Introduction to Papers Inspired by Roger Moore Letter

Stephen Greenspan, PhD

The Fall 2006 issue of *Psychology in Mental Retardation and Developmental Disabilities* contained a lengthy letter by Roger B. Moore (2006) titled "Modification of Individual's IQ Scores is Not Accepted Professional Practice." This was a criticism of an earlier paper in the Spring issue Greenspan, 2006) in which I made the opposite argument, with particular application to death penalty cases. The editor, Sam Thios, kindly offered the opportunity to organize a response to Dr. Moore's thoughtful arguments. I have chosen to do this

through two brief papers. The first paper is by the foremost expert on the Flynn Effect, James Flynn himself, and addresses the major points made in Moore's letter. The second paper takes a "meta" approach, and explores the question "why take IQ scores so seriously in the first place?" In a later paper, I shall argue for a shift, in line with other DSM diagnostic categories, towards sole reliance on clinical judgment, with the basis for such judgments being evidence of a history of profoundly "foolish action" in social and non-social challenging contexts.

References

Greenspan, S. (2006). Issues in the use of the "Flynn effect" to adjust IQ scores when diagnosing MR. *Psychology in Mental Retardation and Developmental Disabilities*, 31, (3), 3-7.

Moore, R.B., Jr. (2006). Modification of individual's IQ scores is not accepted professional practice. *Psychology in Mental Retardation and Developmental Disabilities*, 32 (3), 11-12.

Capital Offenders and the Death Sentence:
A Scandal That Must Be Addressed

James R. Flynn[1]

A thought experiment

Stephen Greenspan (2006) has endorsed my proposal that the IQs of those on death row should be altered when it is clear that they have been scored against obsolete norms (Flynn, 2006). Roger B. Moore, Jr. (2006) has raised the doubts that this proposal usually inspires. Here is a thought experiment that may clarify the issues. Imagine we had a WISC-IV IQ of 70 in a capital case and then the Psychological Corporation issued a bulletin: "We made a mistake in copying the census data for parental years of schooling and our normative sample was one SD substandard on that variable. We have extracted from it a sub-sample that is correctly standardized. It gives IQs on average 5 points lower than the defective norms. In passing, we standardized a total of 13 WISC subtests.

While we originally chose numbers 1-10 of these to calculate full scale IQ, we are now convinced that a better package is numbers 1-7 and 11-13."

I will transpose Moore's objections to that context:

1. Denmark has no such problem.
2. We do not know whether the Psychological Corporation will make such mistakes in the future.
3. When the performances of the original and the altered samples are compared, score differences are not uniform over subtests.
4. That being so, it would be even worse if the Psychological Corporation had both altered their standardization sample and given the new group the Stanford-Binet.
5. When such mistakes have been made in the past, that is, a standardization sample one SD too low and a better

sub-sample derived from it, the difference this made to IQs was not uniform: these cases average at 5 points but vary from 2 to 7 points.
6. The bulletin provides no reason to raise the IQs of American blacks.
7. The Courts often ignore such bulletins. Therefore, this WISC-IV IQ score of 70 should be allowed to stand and the offender should be executed. (I have added the conclusion, of course, but it is what rejecting the bulletin means in practice).

The significance of IQ gains over time

Overwhelming data indicate that the old WISC (normed in 1947-1948) gives IQs about 17 points higher than current Wechsler-Binet tests (Flynn, 2007, Table 1).

1 James R. Flynn is Emeritus Professor of Political Studies at New Zealand's Otago University. He can be contacted at jim.flynn@stonebow.otago.ac.nz

Assume someone gave that test to a boy of 15 in 2002 and he scored an IQ of 87. At 20, he is on death row. Does anyone really believe that this score should be entered into the balance sheet that determines whether or not he will be executed? Recall what IQ scores are all about. They tell you where you stood in relation to your peers at the time you were tested. That means a 15-year old in 2002 should be scored against the 15-year olds of 2002 and not against the lower scoring 15-year olds of 54 years ago. For this boy, the latter are the wrong standardization sample and their substandard norms (compared to his true peers) will inflate his IQ by about 17 points.

Another thought experiment. The Psychological Corporation issues a second bulletin: "Our mistake had nothing to do with census data. Rather than entering the normative data for the WISC-IV standardization sample, some fool hit the wrong key and entered the obsolete WISC-III norms. The good news is that this only inflates IQs by 3 or 4 points." My point is simply this: correcting IQ scores for norms obsolete at the time the subject took the test is simply a *special case* of any other correction we might make when we discovered we had used the wrong norms. Will such corrections be rough? They will indeed, but one thing is certain: worse than a rough correction is no correction at all.

Back to Moore's objections

1. Research in Denmark shows that IQ gains there stopped about 1995. Answer: Therefore, they should not lower scores for any test normed 1995 or after. Note, however, that they should adjust scores for tests normed before that date.
2. IQ gains in America may have waned. Answer: Nothing is certain, but Wechsler-Binet data suggest that they have persisted up to 2002 (when the last test was normed). Since the rate of gain showed no tendency to drop, we should wait for new data before assuming it has done so.
3. When subjects take both current and obsolete tests, the score differences are not uniform over subtests. Answer: The conclusion that Wicherts, et al. (2004) draws from this is that IQ gains cannot be equated with intelligence gains. In my recent book *What is intelligence?*, I concur. My view is that IQ gains signal the investment of our intelligence in new areas, but do not mean an enhanced ability to deal with everyday life. Surely, this is a reason for adjusting scores. If IQ gains meant such enhanced competence, the pool of those who are MR would have been steadily emptied.
4. We should not compare Wechsler tests with Stanford-Binet tests. Answer: If both of these tests are not valid measures of MR, we should stop using one. If we use both and their norms consistently generate different IQ scores, we must try to resolve the conflict. Otherwise, which test you take (the one that gives higher scores or the one that gives the lower scores) turns the death penalty into a lottery.
5. When we compare pairs of earlier and later tests, IQ gains may average at .300 point per years but they vary around that mean. Answer: When results for the WAIS-III are discarded, for reasons to be given, they all fall within .160 to .489. To refuse to adjust scores is to say, in effect, since we cannot be sure where IQ gains lie between .200 and .400, we will pretend they do not exist at all. Well, adopting a rate of 0.3 means you are almost certain to be no more than .100 per year off. Not to adjust scores is to absolutely guarantee that you will be from .200 to .400 off.

6. We do no raise the IQs of blacks simply because they have a negative group difference with whites of 15 points. Answer: No, because we have no reason to think that defective norms are the culprit. This hardly justifies making no adjustment where norms are defective. If you believe that Wechsler-Binet tests unfairly penalize blacks, label them MR when they are not, then you surely ought to raise their scores.
7. The Courts have an equivocal attitude to adjusting IQ scores. Answer: Then we will have to educate them. The tide may be turning. On 28 February 2007, the US Navy-Marine Corps Court of Criminal Appeals stated: "In determining whether an offender meets this definition (author: of MR), standardized IQ scores scaled by the SEM and the Flynn effect will be considered". Web: NMCCA (CODE 07).

I hope that readers will peruse Flynn (2006) and Flynn (2007) for a full statement of my case. Not adjusting scores puts offenders at the mercy of chance: whether someone was lucky enough to take a relatively current IQ test or unlucky enough to take a test with obsolete norms.

Table 1. Twelve Estimates of Recent IQ Gains Over time[2]

Tests Compared	Gains	Period	Rate
(1) WAIS-III (1995) & SB-5 (2001)		6	
(2) WAIS-R (1978) & SB-4 (1985)		7	
(3) WAIS-III (1995) & WISC-IV (2001.75)		6.75	
(4) WISC-III (1989) & SB-5 (2001)		12	
(5) WISC-III (1989) & WISC-III/IV (2001.75)		12.75	
(6) WISC-R (1972) & WISC-III (1989)		17	
(7) WISC-R (1972) & SB-4 (1985)		13	
(8) SB-4 (1985) & SB-5 (2001)		16	
(9) WAIS-R (1978) & WAIS-III (1995)		17	
(10) SB-LM (1972) & SB-4 (1985)		13	
(11) WISC-R (1972) & WAIS-R (1978)		6	
(12) WISC-III (1989) & WAIS-III (1995)		6	

Adapted from Table 3 of Flynn (2007). Used with permission of Syndics of Cambridge University Press.

Enter the WAIS-III

Moore does not attack my other suggestion for adjusting IQ scores. IQs should be lowered by .3 points for every year between when the test was normed and when the subject sat the test. In addition, WAIS-III IQs should be lowered by another 2.34 points. The appendix of Flynn (2007) defends the second suggestion at length, but one cannot anticipate all objections. Therefore, I will confront those put by Larry Weiss (2007) of the Psychological Corporation.

First: The rate of IQ gain may have slowed down since 1978

This appears to provide a beautifully simply escape route for the WAIS-III. Comparing the norms of the WAIS-R (1978) and the WAIS-III (1995) gives a rate of gain of .171 IQ points per year. As Weiss notes, the comparison between the SB-4 (1985) and SB-5 (2001) gives .173 or almost exactly the same rate. He cites a comparison between an earlier and later version of the Differential Abilities Scale (DAS) that adds confirmation. So why assume there is anything wrong with the norms of the WAIS-III?

The norms of the WAIS-III are indicted by a whole matrix of comparisons interrelated in such a way that weakening the problems arising from one comparison only intensifies the problems arising from others. Let us state Weiss's full hypothesis: (1) There is a prima facie case that the norms of the SB-4 and SB-5 are valid -- otherwise why cite them as validating the norms of both the WAIS-R and WAIS-III; (2) The post-1978 rate of gain fell off to .171 IQ points per year.

Table 1 presents 12 comparisons of a variety of Wechsler and Stanford-Binet norms. Look at comparison (1): the norms of the WAIS-III produce IQ scores 5.5 points higher than the also (assumed to be) valid SB-5. Yet the two tests were normed only 6 years apart and therefore, according to Weiss, the difference should be only about one point (6 x his rate of .171 = 1.03 points). So the SB-5 now indicts the WAIS-III norms as too lenient by almost 4.5 points. Look at comparison (3): the norms of the WAIS-III produce IQ scores 3.10 points higher than the WISC-IV. The two tests were normed 6.75 years apart and Weiss's hypothesis leads us to expect a difference of 1.15 points (6.75 x .171 = 1.15). That would indict the WAIS-III norms as too lenient by almost 2 points. This range of 2 to 4.5 points indicates that my value for the inflation of IQs by the WAIS-III is a bit on the low side.

Worse, the hypothesis of gains dropping to .171 points per year after 1978 is overwhelmingly falsified by the total data. There are 8 comparisons that refer to the post-1978 period, namely, all but (6), (7), (10), and (11), and they give an average rate of gain of .355. Weiss picked out the values that rank second and third lowest. If I picked out those that rank second and third highest, we would posit a rate of .474 IQ points per year.

Weiss's hypothesis focuses our attention on the 5.5-point discrepancy between the two tests actually being used on death row. It is a discrepancy reinforced by the comments of the lawyers in every capital case I have encountered: Use the SB-5 for a low IQ that will get the offender off; use the WAIS-III for a high IQ that will get the offender executed. The WAIS-III was normed 6 years earlier than the SB-5 and therefore, the legitimate expectation would be that WAIS-III IQs would give scores 1.8 points higher (6 years x .300 = 1.8). This leaves someone's norms out by about 3.7 points (5.5 - 1.8 = 3.7). Weiss notes that different test content can generate different rates of IQ gain. But this would be a case of something far more serious: different test content generating different IQ scores at the crucial level of MR -- even were the tests normed on the same day and obsolescence was not a factor.

If test content is the culprit, we will be in deep water indeed. At present, there is no need to posit such a problem in that the hypothesis of substandard WAIS norms is compatible with all existent data. The most urgent thing to do in this area is to test the 5.5-point discrepancy between the WAIS-III and SB-5 by replication. It creates a scandal that cannot be ignored. As long as it stands, can anyone deny that we have a problem as to which set of norms to trust? The Weiss hypothesis suggests trusting both: that is not an option.

It may be said, what of the SB-5's norms? As we shall see, the jury is out until we have data from tests normed after the SB-5. But we can do a preliminary evaluation because Table 1 offers comparisons in which the SB-5 figures as the more recent test. Setting comparison (1) aside because it includes the WAIS-III, comparison (4) with the WISC-III gives a rate of gain of .417 IQ points per year; and comparison (8) with the SB-4 gives .173 points. These give an average rate of .295 or almost exactly what we would expect. For now, the norms of the SB-5 get a pass, although if the test were taken today, scores would have to be lowered by 1.8 points for six years of obsolescence.

In passing, Weiss mentions better representation of low-scoring subjects as a possible reason for discrepancies in the measured rate of IQ gains. However, the architects of both the WAIS-III (Tulsky & Zhu, 1997) and SB-5 (Roid, 2003) went to new lengths to insure that such groups were not omitted. So the comparison between these two tests should be unaf-

fected. As for other comparisons, my own calculations suggest that such improvements affected mean IQ by about one-half of a point (Flynn, 1998, Table 2).

Second: Scholars have made positive comments about the WAIS-III norms

The scholars Weiss cites evaluated the WAIS-III prior to the year 2000. The most damning evidence against the WAIS-III became available only after that date, namely, comparisons of its norms with those of the SB-5 and the WISC-IV. You simply cannot assess a set of norms in isolation. They must be analyzed in a matrix, each member of which is a comparison of two sets of norms separated in time. In addition, the norms being assessed must appear as the earlier set in some comparisons and the later set in others.

Look at the pattern the WAIS-III exhibits in Table 1. In comparisons (1) and (3), where it is the earlier test, it produces huge rates of gain, namely, rates of 0.917 and 0.459 points per year for an average of 0.688. In comparisons (9) and (12), where it is the later test, it produces virtually nil gains, that is, rates of +0.171 and - 0.117 for an average of 0.027.

What would explain this discrepancy? As performance rises, representative samples of Americans set tougher and tougher norms. Therefore, if we give the same group of subjects two tests, one normed earlier and the other normed later, they will get higher scores on the earlier test with the easier norms. If a particular test has too easy norms even when standardized, it will show a peculiar pattern in comparisons. Posit that the norms of the WAIS-III were too lenient and produced too high scores even in its own day. Then the mean IQ of a group that takes it as the earlier test will soar far above the mean IQ of the same group on a later test and produce an inflated estimate of IQ gains. And the mean IQ of a group that takes it a later test will show almost no difference compared to their mean IQ on an earlier test and produce a deflated estimate of IQ gains. Such a pattern cries out for a diagnosis of too lenient norms and inflated scores.

In passing, note what adjustment would bring the WAIS-III results into line - if taken in isolation from the whole matrix. If we assume that its norms inflate scores by 2.45 points, the values in Table I alter as follows:

(1) 5.50 - 2.45 = 3.05 points gained over 6 years = rate of .5083

(3) 3.10 - 2.45 = 0.65 points gained over 6.75 years = rate of 0.0963
Average where WAIS-III is the earlier test becomes rate of .3023.

(9) 2.90 + 2.45 = 5.35 points gained over 17 years = rate of 0.3147

(12) - 0.70 +2.45 = 1.75 points gains over 6 years = rate of 0.2917 Average of where WAIS-III is the later test becomes .3032.

Is the WAIS-III unique?

In other words, assuming that the WAIS-III norms inflate scores by a slightly higher value than I have suggested brings its earlier/later test results into alignment. Why have I foregone 2.45 in favor of 2.34 points? Out of respect for the total matrix. The lesser adjustment brings the minimum and maximum WAIS-III values closest to the minimum and maximum values found in the 8 comparisons that do not involve the WAIS-III. Since we are assuming that these have roughly representative norms, it seemed the fairest thing to do.

A check: does any other set of norms show the same problems as those of the WAIS-III? We can only test those that figure as both earlier and later in comparisons. I will add a comparison not included in Table 1 because it is less recent than those therein. Comparison of the WAIS (1953-54) and the WAIS-R (1978) shows a rate of gain of .328 IQ points per year (Wechsler, 1981, p. 47).

We can now examine the norms of three additional tests:

1. The WAIS-R. It figures as the later test in two comparisons, namely, with the WAIS (.328) and the WISC-R (.160) for an average rate of .244. It figures as the earlier test in two comparisons, namely, with the SB-4 (.489) and the WAIS-III (.171) for an average of .330. Discrepancy: .330 - .244 = .086.
2. The SB-4. It figures as the later test in three comparisons, namely, with the WAIS-R (.489), the WISC-R (.227), and the SB-LM (.116) for an average of .277. It figures as the earlier test in one comparison, namely, with the SB-5, which gives a rate of .173. Discrepancy: .277 - .173 = .104.
3. The WISC-III. It figures as the later test in one comparison, namely, with the WISC-R, which gives a rate of .312. It figures as the earlier test in three comparisons, namely, the SB-5 (.417), the WISC-IV (.332), and the WAIS-III (-.117) for an average of .211. Discrepancy: .312 - .211 = .101.

I have included the WAIS-III in these comparisons because, at present, we are testing, rather than discounting, its validity.

The norms of other tests create discrepancies ranging from .086 to .104. What kind of discrepancy do the norms of the WAIS-III create? As we have seen, as the later test it gives .027 and as the earlier test it gives .688. The discrepancy is .661 and over six times as great as any other set of norms. I invite anyone to explain this

Figure 1.

Note the inflated rate of gain (.688) when the WAIS-III is the earlier test and the virtually nil rate (.027) when the WAIS-III is the later test in a combination.
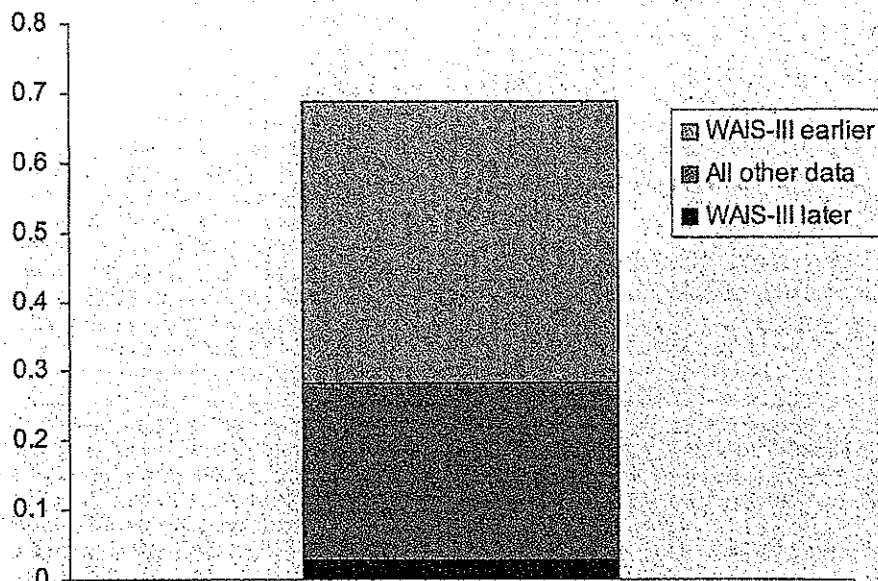
Adapted from Figure A2.1 of Flynn (2007). Used with permission of Syndics of Cambridge University Press

away by assumptions about the rate of gain altering from one decade to another. They will spend much time and find that nothing works.

This issue cannot be settled by appeals to authority. However, even though my suggestion that the WAIS-III norms inflated IQs by 2.34 points is barely in print, one distinguished scholar has been convinced. Merrill Hiscock, Professor of Psychology at the University of Houston, and one of his students analyzed scores drawn from the Wechsler manuals for the WAIS, WAIS-R, and WAIS-III to estimate the effects of aging. They found the bizarre result that the annual rate of change from the WAIS to the WAIS-R was 87% higher than that from the WAIS-R to the WAIS-III -- until they made my recommended adjustment of 2.34 points to reduce inflated WAIS-III scores. That done, they were able to offer a coherent interpretation of both the effects of aging and cohort gains implicit in the data (Dickinson & Hiscock, under review).

Third: The rigorous selection of the WAIS-III standardization sample

The Wechsler and Stanford-Binet organizations do everything humanly possible to get representative samples. What a poor reward when a scholar like myself suggests that they have simply had bad luck! When that happens, aware of their impeccable procedures, it is inevitable that the test

publishers will reject the suggestion.

However, they cannot do the impossible. To get norms for the WAIS-III and the older age groups of the SB-5, you have to select a representative sample of American adults. With schoolchildren, as long as you locate a good sample of schools, you are home free. Adults are not waiting for you gathered together in one institution. They can be located only if you visit a representative sample of their work places or homes, and no one can do that. School children have to participate if their school cooperates. Adults can self-select either by being too busy (high IQ subjects) or too uninterested (low IQ subjects). These distortions cannot always be detected by using stratification variables such as SES or years of schooling. It is likely that adult samples can only be accurate to within a few points. See Flynn (2006b) for a detailed discussion. There is a presumption in favor of norms for school-age children but no such presumption for adults.

Fourth: IQ gains measured by other kinds of tests should be taken into account.

Weiss notes an estimate yielded by two recent versions of the Differential Abilities Scale. However, as he points out, different tests produce different rates of gain, for example, data from many nations show that gains on Raven's Progressive Matrices are far larger than those on Wechsler-Binet

tests (Flynn, 1987). The DAS is a different kind of test than the Wechsler-Binet tests, as are the Primary Mental Abilities Test, the Armed Forces Qualification Test, and so forth. Only confusion will result if we lump their data into our matrix of Wechsler and Stanford-Binet data.

However, I will violate this rule simply to show how little difference it makes to slot an extra data set favorable to the WAIS-III into the total matrix. We will treat the DAS data as constituting an additional comparison in which the WAIS-III was the later test of a pair that yields a rate gain of .170 IQ points per year. We would now have three such comparisons, namely, .170, .171, and -.117 for an average of .075. The value for comparisons in which the WAIS-II was the earlier test would be unaltered at .688. Therefore, the discrepancy would now be .613 (.688 - .075). This is still six times the next largest discrepancy of .104. You cannot simply replicate a particular value here or there and evade the case that the total mass of data constructs brick by brick.

Summing up

There is nothing wrong with using the WAIS-III to test offenders on death row or anyone else. It is a good test and only needs its norms adjusted. In addition to the 2.34 points, its norms are 12 years obsolescent and this dictates subtracting another 3.6 points, making a total of 5.94 points.

This deduction assumes the test is being administered in 2007. Those who are unconvinced must, in all decency, suggest an alternative solution to the problem of the WAIS-III and the SB-5 turning death into a lottery. Figure 1 provides a graphic illustration of why you should look very closely at the WAIS-III. Opting for the SB-5 would not save you the trouble of adjusting scores. Its norms appear to have been accurate at the time it was standardized, but they are now 6 years obsolete and current scores should be lowered by 1.8 points.

References

Dickinson, M. D., & Hiscock, M. (under review). Contribution of the Flynn Effect to age differences in normative scores for WAIS-III and WAIS-R subtests.

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

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

Flynn, J. R. (2006). Tethering the elephant: Capital cases, IQ, and the Flynn Effect. *Psychology, Public Policy, and Law, 12:* 170-178.

Flynn, J. R. (2007). *What is intelligence? Beyond the Flynn Effect.* Cambridge University Press. (available August 2007; description at www.cambridge.org).

Greenspan, S. (2006). Issues in the use of the "Flynn Effect" to adjust IQ scores when diagnosing MR. *Psychology in Mental Retardation and Developmental Disabilities, 31* (3), 3-7.

Moore, R.B., Jr. (2006). Modification of individual's IQ scores is not accepted professional practice. *Psychology in Mental Retardation and Developmental Disabilities, 32* (2).

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

Roid, G. H. (2003). *Stanford-Binet Intelligence Scales, Fifth Edition -- Technical Manual.* Itasca IL: Riverside.

Tulsky, D., & Zhu, J. (1997). Lowering the floor of the WAIS-III. Paper presented at the 105th Annual Convention of the American Psychological Association, Chicago, Illinois.

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

Weiss, Lawrence G. (2007). WAIS-III technical report: response to Flynn. On web at Harcourt Assessment response to Flynn.

Flynn-Adjustment is a Matter of Basic Fairness: Response to Roger B. Moore, Jr.

Stephen Greenspan[1]

In his lengthy letter in the last issue, Roger B. Moore, Jr. (2006) criticized my earlier article (Greenspan, 2006), in which I discussed the use of the "Flynn Effect" to correct for obsolete IQ norms, especially in so-called "Atkins" (death penalty exemption) hearings. The authoritative response comes from the real expert on this matter, James R. Flynn (this issue), but I do want to say a few words in my own defense.

If I implied that use of Flynn adjustment is becoming commonplace, that was a mistake. My purpose was to make members of Division 33 aware of this new practice. It may become commonplace, but it hasn't happened yet, at least for non-Atkins purposes. While Flynn corrections are quite commonly used in Atkins-related testimony, and the practice has met widespread acceptance by trial judges (two points that Moore concedes), the matter is still being litigated on the appellate level, as Moore correctly points out.

Moore, along with many other practicing psychologists, clearly has a problem with the idea that IQ scores may need adjusting. However, the science of psychology has fundamentally accepted the reality of the Flynn Effect (Neisser, 1998), with leading intelligence researchers putting their main efforts into understanding rather than disputing the phenomenon (see Mingroni, 2004). Moore finds support in the fact that the APA did not raise the Flynn effect in its amicus brief in the Daryl Atkins case, but the purpose of that lawsuit was to address the constitutionality of executing defendants with MR, not to determine whether Mr. Atkins himself had MR. The failure to mention the use of Flynn adjustments (which ante-Atkins had not come into widespread use) hardly constitutes proof of APA's opposition.

Moore's most serious attack was on the evidence supporting continued use of the 0.3 FS IQ points per year of norm obsolescence adjustment formula, and here

I will leave it mainly to Professor Flynn to respond. I do need, however, to elaborate briefly on an aspect of my earlier paper where I did not go into sufficient depth, and that has to do with the research methods that have been used. In fact, there have been two very different kinds of studies. The first set of studies (e.g., Flynn, 1984) provides the basis of the 0.3 Flynn adjustment formula; these studies involve the same subjects being tested on an IQ test and then being retested 1-4 years later on a later-normed IQ test. These could involve two editions of the same test (e.g., the WISC-R and the WISC-III), but virtually all possible combinations of Wechsler (adult and child) and Stanford-Binet tests have been compared using this within-subject method, sometimes reported in the test manuals themselves and sometimes by other investigators. An example, involving the WISC-R and WISC-III, shows that Ss taking both showed a 5-point drop in IQ on the WISC-III,

1 Stephen Greenspan is Emeritus Professor of Educational Psychology at the University of Connecticut. He can be contacted at stephen.greenspan@gmail.com.

as compared to their scores on the WISC-R. The obvious explanation is that the norms for the later test were toughened by the test publishers to compensate for the fact that the standardization sample did better. As the WISC-III was normed 17 years after the WISC-R, dividing 5.0 by 17 produces an average FS-IQ increase in the standardization sample of 0.3 points per year (for P-IQ it was around 0.7 points, suggesting that most of the effect is found on Performance subscales). This approximate number has been replicated in comparison after comparison, regardless of the tests used.

The major exception to this finding was the shift from the WAIS-R to the WAIS-III, which is where the idea originated that the Flynn Effect should now be restated at 0.17 points. Flynn (2006) has pointed out that this anomalous finding was a result of major changes in the way the Psych Corp put together its standardization sample for the WAIS-III. Using marketing firms for the first time to assemble its sample, the publisher sought admirably to correct for the "tree stump" phenomenon (whereby a subject who gave not a single correct response would still receive a score in the 40's) by stratifying for low ability. Unfortunately, they appear to have done too good a job, in that the WAIS-III sample had too many low-IQ Ss. It appears that this non-comparable lower-performing sample produced norms that overstated IQ by 2.34 points. Plugging this correction into the same-subject studies brings the Flynn effect factor back to 0.3. Moore cites a letter from a Psych Corp official defending the integrity of the WAIS-III norms as evidence against the use of Flynn's correction formula. Can anyone imagine that Psych Corp would take any other public position, given the mammoth legal liability that would likely accrue from any admission that they may have blown it when compiling the WAIS-III IQ tables?

The other kind of study reported by Flynn (1987) is much less relevant to the U.S. legal arena, and this has to do with the reporting of trends in intelligence test performance in other Western countries. The basic paradigm here has involved performance on military screening batteries, where different cohorts of new recruits are tested in subsequent years on the same batteries. These cross-sectional comparisons have shown gradual increases in performance on these batteries, which may or may not reflect increases in general population intelligence. However, these cross-sectional comparisons are much more sensitive to sampling differences than is the case with US IQ tests (where publishers go to great lengths to ensure comparable samples). Thus, in the two Scandinavian studies cited by Moore (Sundet, Barlaug & Torjussen, 2004; Teasdale & Owen, 2006), upward trends in performance over the years are highly correlated with increases

in numbers of military recruits with some college education. One source of the idea that the Flynn effect may no longer apply is that in these Scandinavian countries, the increases in military scores have stopped, and scores may even be going down. However, a plausible explanation not addressed at all by the authors of those papers is that these societies, and their armies, have become less homogeneous, in large part due to a sizeable influx of less-educated immigrants. Furthermore, these put-together screening batteries are not in any real sense IQ tests, and attaining ceiling on some of the components (such as a Ravens-like measure) may be an artifact of the measures used. More to the point, however, is that IQ trends in Scandinavia have zero relevance to American norms, and the basic point of an Atkins hearing is to see if an individual qualifies for a diagnosis of MR when compared to the US population, and not to the population of Norway, Denmark or any other country.

Moore also alluded to the fact that defendants in Atkins cases, aside from having low IQ also tend to be poor and minority, and if one is to correct for outmoded general norms (which would lower IQs), one should also correct for SES and ethnicity (which would raise IQs). Aside from the fact that the main function of adaptive behavior is to protect against possible false positives, this argument (which is chilling in the sense that a correction proposed by Mercer to reduce discrimination in special education assignment would now be used to increase already high discrimination in execution assignment) misses the point that the purpose of an Atkins adjustment is not to correct for possible changes in actual intelligence in an individual or sub-group but rather to ensure that all individuals (regardless of their demographics) are measured by the same yardstick when being compared to an arbitrary (70 or 75) general population diagnostic criterion. The study by Kanaya, Scullin and Ceci (20023 showed that a test with outmoded norms (the WISC-R at the end of its life) produced inflated scores on the same low (70-85) IQ subjects as compared to the newly-normed WISC-III. In fact, for these low-IQ subjects, the mean drop from the WISC-R to WISC-III was the same five points reported by Flynn for the general population.

The need for making Atkins adjustment stems, of course, from the literalness with which courts, psychologists and lawyers, along with the general public, view IQ numbers and cut-offs. Aside from the notion that IQ doesn't capture all of intelligence, an idea that I first saw expressed by David Wechsler (1974) himself and that has been the focus of my own scholarly program, there is the problem that no serious rationale has ever been put forward to explain why a score ending in zero, or comprising a certain number of SD units, should be

the basis for deciding who has MR or, in this case, who should die. If a decision to execute someone hinges on whether he is a single point above such an arbitrary cutting line (and I recently participated in a hearing in which that was exactly the case), then basic fairness requires that the point be derived from norms that were valid at the time of testing.

James Flynn has made enormous contributions to the field of psychology. The irony is that he is a moral philosopher, not a psychologist, motivated not by a love of psychometrics but by a love of basic fairness. Moore has raised some interesting issues which Flynn has, I believe, answered effectively. But we should not allow statistical game-playing to blind us to what is fundamentally a moral problem.

References

Flynn, J.R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin, 95*, 29-51.

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

Flynn, J.R. (2006). Tethering the elephant – capital cases, IQ and the Flynn effect. *Psychology, Public Policy and Law, 12*, 170-189.

Greenspan, S. (2006). Issues in the use of the "Flynn Effect" to adjust IQ scores when-diagnosing MR. *Psychology in Mental Retardation and Developmental Disabilities, 31* (3), 3-7.

Kanaya, T., Scullin, M.H. & Ceci, S.J. (2003). The Flynn effect and U.S. policies: The impact of rising IQ scores on America society via mental retardation. *American Psychologist, 58*, 778-790.

Mingroni, M.A. (2004). The secular rise in IQ: Giving heterosis a closer look. *Intelligence 32*: 65-83.

Moore, R.B., Jr. (2006). Modification of individual's IQ scores is not accepted professional practice. *Psychology in Mental Retardation and Developmental Disabilities, 32* (2).

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

Sundet, J.M., Barlaug, D.G., & 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. and Owen, D.R. (2000). Forty-year secular trends in cognitive abilities. *Intelligence, 28*, 115-120.

Wechsler, D. (1974). *Selected papers of David Wechsler.* New York: Academic Press.