## Looking Through Flynn's Rose-Colored Scientific Spectacles

Alan S. Kaufman

*Journal of Psychoeducational Assessment* 2010 28: 494 originally published online 14 June 2010
DOI: 10.1177/0734282910373573

Journal of Psychoeducational Assessment
28(5) 494–505

DOI: 10.1177/0734282910373573
http://jpa.sagepub.com
SAGE

# Looking Through Flynn's Rose-Colored Scientific Spectacles

**Alan S. Kaufman**[1]

## Abstract

In the first article of this special issue of the *Journal of Psychoeducational Assessment*, I critiqued Flynn's theoretical explanation of the Flynn effect because he depended too heavily on an apparently huge generational gain on the WISC Similarities subtest; I claimed he was comparing apples with oranges because that subtest changed too much when the WISC was first revised. Four sets of esteemed researchers were invited to respond to my article and also to an article by Zhou, Zhu, and Weiss—Flynn, Sternberg, McGrew, and Ceci and Kanaya. Flynn disagrees strongly with my critique of his theory, a theory that posits striking generational shifts from concrete (utilitarian) thinking to a kind of fluid reasoning that Flynn nicknames "scientific spectacles." In this final article of the special issue, I respond to Flynn's claims, and also to the points made by the other invited respondents. In addition, I respond to the range of opinions expressed by the scholars who were invited to write an essay on whether or not IQs should be adjusted for the Flynn effect in capital punishment cases (Fletcher et al., Hagan et al., and Reynolds et al.). Ultimately, I disagree with Flynn's explanation of the Flynn effect, but I agree with his position that IQs should be adjusted for the effect in death penalty cases.

## Keywords

Wechsler scales, Flynn effect, IQ tests, capital punishment, Cattell-Horn-Carroll (CHC) theory

This special issue of the *Journal of Psychoeducational Assessment* provides a scientific forum for debating the nature and magnitude of the Flynn effect (FE), its causes, and its implications for society. The centerpiece articles (Kaufman, 2010; Zhou, Zhu, & Weiss, 2010) were intended to spark this debate and to encourage a dialogue among researchers, test developers, and theoreticians on an empirical finding that has had broad-based applications throughout the world and striking life-or-death implications within the United States. The four articles that were invited specifically to respond to the two featured papers abound with new empirical analyses (Flynn, 2010; McGrew, 2010); an enriched understanding of how cognitive-developmental research pertains to the FE as well as the clinical implications of the effect (Ceci & Kanaya, 2010); and new theoretical speculation concerning CHC (Cattell–Horn–Carroll) theory (McGrew, 2010) and the components of successful intelligence (Sternberg, 2010). The three articles invited to address the

[1]Yale University School of Medicine, New Haven, CT, USA

**Corresponding Author:**
Alan S. Kaufman, 3033 Curie Street, San Diego, CA 92122, USA
Email: alanskaufman@gmail.com

role of the FE in death penalty cases (Do we adjust IQs—Yes or no?) crisply delineate the wide spectrum of opinions, as well as emotions, on this volatile issue.

Weiss (2010) responds to most of the thorny issues dealt with by the various authors (including the "apples and oranges" question that I raised in my article), and my aim here is to supplement Weiss's observations and opinions, not repeat them.

## Do Changes From WISC to WISC-R Reduce the Size of the Flynn Effect?

I wrote the apples–oranges paper (Kaufman, 2010) to challenge Flynn's (2007, 2009) *interpretation* of the FE, not the effect itself. I stated,

> I am not claiming that the notable changes from the WISC to the WISC-R challenge either the existence of the FE or its magnitude. But I readily challenge Flynn's interpretation of gains on specific subtests such as Similarities and his reliance on those subtest differences to help formulate his explanation of the FE and to develop his theory of intelligence. (Kaufman, 2010)

Ceci and Kanaya (2010) believe that I have challenged not only Flynn's interpretation of the effect but also the body of research that has accumulated on the FE. They are correct that, "Kaufman spends a considerable amount of time on the changes in the instructions that were introduced with the WISC-R." They are decidedly *incorrect* to state (a) "we question the impact of these administrative/scoring changes to account for most of the impact, given the heavy documentation of the Flynn effect on multiple IQ tests and norms over time and around the world," and (b) "More specifically, Kaufman questions the overall finding of higher gains on tests of fluid abilities . . . compared with tests of crystallized abilities" (Ceci & Kanaya, 2010).

Regarding their first point, I do not believe that the changes in administration affected the magnitude of the FE at all. The extensive changes render uninterpretable (and likely inflate) any reported gains on certain WISC subtests, but that is unrelated to the overall gain of 3 global IQ points per decade—a difference that is observed even when different IQ tests, containing completely different subtests, are compared with each other. Indeed, I specifically said,

> The 3-point per decade finding in the United States seems to be robust. It has been identified for the WISC-III and WISC-IV for various combinations of Wechsler's and Binet's scales (Flynn, 2007, table 3) and also for Wechsler/K-ABC comparisons for normal children and those diagnosed with learning disabilities, behavior disorders, mental retardation, and hearing impairment. (Kaufman, 2010)

## The Flynn Effect for Fluid Versus Crystallized Tasks

Concerning Ceci and Kanaya's second point, I do not dispute the repeated finding that the FE is larger for fluid than crystallized tests. That is a given from the abundant research, and that finding has been observed for a wide variety of tests, not just the Wechsler subtests and factors that are alleged to measure fluid thinking or Raven's matrices. I have summarized that research elsewhere (Kaufman & Lichtenberger, 2006, pp. 39-41). However, the fact that gains tend to be larger on fluid than crystallized tasks does not, by itself, provide empirical support for Flynn's (2007, 2009b) theory that society has changed dramatically from a species that wallowed in concrete modes of thought to a present-day group of "abstract whizzes." Furthermore, the worldwide generational gains averaged 6 to 7 points on nonverbal tasks (including Wechsler's Performance Scale and Raven's matrices) versus 3 to 4 points on verbal tasks (Kaufman & Lichtenberger, 2006,

tables 2.4-2.6). That nonverbal-verbal difference of about 3 IQ points, based on multiple data sets from an array of nations (Flynn, 1987, 1998), is tiny compared to the whopping fluid–crystallized difference that he now claims to be true (Flynn, 2007, 2009b). He points to 54-year gains of 23.8 points (Similarities) and 27.5 points (Raven's) as the magnitude of "fluid" gains and contrasts that to half-century "crystallized" gains of 2.9 points, on average, for Vocabulary, Information, and Arithmetic. When converted to a common metric, that translates to a fluid gain of 4.5 to 5 points per decade versus about a 0.5-point gain for crystallized tasks.

And none of these conclusions about fluid versus crystallized gains across generations address McGrew's (2010) cogent arguments that (a) Wechsler's traditional Performance subtests are not very good measures of fluid reasoning, and perhaps don't even tap into fluid thinking at all; and (b) for generations—from the earliest research on the Wechsler–Bellevue to the latest factor analyses of the WISC-IV and WAIS-IV—Similarities loads on verbal/crystallized factors, *not* nonverbal/fluid factors, despite assertions by Flynn (2007) and Ceci and Kanaya (2010) that Similarities is primarily a measure of fluid reasoning.

McGrew (2010) accuses me of inadvertently reinforcing and perpetuating Flynn's incorrect interpretation of the Similarities subtest and Performance IQ "as measures of the novel abstract problem solving or on-the-spot thinking that characterizes fluid intelligence." To my dismay, McGrew is absolutely correct. I was so intent on challenging Flynn's (2007) *theory*, which was dependent on the huge generational gain that he observed for Similarities, that I ignored the incorrect *assumptions* that Flynn made about Similarities and Performance subtests such as Picture Arrangement and Block Design. I was wrong. Nonetheless, the main point of my apples–oranges article does not depend on the cognitive ability that Similarities measures. Even if Flynn had been correct that Similarities assesses novel problem solving, his computation of a 23.8-point gain on this task from 1947 to 2002 is flawed and uninterpretable. So, too, are half-century gains on other subtests, such as Comprehension and Picture Arrangement, because of the enormous administration and scoring changes that were instituted when the WISC was transformed into the WISC-R in 1974. As I stated, and continue to maintain,

> *What the huge changes did do—especially in the administration and scoring of several WISC subtests—is make it unfeasible to compare the subtest performance from one generation to another when the 1949 WISC is used as the anchor for those comparisons.* (Kaufman, 2010)

## Is It Meaningful to Interpret Generational Gains on WISC Similarities?

Flynn (2010) is the only one of the four respondents who disagrees with my basic premise that it is bad science to interpret generational gains on Similarities as meaningful in light of the subtest modifications.

- Sternberg (2010) states,

> As Kaufman points out, what appear to be small changes in test directions can change the psychological construct a test measures . . . We should not assume that because the name of a test stays the same, the actual constructs measured stay just the same as well.

- Ceci and Kanaya (2010) offer clinical, anecdotal evidence in support of my contention:

> We agree that the instructions provided to an individual can have a significant impact on his or her IQ. Indeed, one of us has had firsthand experience with this phenomenon. When testing school children on the WISC-R, it became obvious that some of the children were being penalized on Similarities because they had no idea that the conceptual answers were awarded 2 points while their own perceptual or thematic answers were given only 1 point . . . From this experience, we concluded that the scoring was almost unfair to some children, particularly those with deficits in inhibitory control and impulsivity.

- McGrew (2010) states,

> Kaufman presents convincing task analysis that the major changes in test administration and scoring between the 1949 WISC and 1974 WISC-R may have resulted in significant changes in the underlying construct measured by these two versions of the Similarities test. . . . Kaufman's apples-and-oranges analogy, albeit in the form of clinical and logical task-analysis based arguments, reinforces the problem of unknown scale equivalence, particularly for the Similarities test.

Flynn (2010) states, "Kaufman suggests that for the era of the WISC to the WISC-R (1947-1948 to 1972), my estimate of gains on three subtests may be inflated." In actuality, I simply believe that his estimates of gains are indeterminate. I have no idea if his estimates are too large or too small. I said that the scaled scores on several subtests *cannot be compared* from the WISC to the WISC-R because such comparisons are as meaningful as comparing apples to oranges. He then performed a series of analyses and summarized his results in a table in an effort to prove me wrong. He reduced the observed gains on three subtests "to accommodate [Kaufman's] thesis." However, the results of his analyses prove nothing. Neither numerical sleight of hand nor bombastic rhetoric can adequately address issues that are rooted in logic, rationality, and clinical observations. As Weiss (2010) observed, "Kaufman has wisely advised both Flynn and myself that subtle changes to instructions on the Similarities subtest may have changed the nature of the construct being measured."

Apart from the meaninglessness of interpreting scaled-score gains on tasks whose names have stayed the same (e.g., Similarities) despite changes in the basic construct it measures on WISC versus WISC-R, I also pointed out that "the changes from the WISC to WISC-R may have created a methodological problem for several subtests that prevented control of the practice effect via counterbalancing and thereby overestimated scores on the older versions of these subtests" (Kaufman, 2010). Flynn (2010) then conducted a series of detailed analyses to produce yet another table, as well as a formula, and a belabored discussion to prove me wrong. Once more, his statistical machinations fail to refute my premise, as Weiss (2010) clearly articulates,

> In these models, [Flynn] tests a series of assumptions, adjustments, and prorations of the data and determines that the hypothetical results would not be consistent with the presence of differential practice effects as predicted by Kaufman . . . By modeling a series of 'what if' assumptions, however, Flynn demonstrates that there is no direct way to solve the puzzle of differential practice effects.

Furthermore, my criticisms are pertinent to well-conducted random-design studies with large samples, the kind that have appeared in Wechsler manuals of more recent versions of the WISC and WAIS. Such investigations were not conducted with the WISC and WISC-R; instead, Flynn's analyses and digressions are based on six underwhelming studies:

> In one, the WISC-first group was superior and in the other five, the WISC-R-first group was superior. With small numbers random assignment does not equalize, although to be fair some studies just assigned a sequence of referrals to one group and then assigned the next lot of referrals to the other. (Flynn, 2010)

## Are There Caveats in Interpreting Large Generational Gains on Raven's Matrices?

I challenged Flynn's (2007) whopping 27.5-point gain on Raven's matrices in the United States over a half-century's time for two main reasons. First, Flynn's (2007) estimate was "closely tied to US gains on Similarities" (p. 8); as I have explained, Similarities gains should provide a statistical anchor for nothing. Second, I believe that time-of-measurement effects (referred to as *time lag* by Kausler, 1991) contributed to the size of the apparent gains that have been shown in some matrices studies, for example, investigations that have compared Raven IQs earned in Great Britain from 1938 to 2008 (Flynn, 2009a), Raven-like IQs earned in Norway from the mid-1950s to 2002 (Sundet, Barlaug, & Torjussen, 2004), and so forth (Flynn, 1987, 1998, 2007).

I am not talking about *practice* effects, the kind of IQ gains that occur over an interval of weeks or months simply because of the experience of having taken the same test before. Rather, I am talking about a *cohort* effect, one that affects virtually everyone who is growing up during a specific era. In the 1930s, matrices tests were largely unknown and children or adults who would have been administered such tests would have found them wholly unfamiliar. A whole society would have performed relatively poorly on such test items because of their unusualness. By the 1950s, such tests would have been known by some, not many, and by the 1990s and 2000s, matrices tests and similar item styles proliferate and are accessible to everyone. Therefore, it is feasible that people would score higher on a Raven test from one generation to the next simply because the construct measured by the test would have been a bit different from one decade to the next. Such time-of-measurement or time lag cohort effects exert powerful influences in cross-sectional and longitudinal studies of IQ and aging (Kaufman, 2001b; Owens, 1966) and must be controlled when evaluating true changes in ability between early adulthood and old age.

These time lag effects include *both* instrumentation and real FE gains in IQ. It is the instrumentation aspect of cohort effects that needs to be controlled in FE studies to determine which aspect of the gain is "real" and which aspect concerns the familiarity of the test. However, this key variable has not been controlled in FE studies making it hard to know how much of the generational gain on the Raven's test reflect increased ability. Ceci and Kanaya (2010) rejected my arguments based on their interpretation of simple practice effects, and use Zhou et al.'s (2010) results to support their position. That is not at all the same thing as time lag and their rationale does not refute my argument about uncontrolled cohort effects.

McGrew (2010) presents the fascinating finding that a difference in the approximate size of the FE was observed on the WJ III when the norms were adjusted to reflect current Census statistics. IQs earned by the *same individuals* were compared using the new and old norms. What might have been interpreted as a gain in intelligence (had different groups of people been tested at different points in time) was nothing more than a difference in population sampling. The Census distribution on key background variables is, therefore, another cohort variable that has gone uncontrolled in FE studies and may account for some of the gain attributed to growth of fluid abilities. That does not make the gain less real in terms of 3 IQ points per decade in the United States. But it does affect how one interprets the meaning of that gain.

Ceci and Kanaya (2010) cite data and facts to support the increasing emphasis on abstract thinking from infancy through adulthood in contemporary society. They are concerned that my

objections to Flynn's theory "may not be strong enough to invalidate it." They may be correct and Flynn may be correct. But the answer resides in the data, not in speculation or in research that is tangential to the FE. Ceci and Kanaya (2010) state, "The developmental psychology literature . . . has clearly and consistently shown that small environmental changes can lead to long-term cognitive growth in children even younger than the WPPSI testing age range" and that "today's newborns are . . . living a life with higher fluid demands than those from a generation ago." Those statements are undoubtedly true, but the most pertinent FE data argue otherwise—namely that infants and toddlers demonstrated a *reverse* FE of 5.8 points on the Bayley mental scale in a counterbalanced study of 102 infants and toddlers aged 1 to 42 months tested on the Bayley-II and Bayley-III (Bayley, 1993, 2006) over a 1-week interval (Yang, Zhu, Pinon, & Wilkins, 2006).

## Interpretation of Single Subtests

My biggest point of contention with Flynn (2007, 2009b) is that his current line of research emphasizes interpretation of single subtests. Even his insightful discussion of generational changes on Wechsler Vocabulary (Flynn, 2010) is a one-subtest argument. Texts on the interpretation of clinical tests of intelligence written in the 1960s (e.g., Allison, Blatt, & Zimet, 1968; Glasser & Zimmerman, 1967) emphasized the interpretation of each subtest's unique abilities and inexplicably deemphasized the group factors that had been identified more than a decade earlier by Jacob Cohen (e.g., Cohen, 1952). The texts from the 1970s shifted the focus from single subtests to the abilities and processes that underlie groups of subtests (Kaufman, 1979; Matarazzo, 1972; Sattler, 1974). That shift continues to the present and has been the guiding force behind theory-driven cognitive tests like the WJ III (Woodcock, McGrew, & Mather, 2001), Binet-5 (Roid, 2003), KABC-II (Kaufman & Kaufman, 2004), DAS-II (Elliott, 2007), and CAS (Naglieri & Das, 1997). It has led to four-factor Wechsler Scales, rooted in cognitive neuroscience research, forcing the retirement of the Verbal and Performance IQs.

Emphasizing the unique abilities measured by single subtests like Similarities and Vocabulary represents a return to the interpretive approach advocated in the 1960s when a low score on Picture Completion was invariably interpreted as a difficulty in "distinguishing essential from nonessential details." And whereas Flynn (2007, 2009b) has attempted to combine Similarities with Raven's Matrices and the Wechsler Performance Scale to buttress his fluid reasoning arguments, these tasks make strange bedfellows, as McGrew (2010) has aptly pointed out. Much of Flynn's argument is tied to the unique logical–analogic ability supposedly measured by Similarities (see, e.g., Chapter 2 of *What Is Intelligence?*) and to the alleged 24-point gain on that subtest. More generally, that unique ability fits under the rubric of *verbal reasoning*—a set of skills also measured in slightly different ways by the Crystallized (*not* fluid) subtests of Comprehension, KABC-II Riddles, Binet-5 Verbal Absurdities, and WJ III Verbal Comprehension. McGrew, following Hunt (2000), refers to verbal reasoning as the ability to "*apply culturally approved, previously acquired problem solving methods*" (McGrew, 2010).

Regardless of the name of this aspect of crystallized thinking, it is better to conduct research and build theories on abilities or processes that are shared by groups of subtests—as supported by both theory and factor analysis—than to resort to the interpretation of abilities unique to a single subtest. Consistent with this view, McGrew (2010) predicts that "we will eventually learn that the FE is a differential CHC-ability effect . . . [and] that the global composite FE findings will be found to have masked differential CHC-ability changes across generations." Weiss (2010) echoes a similar "non-CHC" belief: "The rate of gain may be different for different cognitive abilities . . . Little is known about rates of gain for processing speed or working memory tasks."

## Should IQs Be Adjusted for the Flynn Effect in Capital Punishment Cases?

We originally asked several authors to write an essay entitled, "Should IQs Be *Corrected* for the Flynn Effect in Capital Punishment Cases?" We changed "corrected" to "adjusted" based on Leigh Hagan's (personal communication, February 24, 2010) legitimate concern that "the term 'corrected' implies that the obtained score is wrong." Clearly, based on the three essays on capital punishment and Weiss's (2010) discussion of the topic, there is no universal consensus on whether the obtained IQ is wrong when norms are out of date. The opinions expressed are the following.

- Hagan, Drogin, and Guilmette (2010) conclude,

> An all-inclusive declaration about "every individual" does not . . . adequately acknowledge the probabilistic nature of group data and potential inconsistency when applied to individuals. . . . 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**" (bold print added)

- Fletcher, Stuebing, and Hughes (2010) state,

> **IQ test scores should be adjusted for high-stakes decisions that employ these assessments, including capital offense cases** (bold print added). If scores are not adjusted, then diagnostic standards must change with each generation.

- Reynolds, Niland, Wright, and Rosenn (2010) argue,

> 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 . . . **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** (bold print added).

- Weiss (2010) does not take a specific position on the death penalty issue, claiming that,

> **the ethical position of an expert witness providing testimony to the court is not to argue either for or against FE adjustments** (bold print added) but to inform the court about the extant research on the topic.

I respect the diversity of opinion on the topic of capital punishment, and the statistical complexity that surrounds the FE, but I am firmly in the camp with Reynolds et al. (2010), Fletcher et al. (2010), and Flynn (2006, 2007, 2009b) that IQs obtained on outdated norms should be adjusted for the FE in capital punishment cases. Hagan et al. (2010) argue that group data should not automatically be applied to specific individuals, and Zhou et al. (2010) concur, based on their analyses of Wechsler's nonverbal scales:

> The magnitude of the FE . . . varies across ability groups . . . Overall, our findings suggest that the average IQ gain Flynn initially described may only be valid as an aggregated phenomenon. The variation by ability group we demonstrated implies that adjusting an

individual observed IQ by a fixed rate obtained from the overall sample may yield systematic over or under estimates of IQ depending on the individual's ability level.

The fact that the FE differs by ability level was supported by Zhou et al.'s analyses and other research (see Ceci & Kanaya, 2010). However, the bulk of evidence suggests that the FE is at least 3 points for IQs in the range associated with Intellectual Disability. Had the issue been whether or not to adjust IQs for the 3-points-per-decade FE for gifted placement, I probably would have been swayed by the Zhou et al. data. But for life-or-death decisions at the opposite end of the spectrum I believe the FE adjustment is warranted. Perhaps 3 points is a conservative estimate for individuals who score below 70, but what makes more sense if the FE is ultimately shown to be 4 or 5 points for low-functioning adults—to subtract 3 points or to subtract 0 points? I agree with the question posed by Reynolds et al. (2010)—"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?"

I am also unmoved by the group-to-individual argument that Hagan et al. (2010) make against adjusting IQs for the FE, even though this same perspective is shared by (a) Sternberg (2010)—"the FE seems to apply in the aggregate, but it is extremely difficult to apply it in individual cases," and (b) Ceci and Kanaya (2010)—"it is not appropriate to merely subtract 0.3 points for every year that a norm has aged until we know that everyone experiences the same gains on the same subtests and at the same time."

I have made this same group-to-individual argument concerning the inadvisability of applying group data to specific individuals on the relationship of blood lead level to IQ loss (Kaufman, 2001a). That is a different case. Children have different susceptibilities to toxins; the so-called loss of 3 IQ points for children with low levels of blood lead certainly does not generalize to all children. In addition, the lead-IQ studies included in the various meta-analyses varied widely in their results and often found no significant relationship at all. But with the FE, in the United States, a multitude of studies converge around 3 points per decade (Flynn, 2007, 2009b, 2010; Zhou et al., 2010) and the variability around 3 points is relatively minor (Fletcher et al., 2010).

Furthermore, the FE studies are not about the individual, they are about the group. In the best FE studies, counterbalanced random designs, there will be much individual variability in the IQs earned by the same individual on the two versions of the test—some will obtain differences much greater than 3 points and others will show a reverse FE. Such variability is because of multiple factors such as the practice effect and errors of measurement. However, in real-life situations where the application of the FE is of concern, people have one IQ that is the issue, not two, and the only question is the datedness of the norms for that particular score.

In the FE research studies, the goal is to determine the degree of outdatedness of the old norms relative to the new norms. The group data from the counterbalanced studies apply directly to the groups of individuals who constitute the normative samples. In effect, it is a *group-to-group* application, not a *group-to-individual* application. IQs are relative, not absolute, scores. What does it mean when an adult of 35 defines 18 words or solves 7 block designs correctly? It means nothing until it is compared with an appropriate reference group of normal 35-year-olds. When the FE is applied, to make the reference group as relevant as possible based on the degree to which the test norms were outdated at the time of the evaluation, the test scores do not change—the person still defined 18 words and solved 7 designs—but the interpretation of those scores is made more accurate by providing a better yardstick. As Fletcher et al. (2010) point out, "individual scores are not being adjusted; rather, the validity studies are used as a basis for selecting an appropriate normative comparison group."

Hagan et al. (2010) cite research to show that most doctoral programs "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." They further maintain that, "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." Regarding the "it-is-not-common-clinical-practice" argument, I would maintain that the FE has had such a striking impact on clinical practice that it is now unnecessary to teach graduate students to adjust IQs for the FE. As I have previously written (Kaufman, 2009), "One positive outcome of the Flynn Effect is that it has made test publishers more accountable. Historically, they were lazy about revising and restandardizing a test . . . But the Flynn Effect changed all that" (p. 211). The Wechsler scales once were revised every 25 years or so, and now the interval is less than half that. Indeed, virtually all major cognitive batteries have been renormed or undergone a normative update since 2003, so teaching students or other professionals to adjust for the FE is redundant in most clinical situations. Simply teach them to use the test with the most current norms—and that does reflect common clinical practice.

Regarding Hagan et al.'s point about the lack of a legal requirement to apply the FE in death penalty cases, that topic is currently sweeping through the courts on a nationwide basis and the decision is up for grabs. Is the FE a valid scientific construct? Should it be applied to determine a capital criminal's fate? Reynolds et al. (2010) make some pertinent points:

> 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 . . . 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 [that] "the penalty of death is different in kind from any other punishment imposed under our system of criminal justice."

Sternberg (2010) challenges the methodology for evaluating intelligence and suggests that even his own triarchic theory of successful intelligence may require an additional "arch," namely ethical intelligence. He also expresses concerns about the FE and IQ tests in general:

> Both these articles—by Kaufman and by Zhou et al.—render frightening the use of the FE or even of tests such as the WAIS as a basis for establishing competency in criminal proceedings. With the stakes in such cases so high, can we really put enough faith into levels of IQ scores to draw sound conclusions? (Sternberg, 2010)

But what other choice do we have than to deal with the present system and make the best scientific decisions within that system? With the stakes so high, and life-and-death decisions to be made, Sternberg's probing question may be theoretically driven—and it surely reflects high-level analytic thinking and creativity—but it is not practical. Clinical judgments and court decisions will continue to be made based on the WAIS Full Scale IQ, with all its imperfections. And the FE—with all of *its* imperfections—cannot simply be dismissed.

## Conclusions

There is controversy about what causes the FE. There are data to suggest that shifts in population statistics account for part of the gain in IQ scores (McGrew, 2010); other data suggest that the FE is a partly a function of the type of score that is analyzed (Beaujean & Osterlind, 2008; see McGrew, 2010). But even if FE is due primarily to methodology, as opposed to a real gain in IQ,

it makes no difference for the scientific validity of the FE or the strong recommendation by myself, Flynn, Reynolds, Fletcher, and others to adjust the IQs in death penalty cases. The point is that a person tested on an outdated test will earn spuriously high scores as each year goes by, and that amount of spuriousness amounts to about 3 points per decade for Americans. Witness the powerful real-life findings reported by Ceci and Kanaya (2010)—"that the number of children who were diagnosed with mental retardation (MR) nearly tripled on the introduction of the WISC-III as more and more children obtained an IQ of 70 points or below on the newly introduced, harder norm." The FE 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. "If the Flynn effect 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" (Reynolds et al., 2010).

Apart from capital punishment issues, it is time for professionals to be more critical of existing research findings on the FE, as well as potential explanations of the effect. A few studies are conducted in Scandinavian countries (e.g., Sundet et al., 2004; Teasdale & Owen, 2005) and there is a general consensus that there is a reverse FE in Denmark and Norway. These studies of 18- to 19-year-old males did not include females, children, or men older than 20 years, yet conclusions are reached about entire *countries.* The black box behind the FE, as well as its worldwide magnitude by ability level and ethnicity, demands an innovative research program. The kinds of top-notch methodological designs articulated by Rodgers (1998), Beaujean and Osterlind (2008), Ceci and Kanaya (2010), and McGrew (2010) need to set the tone for an in-depth understanding of the FE that moves the field well beyond the controversies that form the basis of this special issue. The case is not nearly closed, despite Flynn's (2010) belief that he can see with perfect clarity through his "scientific spectacles."

## Acknowledgment

I am grateful for the insightful comments made by Jennie Kaufman Singer, PhD, on the three capital punishment articles, on earlier drafts of the present article, and on several other articles that appear in this special issue. Dr. Singer is a clinical psychologist and Assistant Professor, Division of Criminal Justice, Sacramento State University.

## Declaration of Conflicting Interests

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

## Funding

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

## References

Allison, J., Blatt, S. J., & Zimet, C. N. (1968). *The interpretation of psychological tests*. New York, NY: Harper & Row.

Bayley, N. (1993). *Bayley scales of infant development–Second edition*. San Antonio, TX: The Psychological Corporation.

Bayley, N. (2006). *Bayley scales of infant and toddler development–Third edition. Technical manual*. San Antonio, TX: Harcourt Assessment.

Beaujean, A., & Osterlind, S. J. (2008). Using item response theory to assess the Flynn effect in the National Longitudinal Study of Youth 79 Children and Young Adults Data. *Intelligence, 36*, 455-463.

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

Cohen, J. (1952). A factor-analytically based rationale for the Wechsler-Bellevue. *Journal of Consulting Psychology, 16*, 272-277.

Elliott, C. D. (2007). *Differential Ability Scales—Second edition (DAS-II)*. San Antonio, TX: The Psychological Corporation.

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

Flynn, J. R. (1998). IQ gains over time: Toward finding the causes. In U. Neisser (Ed.), *The rising curve: Long-term gains in IQ and related measures*. Washington, DC: American Psychological Association.

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

Flynn, J. R. (2009a). Requiem for nutrition as the cause of IQ gains: Raven's gains in Britain 1938-2008. *Economics and Human Biology, 7*, 18-27.

Flynn, J. R. (2009b). *What is intelligence?* (Expanded Edition). New York, NY: Cambridge University Press.

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

Glasser, A. J., & Zimmerman, I. L. (1967). *Clinical interpretation of the Wechsler Intelligence Scale for Children (WISC)*. New York, NY: Grune & Stratton.

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.

Hunt, E. (2000). Let's hear it for crystallized intelligence. *Learning and Individual Differences, 12*, 123-129.

Kaufman, A. S. (1979). *Intelligent testing with the WISC-R*. New York, NY: Wiley.

Kaufman, A. S. (2001a). Do low levels of lead produce IQ loss in children? A careful examination of the literature. *Archives of Clinical Neuropsychology, 16*, 303-341.

Kaufman, A. S. (2001b). WAIS-III IQs, Horn's theory, and generational changes from young adulthood to old age. *Intelligence, 29*, 131-167.

Kaufman, A. S. (2009). *IQ testing 101*. New York, NY: Springer.

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.

Kaufman, A. S., & Kaufman, N. L. (2004). *Kaufman Assessment Battery for Children–Second edition (KABC-II)*. Circle Pines, MN: American Guidance Service.

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

Kausler, D. H. (1991). *Experimental psychology, cognition, and human aging*. New York, NY: Springer.

Matarazzo, J. D. (1972). *Wechsler's measurement and appraisal of adult intelligence* (5th and enlarged ed.). New York, NY: Oxford University Press.

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.

Naglieri, J. A., & Das, J. P. (1997). *Cognitive assessment system.* Itasca, IL: Riverside.

Owens, W. A. (1966). Age and mental ability: A second adult follow-up. *Journal of Educational Psychology, 57*, 311-325.

Reynolds, C. R., Niland, J., Wright, J. E., & 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*, 477-481.

Rodgers, J. L. (1998). A critique of the Flynn effect: Massive IQ gains, methodological artifacts, or both? *Intelligence, 26*, 337-356.

Roid, G. (2003). *Stanford-Binet Intelligence Scales–Fifth edition.* Itasca, IL: Riverside.

Sattler, J. M. (1974). *Assessment of children's intelligence* (Rev. ed.). Philadelphia, PA: W. B. Saunders.

Sternberg, R. J. (2010). The Flynn effect: So what? *Journal of Psychoeducational Assessment, 28*, 434-441.

Sundet, J. M., Barlaug, D. F., & Torjussen, T. M. (2004). The end of the Flynn effect? A study of secular trends in mean intelligence test scores of Norwegian conscripts during half a century. *Intelligence, 32*, 349-362.

Teasdale, T. W., & Owen, D. R. (2005). A long-term rise and recent decline in intelligence test performance: The Flynn effect in reverse. *Personality and Individual Differences, 39*, 837-843.

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

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III.* Itasca, IL: Riverside.

Yang, Z., Zhu, J., Pinon, M., & Wilkins, C. (2006, August). *Comparison of the Bayley–III and the Bayley–II*. Paper presented at the annual meeting of the American Psychological Association, New Orleans, LA.

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.