy. “Obsolete” means no longer useful. Russell (2010) pointed out that “(i)n psychometric theory, validated tests do not lose their validity due to the creation of newer versions. Time does not invalidate tests due to the improvement of neurological methodology, such as magnetic resonance imaging. This assumption is unscientific, unproven, and if true, would discredit all older neuropsychological and neurological knowledge. In science, no method, theory, or information, once validated, loses that validation merely due to time or the creation of another test or procedure” (p. 60).

Atkins and the Redirected Conversation

In a section of their current article entitled “The Unacknowledged Elephant in the Room,” Cunningham and Tassé (this issue) maintain that in our 2008 article we did not “directly reference capital sentencing applications,” although “Atkins cases are almost certainly the primary intended audience for [Hagan et al.’s] analysis and commentary” (p. 414). In fact, in the very first paragraph of our article, we observed that psychologists’ reports of IQ test data can have a major impact on access to services and even life-and-death decisions” (Hagan et al., 2008, p. 619), and cited Atkins v. Virginia (2002) in support of that assertion.

For the 18 years between Flynn (1984) and the Atkins v. Virginia (2002) decision, the discourse about score shifts over time was not about the death penalty (DP). The Atkins decision changed the conversation. Scientific methods and the data did not change appreciably, although there was more of the latter. The way practitioners reported scores did not change noticeably. What changed was the opportunity to impact the outcome of DP sentencing proceedings.

Prior to Atkins, much of the scholarly discussion focused on psychometric issues and a search for a causal explanation. With the advent of Atkins, practice-focused issues have come to the forefront, including a call by some for subtracting IQ score points from the obtained score. Flynn (2009b) recently observed “the formula of deducting 0.30 points per year is making headway at least among defense attorneys” (p. 134), an acknowledgment of the specialized interest area in which score adjustments have found traction. In a recent capital resentencing, Flynn confirmed the outcome-focused strategy for score adjustments when he testified that psychologists should modify scores depending on their legal and social consequences: “I do it in terms of the real world result” (Ex parte Cathey, 2010, p. 94). Toward that same end, he advised that he would adjust scores to facilitate approval of disability benefits because doing so could extend or improve the claimant’s life but would not adjust scores where outcomes are not as consequential, as in an educational placement decision about a special reading program.

The fact that so many practice-focused publications on the score shifts over time have almost exclusively addressed DP cases strongly suggests that this is an outcome-focused issue. There is no voice calling for the reduction of obtained IQ scores in the high end or mid-range. In contrast with Flynn’s (1987) cornerstone research, his more recent publication (Flynn, 2009a) is silent with respect to the countries that do not have the DP. He has narrowed the conversation further to IQ tests for Americans (Ex parte Cathey, 2010). Overall, the focus of the conversation is narrower while the discourse is more impassioned.

Cunningham and Tassé (this issue) propose a remedy for the IQ score drift over time: “report the corrected IQ scores” (p. 418). Here, the language subtly but significantly shifts from “adjusting” scores to “correcting” them. This is a distinction with a difference. Inherent in “correction” is a change that rectifies an error. This presumes that the psychologist can know the degree and direction of the error with sufficient certainty to modify the evidence for a specific individual of a particular age with a circumscribed ability level on a given test whose norms are of a certain age. Missing is the acknowledgment that the “true score” remains theoretical due to other sources of error, such as administration and scoring miscalculation, practice and feedback effects, and compromises caused by the test-taker’s poor attention and effort.

The next step down the slippery slope is to reify the process of modifying scores by designating it with the formal moniker “Flynn Correction.” Reynolds, Niland, Wright, and Rosen (2010) advanced the cause to this next level and insinuated a threat for psychologists who do not “correct” scores by declaring “if the FE is real, the failure to apply the Flynn Correction as we have described it is tantamount to malpractice” (p. 480).

Fletcher et al. (2010) fully support changing obtained IQ scores, pointing out that “correcting an IQ score is a simple procedure that avoids having to change standards” (p. 470). They propose that it is a simple process to multiply the number of years post-norming by the purported Flynn Correction (0.3 points/year) to establish a “contemporary sample” mean against which the obtained score could be interpreted.

While the calculation is simple, it is simply incorrect, unwarranted, and unsupported for several reasons. In spite of Fletcher et al.’s (2010) pronouncement that “correcting an IQ score is not a violation of test administration,” (p. 470) it is well established that psychologists should follow the test manual instructions or risk violating the standardization. None of the test manuals, including Reynolds’ (Reynolds Intellectual Assessment Scales; Reynolds & Kamphaus, 2003), directs psychologists to engage in post-hoc score revisions.

Unlike the standard error of measurement, which provides the probability of a score falling within a particular range based on known rates of variance, the proposal to “correct” IQ scores provides no such assurances that the new score is more reliable for a particular individual. As noted in Justice Jeanette Knoll’s recent concurrence in the Supreme Court of Louisiana case of State v. Dunn (2010), even if one were to accept the notion of a phenomenon that affects every individual’s IQ score, “it most certainly does not follow that it inflates every individual’s score equally or consistently on a year to year basis. Yet proponents of the Flynn effect would have courts apply a 0.3 points per year across the board” (p. 20).

The assumptions necessary for recalculating means, short of full periodic renorming, simply are not borne out: “It is clear that we still have a long way to go before we understand the exact nature and magnitude of the FE for all types of individuals, across time, and across tests. Until we do, discussions regarding its cause(s) and interpretation may be futile, as will solutions for trying to ‘fix’ it” (Ceci & Kanaya, 2010, p. 446).
Conclusions

We agree that mean IQ scores shift over time. However, the magnitude and direction of that shift are not predictable (Russell, 2010). In addition, a number of other specific factors affect the shifting of scores, including where the score falls on the distribution (Zhou et al., 2010), the age of the examinee (Kaufman, 2010), and the changes in the iterations of the tests themselves (Kaufman, 2010). Thus, unlike other psychometric concepts that affect score reliability but that are measurable and predictable (e.g., the standard error of measurement), the degree to which group mean score shifts may impact an individual’s score is incalculable.

References

Ex parte Cathey, Cause No. 713189-B, 176th Judicial District Court, Harris County, Texas (2010).