Considerations on the Flynn Effect

Lawrence G. Weiss

Journal of Psychoeducational Assessment 2010 28: 482 originally published online 7 July 2010
DOI: 10.1177/0734282910373572

Considerations on the Flynn Effect

Journal of Psychoeducational Assessment
28(5) 482–493

DOI: 10.1177/0734282910373572
http://jpa.sagepub.com

Lawrence G. Weiss[1]

Abstract

Flynn has proposed a grand integrative theory, which he calls "scientific spectacles," to explain the phenomenon of rising IQ scores across multiple decades known as the Flynn effect (FE). In his theory, he purports that modern society has placed increasing value and emphasis on the application and education of scientific principles—which include abstract fluid reasoning—and this has mirrored larger score increases in the abstract fluid reasoning tasks included as part of most intelligence tests. He highlights huge gains in the Wechsler Similarities subtest over time as one linchpin supporting his theory that fluid reasoning is rising faster than other aspects of intelligence, but further points to large increases in Wechsler performance tasks (i.e., perceptual organization and reasoning) relative to crystallized tasks, and to the many studies of Raven's Progressive Matrices, which is widely held to be a fairly pure measure of fluid reasoning. As with any new theory of this magnitude, Flynn's grand proposal has invited numerous criticisms that must now be addressed—and this is as it should be in the spirit of scientific progress. Each of these issues is multilayered and has many legitimate perspectives. The present article attempts to carefully consider each issue and perspective in sequence.

Keywords

Flynn effect, capital punishment, methodology, theory

Construct Considerations

In this issue, Kaufman challenges Flynn's (2007) scientific spectacles theory as the primary cause of the Flynn effect (FE) by questioning the reality of the huge reported gains in Similarities. As one of very few, if not the only, person who can recall working on the revision of WISC to WISC-R, Kaufman has wisely advised both Flynn and me (Flynn & Weiss, 2007) that subtle changes to instructions on the Similarities subtest may have changed the nature of the construct being measured. Hence, Kaufman's (2010) analogy that comparing WISC and WISC-R Similarities scores is like comparing apples and oranges.

This is essentially a question of what cognitive response processes are required to complete the task and if the change in administration instructions invokes substantially more (or less) abstract reasoning for enough people to shift the norms. With the possible exception of brain imaging studies, cognitive response processes are difficult to observe directly and, like all latent

[1]Pearson Clinical Assessment, San Antonio, TX, USA

Corresponding Author:
Lawrence G. Weiss, Pearson Clinical Assessment, 19500 Bulverde Road, San Antonio, TX 78259-3701, USA
Email: larry.weiss@pearson.com

traits, must be inferred. If the changes to Similarities instructions were sufficient to change the nature of the construct being measured, then one would expect to see Similarities "behaving" differently across versions in terms of the pattern of convergent and divergent correlations with other tasks—a point which none of the articles in this issue directly tackle. The different versions of Similarities would need to be shown to occupy somewhat different spots in the nomological network of cognitive tasks, or load differently on factor analyses. Until such evidence is forthcoming argumentation that these tasks invoke different cognitive constructs in the brain are unconvincing.

Kaufman (2010) provides an important service by reminding FE researchers not to take for granted the comparability of scores across versions, and this advice will be increasingly important as new FE analyses include the fourth edition of the Wechsler tests, which have undergone substantial changes designed to increase the amount of fluid reasoning and working memory. Not only have new subtests designed to measure fluid intelligence (e.g., Matrix Reasoning) been added to the performance/perceptual reasoning scale, but also, more to Kaufman's point, subtle changes within subtests were designed to shift the mix of response processes invoked by the task. As one example, the WAIS-IV Digit Span subtest has added a new sequencing component to the traditional digits forward and backward tasks that have been used in all previous versions. This new component was designed to invoke increased demands on working memory relative to short term memory. Thus, the comparability of WAIS-III to WAIS-IV Digit Span scores could involve some amount of "mixed fruit" and measurement invariance would need to be investigated before comparing or combining these scores.

Ceci and Kanaya (2010) argue that improved instructions on Similarities, which prompt test takers for more abstract responses, clearly aid some impulsive subjects and account for some but not most of the score increases. This is a good point in that impulsive individuals are by definition a minority of the general population, yet there is no reliable way to know to what extent the extra coaching changes the response processes of normal subjects. In this issue, Ceci and Kanaya argue that Kaufman's objections are not sufficient to negate the cumulative findings relating increases in a variety of fluid tasks to the FE. Clearly, the issues surrounding Similarities must be viewed in perspective of the larger body of FE literature. As Ceci and Kanaya note, tasks that load higher on fluid reasoning (e.g., Wechsler Performance/Perceptual Reasoning subtests and Raven's Progressive Matrices) show a larger FE than tasks that load higher on crystallized abilities and there are few, if any, exceptions to this finding. These authors are careful to avoid describing measures as categorically fluid or crystallized and describe them instead as loading higher in one area than another. This is an important point that will be returned to later.

McGrew (2010) takes Kaufman's criticism of Flynn's scientific spectacles theory to a deeper level by questioning if Similarities indeed measures fluid reasoning, and further, by challenging the basic assumption held by most if not all FE researchers that the Wechsler Performance/Perceptual Reasoning subtests are good measures of fluid reasoning or even of psychometric "g." McGrew recategorizes the Wechsler subtests in terms of fit to the Cattell–Horn–Carroll (CHC) model of intelligence (which forms the basis of the Woodcock–Johnson test) and reports construct underrepresentation in the Wechslers. As has been elaborated elsewhere, however, the Wechsler third and fourth editions are not based on CHC theory, but rather, on current research in cognitive neuroscience (Weiss, Saklofske, Coalson, & Raiford, 2010). McGrew further examines the g loadings of Wechsler subtests when factor analyzed together with subtests from the Woodcock Johnson, Kaufman, and ITPA (Illinois Test of Psycholinguistic Abilities) tests and finds that other tasks have higher g loadings than the Wechsler tasks. McGrew equates high g loadings to fluid reasoning and criticizes Kaufman (2010) for stating that Similarities taps fluid intelligence, and Zhou, Zhu, and Weiss (2010) for describing the Wechsler Performance/Perceptual subtests as measures of fluid intelligence.

Overlooking the blog-like language (e.g., "rusty" and "corroded" Wechsler subtests), McGrew contributes meaningfully to the current discussion by emphasizing the importance of examining measurement invariance between versions before comparing them. However, his assumption that high g loadings indicate a task is fluid—although not without precedent—might require more scrutiny. Using this assumption and McGrew's data would lead one to conclude that Vocabulary—which is universally agreed to be a measure of crystallized knowledge—is a good measure of fluid reasoning because the amount of shared variance with the underlying g component (h^2) is 55%, approaching the value of 61% for Raven's Progressive Matrices (RPM; see sample 1b, McGrew, 2010). Further, the g loadings for Vocabulary are rated as high in six of the seven samples analyzed by McGrew, and it has higher g loadings than the WAIS-IV Matrix Reasoning subtest (a Raven's like task), which McGrew and others classify as fluid.

Flynn (2010) notes that vocabulary knowledge (as measured by the Wechsler Vocabulary subtest) appears to be progressing faster across generations for adults than children, and spends considerable time speculating on the societal causes for this finding. Regardless of the possible causes, the finding is interesting in light of the high g loadings observed for Vocabulary in the McGrew data and numerous other sources. As has been discussed elsewhere (Weiss et al., 2010), although considered to be a form of crystallized knowledge, vocabulary scores reflect more than words taught in school as they appear to be reciprocally interdependent with several other cognitive abilities. Language development and general cognitive development have been historically difficult for researchers to untangle in infants and toddlers as words initially give us the ability to describe the world around us, and eventually the building blocks to reason with abstract verbal material. While exposure to vocabulary words clearly plays a critical moderator role in language development, an advanced vocabulary also reflects an advanced ability to understand higher level concepts. For adolescents and adults, the ability to represent in mind a complex idea or set of ideas with a single word allows more efficient use of limited working memory space for fluid reasoning and problem solving. Further research into the dynamic interplay between vocabulary and other cognitive processes may help explain the finding that the Wechsler Vocabulary subtests share almost as much variance with the underlying g component as do matrices tests such as the RPM, which is considered by many to be a proxy for general intelligence and the purist measure of fluid reasoning presently available.

For any cognitively complex task the issue of construct definition is not categorical, and can better be framed in terms of how fluid it is in relation to the other tasks to which it is compared. However, all such nomological networks are essentially relativistic approaches where results can vary based on the model inputs—as demonstrated by McGrew's data. The search for a pure marker of fluid reasoning has been elusive because one cannot reason in a vacuum—some form of stimuli must be presented to the subject and various other abilities are necessarily invoked as the stimulus is perceived. Even the Raven's task requires perceptual organization, visual–spatial reasoning and sequencing abilities—not to mention secondary loadings on working memory. Perhaps future researchers should look toward brain lesion mapping technology to clarify factor analytic studies of cognitive response processes linked to intelligence test tasks; indeed, some researchers are already beginning to conduct such studies (Glascher et al., 2009). However, at present, there is insufficient overlap between the psychometric and neurological literatures with respect to the structure of intelligence.

None of this discussion is intended to suggest that the Vocabulary subtest measures fluid reasoning. Rather, the goal is to illustrate how psychometric analyses of "g" loadings and factor structures are insufficient tools for elucidating the neurologic complexity of the human mind, and, more to the point, that Flynn's attempt to interpret subtests in isolation can be misleading because of the reciprocal ways in which narrow abilities support one another during higher order

problem solving. Building theories about society induced changes in intelligence based on single subtests may not prove sustainable.

Methodological Considerations

Quite apart from the debate about constructs, Kaufman (2010) raises an important methodological point when he hypothesizes that the changes in instructions (i.e., more coaching for higher-level abstract answers on WISC-R) may have created a differential practice effect which served to inflate only one group in a counterbalanced design. Thus, the group that took WISC-R first would have an unfair advantage when administered WISC a few weeks later due the coaching received on WISC-R that was not available to members of the WISC standardization sample. This may be a more enduring concern than the apples and oranges debate about constructs shifting with changes in subtest instructions; however, there is no direct evidence that differential practice effects exist in this case.

In this issue, Flynn attempts to tackle the thorny question of differential practice effects raised by Kaufman. Flynn (2010) explores several models of what the data should look like if Kaufman were correct and differential practice effects existed. In these models, he 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 (2010). Some of these assumptions and adjustments might be debated (e.g., "Let us assume that going from the WISC-R to the WISC inflated scores by 10 points on the three subtests supposed to be susceptible," and "I have multiplied Kaufman's values by 0.67 and rounded off"). 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. Most researchers consider differential practice effects to be fatal flaws in counterbalanced research designs. This is an important point because so many FE studies are based on this design.

Overall, there has been too much reliance on counterbalanced studies among FE researchers—largely because they are conveniently provided by test authors and publishers who must inform practitioners how test scores differ across versions. However, the constant press to improve these tests for clinical practice based on ongoing theory and research, and the concomitant changes in constructs, create some limits on the application of counterbalanced cross-sectional studies for estimating longitudinal change in intelligence across generations. If differential practice effects are suspected, perhaps future researchers might consider matched control designs in which two samples of current examinees equivalent on educational level and other relevant demographics are assigned to take either the old or new version, not both. To guard against a single poorly matched sample, the sample taking the current version can be repeatedly drawn from the much larger standardization sample and the results averaged. Alternatively, the samples could be matched on an independent measure of ability.

Studies designed specifically to evaluate the amount of change in IQ across generations would be improved by use of the same measure; yet, this approach would eventually place FE research out of step with advances in the conceptualization of intelligence. If different versions will continue to be used in this line of study then FE researchers must begin to use more sophisticated approaches. Previous methods used by Flynn and Weiss (2007) involved (a) using FE theory to estimate how much a deleted subtest would have increased the IQ score if it had been retained in the new version and (b) dropping new subtests from FE analyses on the grounds that they were not included in the older version. These methods are partially self-fulfilling FE prophesies and should be used with caution, if at all. More sophisticated approaches are available and should be employed. One possible approach would be common item equating of earlier and later editions for those subtests where the administration procedures have not meaningfully changed. This

approach requires sufficient data in the tails of the sample distributions to adequately inform equating of extreme scores. Another possible approach is measurement invariance, which employs joint sample factor-analytic techniques to examine the invariance of the construct across samples and the differences in latent means. Bowden, Lange, Weiss, & Saklofske (2008) and Georgas, Weiss, van de Vijver, and Saklofske (2003) provide examples of methods used to compare construct equivalence and test norms across national populations. The methodology employed by Wicherts et al. (2004) is also notable. Through newer types of analyses such as these, the field of FE research can be placed on more solid methodological footing.

Empirical Considerations in Theory Building

Zhou et al. (2010) and Weiss (2007) criticize Flynn (2006) for stating that WAIS-III norms are substandard solely because they do not conform to predictions based on FE theory for the amount of increase from WAIS-R, and for recommending that psychologists adjust obtained WAIS-III scores for the substandard norms in high-stakes evaluations. In a postscript to his article in this issue, Flynn defends this position citing two samples reported by Floyd, Clark, and Shadish (2008). However, both of these data points appear to be from the same narrow and atypical sample of college students, half of whom have a learning disability or ADHD, and which is unrepresentative of the general population. As pointed out previously (Weiss, 2007), data cited in Flynn's (2006) article show that the average yearly difference between Stanford–Binet (SB)-4 and SB-5 scores is identical to the difference (0.17 points per year) reported between WAIS-R and WAIS-III scores (see Flynn, 2006, table 1). Thus, the SB-5 normative data provide a cross validation of the WAIS-III norms using independent research groups and sampling methods—as suggested by Flynn in his postscript.

For general clinical practice, Flynn's concerns about the WAIS-III norms became mute in 2008 with the publication of the WAIS-IV, and remain of interest only to FE researchers and forensic psychologists who may testify in court cases involving archival WAIS-III scores. Yet the general point about the appropriateness of adjusting norms to fit FE predictions remains important for future practice because Flynn proposes use of a matrix of previous FE effects to rate the degree of "eccentricity" of any newly published IQ test. Fletcher, Stuebing, and Hughes (2010) contribute, perhaps unintentionally, to this line of thinking by eliminating two WAIS-III studies as outliers based in part on Flynn's criticism of the WAIS-III, and then reporting that the meta-analytic mean size of the FE across the remaining 14 studies is similar to Flynn's unweighted mean.

All major publishers of IQ tests currently follow best practices in sampling techniques. If the next edition of the Wechsler, Kaufman, Woodcock–Johnson, Stanford–Binet, Raven, Reynolds, Das–Naglieri, or other intelligence test is judged substandard by Flynn solely because it does not conform to the theory that bears his name, and adjustments are made to force the normative scores to fit FE theory, then FE theory would become a closed, self-reinforcing system that cannot be tested or disproved and cannot evolve based on an accumulation of reliable new evidence. In this regard, Kaufman (personal communication, June 6, 2009) and McGrew (2010) are in agreement with Zhou et al. (2010) and Weiss (2007). The FE is not a law of nature—like the boiling point of water—to which all new measuring instruments must be properly calibrated, else rejected as faulty.

This is a different issue than adjusting scores for obsolete tests. This distinction is important. For example, Fletcher et al. (2010) note that Weiss (2007) has been misinterpreted by Hagan, Drogin, and Guilmette (2008) as opposing FE adjustments in general when that post was actually a response to Flynn's charge that the then current WAIS-III norms were substandard. Fletcher and colleagues are correct. I oppose adjusting current IQ test scores if based solely on the grounds

that the norms do not align with FE predictions. Adjusting scores of obsolete tests in high-stakes evaluations is another matter that will be discussed below.

Peeking Inside Flynn's Black Box

In this issue, Zhou et al. explore several methodological issues in the examination of possible differential FE by ability level using two statistical approaches with different sets of assumptions—analysis of covariance (ANCOVA) and equal percentile equating. The ANCOVA analysis allowed the four samples to be combined and nonsignificant trends suggested that the rate of change in Performance IQ (PIQ) scores may be larger at the middle and lower portion of the distribution and smaller in the upper portion of the distribution of scores. The equating analysis was necessarily conducted on the four samples separately, and suggested larger rates of change in the lower portions of the distribution for two of the samples (WPPSI-R/WPPSI-III and WAIS-R/WAIS-III), but larger rates of change in the upper portion of the distributions for the other two samples (WISC-III/WISC-IV and WAIS-III/WAIS-IV). Thus, the results of the ANCOVA and equating analyses may lead to different conclusions. McGrew added construct-level concerns (see above) to the list of methodological issues discussed by Zhou et al.

Reynolds, Niland, Wright, and Rosenn (2010) assessed the strengths and weaknesses of each method and concluded that "the larger sample size and 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." Further, and as noted above, the equating method requires large enough sample sizes to adequately equate low frequency scores in the tails. At the same time, however, the ANCOVA method adds extraneous error variance related to variables intervening between the two testing periods. Further studies are needed to resolve these methodological issues.

Zhou, Gregoire, & Zhu (2010) extended the current findings to include FSIQ. Studies that explore Flynn's black box are important because they can be used to advance several different research questions including the causes of the effect, the size of the effect for different groups, and whether adjustments can be applied to individuals based on group membership. These issues are considered below.

Theoretical Considerations

Except for Flynn, there is general agreement among the contributors to this special issue that we know precious little about the causes of the effect. Yet the temptation to speculate is almost irresistible and convincing rationales readily can be built based on logic and historical correlations. For example, recently, Steen (2009) has written a compelling and insightful analysis of how medical illnesses and cures correlate with gains in human intelligence over time. Furthermore, several authors are creating trendy theories based on idiosyncratic views of society and history (increasing use of computer games, improvements in nutrition, changes in education policy, etc.) and then mining the data for isolated bits of evidence (e.g., single subtests) to support them. The field does not need further armchair speculation on the causes of FE.

Although not writing about FE research in particular, Rogers (2010) asks in general if researchers are "putting the heart before the course" (p. 9). Theoretical knowledge in the field is best advanced, he argues, by first developing competing models of the causal mechanisms for the effect and then systematically testing them. Further studies of group differences in rates of gain can inform hypotheses about FE causes, but would be more convincing when the hypotheses are articulated in advance of the results. Reasonable hypotheses that could be supported or refuted by analyses of FE by group membership include that higher ability people benefit more from certain

causes because they have a greater capacity to take advantage of them Alternatively, people in lower socioeconomic (SES) groups may benefit more because some causes represent societal improvements to which higher SES groups already had access—a hypothesis that seems to underlie several current theories; however, separating low-SES from low-ability groups is a methodological concern. As Ceci and Kanaya (2010) have pointed out, cognitive development depends on key environmental inputs at critical developmental stages, and so another reasonable hypothesis might be that children of a certain age benefit most from a particular type of cause.

It is natural based on their graduate training in experimental design for research psychologists to discount logical analysis and historical correlation as noncausal. Consider, though, that investigating FE causes is more akin to a combination of archeology and political science than classical experimental design and in those fields logical analysis backed up by a preponderance of circumstantial evidence can be considered scientific proof. The trick is to systematically look for evidence that should not be present if the hypothesis is true. For example, if certain groups of ancient people were thought to be hunter-gatherers, then their remains should never be found in the vicinity of farming tools and so researchers must systematically dig in farming villages to disconfirm the hypothesis. So, FE researchers also should be mining the data in search of disconfirming evidence for particular theories. Given the necessarily archeology-like nature of investigating FE causes in past generations there needs to be room for consideration of logical analysis and historical correlation, but it could be more a piori and systematic.

In investigations into causes it is helpful to keep in mind that the FE phenomenon is essentially a generational effect and that different generations experience different societal risks (e.g., war, depression, famine, racism, sexism) and benefits (periods of economic prosperity with increased educational and medical resources) that also may be differentially experienced by different subgroups of the same generation, and may or may not be permanent and experienced by succeeding generations. This focus on differential generational effects may also apply to discussions of rates of gain over time as some age cohorts may have larger or smaller rates of gain as a result of impermanent societal conditions in some countries. These differential effects are important because most discussions of rates of gain have emphasized presumably permanent societal improvements, especially in the areas of education and medicine. This means that hypotheses about FE causes need to specify the generations in which they are expected to show the effect in addition to the subgroups affected.

For example, Ceci and Kanaya (2010) suggest that researchers further examine the role of ethnicity, and especially the interaction of ethnicity and ability level. Weiss, Chen, Harris, Holdnack, and Saklofske (2010) recently reported data showing that the IQ gap between racial/ethnic groups has dramatically narrowed by 9 to 10 points for adults across successive generations between 1917 and 1991. Testable hypotheses related to interactions among ethnicity, ability level, and education level can be generated and submitted to analyses in systematic attempts to disprove particular armchair speculations about causal mechanisms. This is how empirical evidence informs theory building. As will be argued below, however, such subgroup differences may be irrelevant to adjusting individual IQ scores in forensic practice.

A Time to Kill?

Some states in the United States allow the execution of individuals convicted of murder, but not if they are intellectually deficient. The finding of intellectual deficiency in these cases often rests on historical records of IQ test scores such as when the felon was in school, and those tests may not have been current at the time they were administered. Adjusting the IQ score on record for the amount of obsolescence at the time of administration (at the rate of 0.3 points per year) results in lowering the score, sometimes below the cut-off of 70, such that the prisoner cannot be

legally put to death. The appropriateness of making such adjustments is the topic of the second set of articles in this special issue.

The Zhou et al. (2010) study drives directly through the heart of the controversy of making FE adjustments in postconviction capital murder cases. Those who are not in favor of making such adjustments often argue that the FE is an aggregated phenomenon and cannot be reliably applied to individuals. Those who are in favor of such adjustments often respond that they are not really adjusting an individual's score but the cut off for intellectual deficiency—which can only be defined in the aggregate. It is perhaps predictable then that each contributor to this special issue interprets Zhou et al.'s finding through his or her own pair of "scientific spectacles." The Zhou et al. suggestion of possibly larger effects in the lower ability range is interpreted as supporting the view that the FE is only valid as an aggregated phenomenon because of its wide individual variation as argued in the articles by Hagan, Drogin, and Guilmette (2010), Ceci and Kanaya (2010), and Sternberg (2010). This same finding is interpreted by other authors of this special issue as indicating that the aggregated, mean adjustment is conservative and can therefore be applied with confidence to lower functioning individuals by Fletcher et al. and Flynn. Reynolds and colleagues appear to agree with the latter camp stating that while future research may confirm that a larger correction needs to be applied for lower IQs, best practice for now is the application of a constant effect.

In discussions about FE adjustments, the key issue centers on which generation constitutes an appropriate normative reference group for the individual being tested. A person who was born in 1978 and tested in 2010 at age 32 using a current IQ test will be compared with a normative reference group of 30- to 34-year-olds born between 1976 and 1980. In this case, the person is being compared with the generation to which he or she belongs. If the test used was 20 years old at the time this person was tested, then he or she would be compared with a group of 30- to 34-year-olds who were born between 1956 and 1960—clearly not the same generation. If generational effects exist—as all contributors to this special issue agree they do—then this is clearly not the optimal normative reference group for this individual. Consequently, an adjustment to the person's score that takes into account changes in the normative reference group may be appropriate. This example makes clear that the FE is related to changes in the score distribution of the reference sample.

It is helpful to keep in mind that adjusting an individual's score on an obsolete test is really a back-door method to adjusting the cut-off score for intellectual deficiency based on a more current reference sample. The reason it is done this way is that the courts will not accept a change to the criteria for mental retardation, but some will consider arguments about the precision of the defendant's score. Adjusting the cut-off is clearly a normative issue and norms are an aggregated phenomenon by definition. Viewing the issue through this pair of scientific spectacles obviates the debate about the probabilistic nature of group data when applied to individuals.

Notwithstanding the above, concerns with the precision of FE adjustments exist and must be acknowledged. Until new norms are collected for the obsolete test we have no hard data on the new normative cut-off, but we may base estimates on historical rates of gain as identified by Flynn. Such estimations assume that the historical rate of change will continue although this cannot be known at the time the adjustment is made. Once the test is revised and new norms collected then the actual rate of gain is known and may be used to make a more precise adjustment. Of course, this assumes that the new normative sample is collected and stratified according to best practices.

The problem with reverse engineering an adjustment to the cut-off by adjusting an individual's score is that the adjustment is made based solely on the mean rate of gain. As every test developer knows, the variance is equally important! When it comes to normative comparisons, the mean tells us very little without reference to the variance. At the risk of stating the obvious, if a person

scores 12 points on a test whose mean is 20 points, we know only that they are below average—we have no information about how far below average this person scored until we know the variance. If the standard deviation (which is derived from the variance) is 8, then this person is 1 standard deviation below the mean with an IQ score of 85 and clearly not intellectually deficient. However, if the standard deviation is 4 points, then this individual is 2 standard deviations below the mean and on the cusp of mental retardation with an IQ score of 70. To complicate the matter, if the rate of gain is not normally distributed in the population—as implied by most of the causal theories—then the distribution of gain scores would be skewed creating different gains by IQ level as implied by Zhou et al.

Although it is likely that the Zhou et al. study will be used by others as evidence of the risks attending adjustments to individual scores, it may be better to view this article as one of the early studies in a line of research designed to improve the precision of FE adjustments based on a better understanding of the variance surrounding the mean rate of IQ gain. As stated by Zhou et al.,

> . . . forcing an IQ adjustment using a fixed rate could cause misleading results and potentially misclassify a proportion of examinees. At the same time, clinical practice must continue even as research continues that may impact practice. Although the evidence for differential adjustments based on ability level is still nascent, early indications appear to favor slightly larger adjustments in the lower ranges of scores where high-stakes legal evaluations are most likely to occur.

Research into Flynn's black box needs to continue to better understand the variance issue and also inform potential causes of the FE. As suggested above, this research should not be based on single subtests but on total scores or factor-based composites. As this research emerges, we may be able to improve the precision of FE adjustments. As practice must continue, however, Flynn (2010) and Reynolds et al. (2010) may be correct that 3 points per decade is conservative in the low ability range.

Ethical Considerations

Sternberg's (2010) article is unique among the articles in this issue in that it does not directly address either feature article, but questions a deep and perhaps ego-syntonic assumption among FE researchers that increasing IQ scores are good for society. In a selective review of recent history, Sternberg notes the many highly intelligent people who have done great harm to society, and proposes that increasing IQ without wisdom is perhaps unwise. It should be obvious to most readers that in his passion to make a point Sternberg minimizes the many intelligent people who have done great good for society, and ignores a line of research relating IQ with positive life outcomes for individuals (Gottfredson, 1997; Gottfredson & Saklofske, 2009). His article is important, however, because he raises the possibility of studying ethical intelligence—which is the ability to apply ones intelligence for good, or at least to avoid doing harm to others. A discussion of ethical intelligence is particularly apropos to this special issue given the dire consequences of adjusting or not adjusting IQ scores for FE in capital murder cases.

It has sometimes been said that science cannot solve moral problems in society. But that has never stopped us from trying. Apart from abortion, the death penalty may be one of the more disputed moral problems in society today and this creates an ethical dilemma for psychologists testifying in such cases and for researchers studying the issue of FE adjustments. As a moral philosopher by training and profession, perhaps Dr. Flynn can appreciate this angst. Amnesty International, which has long been opposed to the death penalty, issues a widely accepted annual report on state-sponsored executions. The current report (Amnesty International, 2009) states

that executions in the United States were at a 3-year high ($n = 52$) in 2009, with almost half ($n = 24$) occurring in Texas. Only four countries executed more people than the United Sates in 2009, and those were China, Iran, Iraq, and Saudi Arabia. The United States is the only country in the Americas that executed anybody in 2009. The only European nation with the death penalty still on its books is Belarus. In 2009, the Constitutional Court in Russia renewed a moratorium on death sentences, and Kenya commuted the death sentences of 4,000 condemned prisoners. The report concludes that the "path toward full abolition of the death penalty is irreversible." If this occurs, the issue of adjusting defendant's scores for FE—at hotly debated as it is—may one day be nothing more than a minor historical footnote in the history of psychology.

Even so, the death penalty is a legal reality at present and I must admit that several times during the course of writing this article I have stopped to consider how my words might be used in a death penalty case, and I cannot help but wonder if my colleagues have done the same. I suspect that all of us who have contributed to this special issue have valiantly attempted to be dispassionate scientists studying data but that all of us have our own private views on the death penalty. Furthermore, my clinical intuition tells me that the moral spectacles though which each of us views the death penalty issue has likely swayed our interpretation of ambiguous data in ways which we may not even be aware. For instance, both opponents and supporters of FE adjustments have found solace in the Zhou et al. (2010) results. Furthermore, each of the death penalty articles stakes out one side of the issue and none of the authors acknowledge the legitimacy of any competing point of view. Perhaps this is out of fear of attorneys taking their words out of context and making the witness appear indecisive in court. Still, the ethical position of an expert witness providing testimony is not to argue either for or against FE adjustments but to inform the court about the extant research on the topic. This is an important point that all respondents have missed.

Knowns and Unknowns

What conclusions can the field agree on about the extant research into FE?

- Dr. Flynn is to be commended for a meaningful life's work.
- The Flynn effect (FE) is real.
- The FE has been shown to be near 3 IQ points per decade on average across a large number of studies, countries, and tests.
- A smaller number of studies have reported smaller rates of gain in particular countries or tests.
- It is unknown if the 3-point per decade rate of gain will continue in the future.
- Changes in tests on revision complicate the measurement of rates of change.
- Measurement invariance should inform comparisons of scores across tests or test versions.
- The rate of gain may be different for different cognitive abilities. Verbal and crystallized tasks often show smaller gains than visual spatial and fluid reasoning tasks. But there is some disagreement about which tasks measure fluid reasoning.
- Little is known about rates of gain for processing speed or working memory tasks.
- The causal mechanisms are unknown. Several viable theories have been advanced, most involving some form of improvement in education or health, yet proofs have been limited to logical rhetoric and historical correlation.
- Adjustment of scores on obsolete tests for high-stakes evaluations is controversial. The primary area of disagreement concerns the appropriateness of adjusting individual scores based on group data.

- There is no definition of when a test becomes obsolete. When asked privately, most FE researchers have 10 years in mind.
- Adjustments for routine clinical practice are not recommended.
- A small number of studies have suggested differential FE by ability level, but not enough is known about this at present.
- Advanced statistical methods are necessary to improve FE knowledge.
- FE research should continue.
- We will know more about the FE in 10 years than we know today.

Acknowledgment

I appreciate the careful review and comments about methodology by Mark Daniel, PhD, Senior Scientist at Pearson Clinical Assessment, on an earlier draft of this article.

Declaration of Conflicting Interests

The author is employed at Pearson Clinical Assessment which develops and publishes numerous psychological and educational tests including the Wechsler, Kaufman, and Ravens tests that are mentioned in this article.

Funding

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

References

Amnesty International. (2009). *Amnesty International Report 2009: State of the world's human rights*. London, England: Author.

Bowden, S. C., Lange, R. T., Weiss, L. G., & Saklofske, D. H. (2008). Invariance of the measurement model underlying the Wechsler Adult Intelligence Scale-IV in the United States and Canada. *Educational and Psychological Measurement, 68*, 1024-1040.

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.

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?* 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.

Flynn J. R., & Weiss, L. G. (2007). American IQ gains from 1932 to 2002: The WISC subtests and educational progress. *International Journal of Testing, 7*, 1-16.

Floyd, R. G., Clark, M. H., & Shadish, W. R. (2008). The exchangeability of IQs: Implications for professional psychology. *Professional Psychology: Research and Practice, 39*, 414-423.

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

Georgas, J., Weiss, L. G., van de Vijver, F. J. R., Saklofske, D. H. (2003). *Culture and children's intelligence*. San Diego, CA: Academic Press.

Glascher, J., Tranel, D., Paul, L.K., Rudrauf, D., Rorden, C., Hornaday, A., Grabowski, T., Damasio, H., Adolphs, R. (2009). Lesion mapping of cognitive abilities linked to intelligence. *Neuron, 61*(5), 681-691.

Gottfredson, L. S. (1997). Why g matters: The complexity of everyday life. *Intelligence, 24*, 79-132.

Gottfredson, L. S., & Saklofske, D. H. (2009). Intelligence: Foundations and issues in assessment. *Canadian Psychology, 50*, 183-195.

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.

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.

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.

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.

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.

Rogers, J. L. (2010). The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *American Psychologist, 65*, 1-12.

Steen, R. G. (2009). *Human intelligence and medical illness: Assessing the Flynn effect*. New York, NY: Springer.

Sternberg, R. J. (2010). The Flynn effect: So what? *Journal of Psychoeducational Assessment, 28*, 434-440.

Weiss, L. G. (2007). *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

Weiss, L. G., Chen, H., Harris, J. G., Holdnack, J. A., & Saklofske, D. H. (2010). WAIS-IV use in societal context. In L. G. Weiss, D. H. Saklofske, D. L. Coalson, & S. E. Raiford (Eds.), *WAIS-IV clinical use and interpretation: Scientist-practitioner perspectives*. 97-140, San Diego, CA: Elsevier.

Weiss, L. G., Saklofske, D. H., Coalson, D. L., & Raiford, S. E. (2010). Theoretical, empirical, and clinical foundations of the WAIS-IV. In L. G. Weiss, D. H. Saklofske, D. L. Coalson, & S. E. Raiford (Eds.), *WAIS-IV clinical use and interpretation: Scientist-practitioner perspectives*. 61-96, San Diego, CA: Elsevier.

Wicherts, J. M., Dolan, C. V., Hessen, D. J., Oosterveld, P, van Baal, G. C. M., Boomsma, D. I., & Span, M. M. (2004). Are intelligence tests measurement invariant over time? Investigating the nature of the Flynn effect. *Intelligence, 32*, 509-537.

Zhou, X., Gregoire, J., & Zhu, J. (2010). The Flynn effect and the Wechsler Scales. In L. G. Weiss, D. H. Saklofske, D. L. Coalson, & S. E. Raiford (Eds.), *WAIS-IV clinical use and interpretation: Scientist-practitioner perspectives*. 141-166, San Diego, CA: Elsevier.

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.