The WAIS-III and WAIS-IV: Daubert motions favor the certainly false over the approximately true

Article *in* Applied Neuropsychology · February 2009

DOI: 10.1080/09084280902864360 · Source: PubMed

1 author:

James Flynn
University of Otago
82 PUBLICATIONS **6,979** CITATIONS

SEE PROFILE

APPLIED NEUROPSYCHOLOGY, *16*: 98–104, 2009

ISSN: 0908-4282 print/1532-4826 online
DOI: 10.1080/09084280902864360

Psychology Press
Taylor & Francis Group

The WAIS-III and WAIS-IV: *Daubert* Motions Favor the Certainly False over the Approximately True

James R. Flynn
University of Otago, Dunedin, New Zealand

Daubert motions oppose adjusting IQ scores. They argue that the rate of IQ gains over time (the Flynn Effect) cannot be set at 0.3 points per year with scientific exactitude; therefore, the adjustment formula that rate implies is inadmissible in capital cases. This ignores the fact that there is universal agreement in the scientific community that there have been substantial gains and that, therefore, the worst possible option is to simply leave inflated IQ scores unadjusted. That would undermine equity entirely.

New data from the WAIS-IV are included in a meta-analysis of 14 combinations of Wechsler and Binet IQ tests. The overall average is a rate of 0.311 points per year; the average within Wechsler tests is 0.299 point per year. A new estimate of the extent to which the WAIS-III inflated IQs, even at the time it was normed, yields 1.65 points (rather than 2.34 points). However, two new studies comparing the WAIS-III to the Woodcock-Johnson III and the Kaufman Adolescent and Adult Intelligence Scale give huge estimates. It is recommended that WAIS-III scores be set aside and subjects tested on the WAIS-IV and the Stanford-Binet 5.

Key words: *Atkino cases, Daubert motions, Flynn Effect, Q and death penalty, Q gains WAIS to WAIS-IV*

The court has before it a capital case. Mr. Smith was murdered at his home and the coroner established the time of death at 10 a.m. The defendant admits he entered the home but claims that Smith was already dead. There is a damning piece of testimony against him. A newspaper boy saw him entering the home and heard the town clock strike ten. However, the defense presents three witnesses who passed the town clock on the morning in question and noticed that, as usual, the town clerk was some days late in resetting it to mark the start of daylight saving time. They were all amused as they checked their watches. One put the actual time at 10:55 a.m., another at 11:00 a.m., and the third at 11:05 a.m. In any event, the actual time supported the defendant's testimony.

The prosecution argues that the three defense witnesses have no standing as "expert witnesses" and cites the criteria set by the *Daubert* case:

1. The witnesses' technique was to check the clock against a wristwatch. The scientific community in general has not yet accepted this method of adjusting time as measured by town clocks.
2. The theory that lies behind the technique is that reset watches are more likely to capture the actual time (under daylight saving time) than nonreset town clocks. The prosecution is unaware of anyone who has rigorously tested that theory.
3. Articles defending the theory have appeared in peer-reviewed administrative journals (that assess the consequences of lazy town clerks), but not in peer-reviewed chronological journals devoted to the science of measuring time.
4. The method of correcting the town clock has a margin of error that is not precisely measurable. Here, the prosecution is absolutely correct. The three defense witnesses all offer different corrections that indicate the town clock is slow by anywhere from 55 to 65 minutes; for all we know, a

Address correspondence to James R. Flynn, University of Otago, Box 56, Dunedin, New Zealand. E-mail: jim.flynn@stonebow.otago.ac.nz

fourth witness would put the correction at 50 to 65 minutes.

6. Following on from (4), there are no established standards for applying the technique. Adding an hour on to the time of a town clock not reset for daylight savings ignores all sorts of complications: a drunken town clerk might set the clock back an hour, experts note that the task of coordinating any mechanical clock or watch with Greenwich time is complex, and so forth.

The prosecution also makes a point of law. At the time the Supreme Court accepted town clocks as the measure of time, they were all set "in conformity with professional practice" and vouched for by experts as the most accurate measure possible. This was intended to void experts wrangling about the "real time," and ever since, courts have been reluctant to question their reliability and invite a new battle of experts.

I take it that anyone would regard all of this as bizarre. It misses the point: failing to revise the town clock means putting the time at something that is certain to be misleading. And while taking the average of the three more reliable times leaves us uncertain as to the exact time, it must be done to avoid grievous error.

An additional absurdity: taking town clocks at their face value makes a lottery out of who lives or dies. Surprisingly, there was an identical case in the next town, where the clerk was assiduous and kept the town clock updated. So whether you live or die depends on whether your town has a lazy or conscientious town clerk.

IQ TESTS AND OBSOLETE NORMS

There was nothing defective about the town clock. But clocks are not like the tape measure that allows me, if it is not defective, to measure height whenever I wish. They require maintenance from time to time and after each intervention they must be accurately reset. The fact that we cannot take IQ scores at face value has nothing to do with the quality of IQ tests. No IQ test has ever classified anyone as gifted, or normal, or suffering from mental retardation. We use the test norms to do that, and today, every test publisher accepts that we must reset the test norms periodically to keep test scores from being deceptive.

You must understand that the test score sends a message that has nothing to do with the test in isolation. To use an IQ test at all, you must administer it to a standardization sample, that is, a representative sample of Americans of all ages, or at least all ages in the age range within which the test is to be used. The average performance is by definition a score of 100. A score of 70 is by definition one that separates off the bottom 2.23 percent of the population (in statistical terms it is two standard deviations below the mean). If the score of 70 did not do that, no one would trust it, because the psychological community believes that the bottom 2.3% captures, roughly, the group that suffers from mental retardation.

For particular individuals, they must be compared to their peers—that is, to people that are of the same age. They must be of the same age because no 6 year old can be expected to match the performance of a 12 year old, and no 70 year old can be expected to match the performance of a 35 year old (unless they are superior for their age). They must be of the same age *at the same time*, because even people who suffer from mental retardation can rise on the percentile scale if you compare them with people of the past. Common sense tells us that someone at 70 on today's norms could be average if compared to people of the same age from the Stone Age. Overwhelming evidence tells us that an American at 70 on today's norms would be average if compared to people of the same age in 1909. The average American has gained 30 IQ points over the century, and this appears to be true at all levels of the IQ scale (Flynn, 2009).

There is an interesting issue here. The psychological community might want to debate whether 20th-century IQ gains are real intelligence gains even among the bottom 2.23%. But if this were literally true, only those with IQs below 40 on current norms would strike clinical psychologists as mentally retarded, and this is far from the case. In addition, there is a test meant to measure whether people can cope with everyday life, the Vineland Adaptive Behavior Scales. During a period (1989–2002) in which American schoolchildren gained over four points on the Wechsler Intelligence Scale for Children (WISC), they made (at best) no gains on the Vineland (Flynn, 2009, pp. 126–127; Vineland, 2006).

But even if we someday decide that the number of mentally retarded is dwindling, we would still have to adjust IQ scores so that individuals were being compared to their own age group at the same time. Consider, for instance, identical twins convicted of a capital offense. In 1975 at school, one takes the then new WISC-R whose norms were relatively current. The standardization sample was tested in 1972, so there is only a three-year lag between himself at age 11 and the 11 year olds who normed the test. He gets an IQ of 67 and lives. In 1975, his twin happens to attend a different school. There he takes the old WISC whose norms have not been updated since 1947–1948 (when its standardization sample was tested). So now there is a 27.5-year lag between himself at age 11 and the 11 year olds who normed the test. Thanks to being compared to 11 year olds from the distant past, when average performance on the test was worse, he gets an IQ of 74.35 and dies. But the extra points have nothing to do with

his mental competence—it is entirely the work of the obsolete norms!

No one felt they could make life and death a lottery in terms of whether or not a town clerk remembered to reset a clock. Do we really want to make life and death the same kind of lottery? To make death depend on whether a school psychologist had been prompt in buying the latest version of the WISC, or whether, perhaps because of a limited budget, decided to use up copies of an older version is unacceptable. No prosecutor or prosecution expert has had the courage to address that question. In sum, whatever we eventually decide about our criterion for mental retardation, we cannot in the meantime "tolerate the infliction of a sentence of death under legal systems that permit this unique penalty to be so wantonly and so freakishly imposed" (Justice Stuart in *Furman v. Georgia*, 1972).

ADJUSTING OBSOLETE IQ SCORES

The town clock example was an answer to a recent prosecution motion (November 2008) to exclude adjusting IQ scores in a capital case. The case is Leon Anthony Winston (Petitioner) v. Loretta K. Kelly (Warden) in the United States District Court for the Western District of Virginia (Roanoke Division) C.A. NO. 7:07ev00364. The warden moved that a stay of execution be vacated and supports this motion with a memorandum. The memo argues that the report of a psychologist who wants to adjust the petitioner's IQ scores in the light of the "Flynn Effect" (massive IQ gains over time) should be precluded. It describes itself as a "*Daubert* motion," citing a case that set standards for the admissibility of evidence and argues that IQ adjustments do not qualify as reliable evidence.

I will not repeat my rebuttal of the main drift of this *Daubert* motion, but briefly, it is that no court would argue that a piece of evidence *known* to be radically deceptive be left to stand because evidence that casts light on it, and does much to obviate its false implications, is not as precise as we would like. We cannot adjust the time given by the town clock except within a range, but we have enough evidence to know that we must make a rough adjustment to avoid a great injustice.

However, there is some detail in this motion worth picking out for comment. I advocate adjusting WISC and Wechsler Adult Intelligence Scale (WAIS) scores as follows: for every year between the year when a person took a test and the year when the test was normed, deduct 0.3 IQ points from the IQ score. Recall the example of the identical twins. The one who took the WISC-R when its norms were only three years out of date (1975 as compared to 1972) would have 0.9 points deducted, lowering his score from 67 to 66.1. The other who took the WISC when its norms were 27.5 years out of date (1975 as compared to the 1947–48) would have 8.25 points deducted, lowering his score from 74.35 to also 66.1. Now that we adjusted their scores, the identity of their performance is clear and both will live.

The *Daubert* memo emphasizes that the formula for adjusting IQ scores assumes a precision, which the evidence for the rate of IQ gains over time lacks. That rate is measured by giving the same groups of subjects both an older and a newer test and seeing how much better they do on the older test thanks to its more obsolete norms. If the two tests were normed 10 years apart, we would expect that the older scores would be inflated by an extra 3 IQ points (a rate of 0.3 points per year $\times$ 10 years = 3 points). The memo cites a table reproduced in Flynn (2009), Table 3 that uses multiple comparisons to estimate the rate of recent IQ gains. I hereby update this table to correct one value in the light of new evidence and to add two new comparisons. But the memo's point stands. The 14 comparisons in Table 1 give an average estimate of 0.311 points per year, so close to 0.3 as to make no difference. But the range is from a huge estimate of 0.917 points per year to one maverick negative estimate.

However, the first thing to note about Table 1 is the atypicality of comparisons where there are seven or fewer years separating the norms of the pair of tests. We see that these comparisons yield all of the extreme values. The reason is that any comparison may be a point or two off and that such a variation over six years influences the rate by twice as much as a comparison over 12 years. But more impressive are the values in bold. It is well known that rate of gain can differ from one kind of test to another, so let us compare like with like.

The bold highlights comparisons where either a later form of the WISC has been used to check an earlier form of the WISC or a later form of the WAIS has been used to check an earlier form of the WAIS. These show rates of gain averaging at about 0.3 points per year with admirable consistency. The final column in the table shows how little the points gained over time would have to differ from the actual gain, if we were to bring all results perfectly in line at a rate of 0.3 points per year. The values for all 14 comparisons average at only 1.23 points. The values for the four within-test Wechsler comparisons average at 0.39 points.

The *Daubert* memo quotes me as saying that we will only be sure that the WAIS-III (normed in 1995) has become obsolete at 0.3 points per year when the WAIS-IV results are published. The WAIS-IV was renormed in 2006 and Table 1 shows a gain over the 11 years at 0.306 points per year. It is now 2009 and the demand will be for results that go beyond 2006. This demand is not relevant to any test normed before 2006. We already have the evidence needed to adjust their scores for obsolescence. As for the WAIS-IV itself, once

TABLE 1
Fourteen Estimates of Recent IQ Gains Over Time

Tests Compared	*Gains*	*Period Years*	*Rate*	*Ideal Gain*	*Ideal vs. Real*
1. WAIS-III (1995) & SB-5 (2001)	+5.50	6	+0.917	1.80	3.70
2. WAIS-R (1978) & SB-4 (1985)	+3.42	7	+0.489	2.10	1.32
3. WAIS-III (1995) & WISC-IV (2001.75)	+3.10	6.75	+0.459	2.03	1.07
4. WISC-III (1989) & SB-5 (2001)	+5.00	12	+0.417	3.60	1.40
5. WISC-III (1989) & WISC-IV (2001.75)	+4.23	12.75	**+0.332**	3.83	**0.40**
6. WISC-R (1972) & WISC-III (1989)	+5.30	17	**+0.312**	5.10	**0.20**
7. WISC-R (1972) & SB-4 (1985)	+2.95	13	+0.227	3.90	0.95
8. SB-4 (1885) & SB-5 (2001)	+2.77	16	+0.173	4.80	2.03
9. WAIS-R (1978) & WAIS-III (1995)	+4.20	17	**+0.247**	5.10	**0.90**
10. SB-LM (1972) & SB-4 (1985)	+2.16	13	+0.166	3.90	1.74
11. WISC-R (1972) & WAIS-R (1978)	+0.90	6	+0.150	1.80	0.90
12. WISC-III (1989) & WAIS-III (1995)	−0.70	6	−0.117	1.80	2.50
13. WAIS-III (1995) & WAIS-IV (2006)	+3.37	11	**+0.306**	3.30	**0.07**
14. WISC-IV (2001.75) & WAIS-IV (2006)	+1.20	4.25	+0.282	1.28	0.08
Average of all 14 comparisons			+0.311		1.23
Average of 4 WISC/WISC & WAIS/WAIS comparisons			**+0.299**		**0.39**

Note. This table is useful for analyzing whenever the norms of a given test seem eccentric. For example, if a test has substandard norms, it will inflate estimates when paired with a later test and deflate estimates when paired with an earlier test. Use the Ideal versus real column to assess the WAIS-III: (1) It is paired with a later test in (1), (3), and (13) and these show deviations of 3.70, 1.07, and 0.07 toward too many points gained; (2) It is paired with an earlier test in (9) and (12) and these show deviations of 0.90 and 2.50 toward too few points gained; (3) the sum of the deviations is 8.24 and divided by 5 = 1.65, as the number of points by which the WAIS-III inflated IQ scores even at the time it was standardized.

again, we cannot be sure. but every one of us, lacking time machines to go into the future, uses the recent past to make a rough prediction, unless there has been some clear sign of a change that would undermine continuity. If someone has a case that IQ gains in America should cease, let them bring it forward. Their data should be about America and not about, say, Scandinavia (where gains have stopped) or six nations where we know they are persisting, some at a rate greater than 0.3. We do not predict temperatures in America on the basis of data from the North Pole.

ANALYSIS OF GAINS FROM THE WAIS-III TO WAIS-IV

Most cases turn on the adjustment of an obsolete IQ score from some version of the WISC taken at school or on some version of the WAIS taken on death row. Since the WAIS-IV data is just at hand, I will first analyze WAIS gains over the last half century.

We begin with a group to whom the Wechsler organization gave both the older WAIS (1953–1954) and the newer WAIS-R (1978). The group was aged 35 to 44 and numbered 80 (Wechsler, 1981, p. 47). Some of the eleven subtests used to compute Full Scale IQ had been revised, but no subtest was dropped or added. Therefore, the comparison is straightforward: they got a mean IQ of 111.3 on the earlier test and 103.8 on the later, giving an IQ gain over 24.5 years of 7.5 points.

The group that took both the (by now) older WAIS-R (1978) and the newer WAIS-III (1995) ranged from ages 16 to 74 and numbered 192 (Wechsler, 1997b, pp. 78–79). The list of 11 subtests that were used to compute Full Scale IQ had changed, but they gave the comparison group all 11 of the old WAIS-R subtests. That was fortunate because it meant that the true obsolescence of the WAIS-R could be measured. As Flynn and Weiss (2007) observe, comparing one basket of subtests to another distorts results. Therefore, I calculated the standard score total the group got on the same 11 WAIS-R and WAIS-III subtests. Using these totals and the WAIS-R conversion tables (Wechsler, 1981, pp. 93–109), I calculated Full Scale IQs for the two tests over all the ages covered. Since these gave a 4.2-point difference on average (with little variation), I simply subtracted that from their WAIS-R mean to get a WAIS-III mean. A Full Scale IQ of 105.8 on the earlier test and 101.6 on the later gave a gain over 17 years of 4.2 points.

The group that took both the WAIS-III (1995) and the newer WAIS-IV (2006) ranged from ages 16 to 88 and numbered 240 (Wechsler, 2008, p. 75). The list of subtests used to compute Full Scale IQ had not only changed, but had dropped from 11 to 10. But, once again, they gave the comparison group all 11 of the old WAIS-III subtests, and once again that was

fortunate because it meant that the true obsolescence of the WAIS-III could be measured. I calculated the total standard score the group got on the same 11 WAIS-III and WAIS-IV subtests. Using these totals and the WAIS-III conversion table, I calculated Full Scale IQs for the two tests.

When you estimate the mean IQ of all the members of a group from one total standard score, you must simulate the range of scores of its members. The best way to do this is to use a spread of standard scores that surround that total. In this case, a simplistic one-point conversion would have inflated the estimate of IQ gains over time. The proper method is described at the bottom of Table 2 and gave a 3.37 point difference. I subtracted that from the WAIS-III mean to get a WAIS-IV mean. A Full Scale IQ of 102.90 on the earlier test and 99.53 on the later gave a gain over 11 years of 3.37 points. Table 2 traces IQ gains all the way from the original WAIS (1953–1954) through the WAIS-IV (2006) and shows a total gain of 15.07 points for a rate of 0.287.

THE FORMULA OF 0.3 POINTS PER YEAR

The WISC was also standardized four times over a similar period and I have analyzed that data elsewhere (Flynn, 2009, pp. 180–181). The similarity of the gains on these two Wechsler tests, which is to say the similarity of the gains of children and adults, is striking:

1. WISC 1947–1948 to 1972 (24.5 years) = 7.63 points or a rate of 0.311 per year; WAIS 1953–1954 to 1978 (24.5 years) = 7.50 points or a rate of 0.306 per year
2. WISC 1972 to 1989 (17 years) = 5.37 points or a rate of 0.316 per year; WAIS 1978 to 1995 (17 years) = 4.20 points or a rate of 0.247 per year
3. WISC 1989 to 2001.75 (12.75 years) = 3.83–4.63 points or a rate of 0.300–0.363 WAIS 1995 to 2006 (11 years) = 3.37 points or a rate of 0.306 per year

TABLE 2
Gains from the WAIS (1953–54) to WAIS-R (1978) to WAIS-III (1995) to WAIS-IV (2006)

Subtest	*W*	*W-R*	*Gain*	*W-R*	*W-III*	*Gain*	*W-III*	*W-IV*	*Gain*	*TG*
Vocabulary	11.9	10.1	1.8	10.8	10.2	0.6	11.0	10.0	1.0	3.4
Similarities	11.9	9.7	2.2	11.3	10.4	0.9	11.0	10.3	0.7	3.8
Arithmetic	11.3	10.3	1.0	10.1	10.4	−0.3	9.7	9.7	0.0	0.7
Digit span	10.4	9.8	0.6	10.4	10.3	0.1	10.4	10.1	0.3	1.0
Information	11.4	10.3	1.1	10.5	10.5	0.0	10.4	9.9	0.5	1.6
Comprehension	12.0	10.2	1.8	11.0	10.5	0.5	10.5	10.1	0.4	2.7
Picture completion	11.2	9.4	1.8	11.1	10.7	0.4	10.6	9.7	0.9	3.1
DS-Coding	11.6	9.8	1.8	11.8	10.6	1.2	10.0	9.8	0.2	3.2
Block design	10.9	9.9	1.0	11.4	10.7	0.7	10.5	10.2	0.3	2.0
Picture arrangement	11.1	10.3	0.8	11.1	10.5	0.6	10.6	9.7	0.9	2.3
Object assembly	11.5	10.2	1.3	11.3	10.4	0.9	—	—	—	—
Matrix reasoning	—	—	—	—	—	—	10.9	10.3	0.6	—
Sum SS	125.2	110.0	15.2	120.8	115.2	5.6	115.6	109.8	5.8	

Full Scale IQ gains
1. Comparing 111.3 (WAIS) & 103.8 (WAIS-R) = 7.5/24.5 years = 0.306
2. Comparing 105.8 (WAIS-R) & 101.6 (WAIS-III) = 4.2/17 years = 0.247
3. Comparing 102.9 (WAIS-III) & 99.53 (WAIS-IV) = 3.37/11years = 0.306

Average rate from 1953–1954 to 2006: 15.07/52.5 = 0.287 IQ points per year

Examples of calculations for WAIS-R and WAIS-III at ages 20–34
1. Ages 20–24: WAIS-R SS = 120.8 = IQ 106.8; WAIS-III SS = 115.2 = IQ 102.2 (Wechsler, 1981, p. 97). Difference = 4.6 IQ points
2. Ages 25–34: WAIS-R SS = 120.8 = IQ = 103.8; WAIS-III SS = 115.2 = IQ 100.2 (Wechsler, 1981. p. 99). Difference = 3.6 IQ points, and so forth.

Calculations for WAIS-III and WAIS-IV: all ages (Wechsler, 1997a, pp. 197–198)
1. The conversions that surround (+/−10 points) the WAIS-IV raw score total of 109.8 are 100 SS = 94 IQ and 120 SS = 105 IQ. Therefore, a range of 20 SS = 11 IQ points. 9.8/20 = .49 × 11 = 5.39 IQ points; and 94.0 + 5.39 = **99.39** as WAIS-IV IQ.
2. The conversions that surround (+/−10 points) the WAIS-III raw score total of 115.6 are 106 SS = 97 IQ and 126 SS = 109 IQ. Therefore, a range of 20 SS = 12 IQ points. 9.6/20 = .48 × 12 = 5.76 IQ points; and 97 + 5.76 = **102.76** as WAIS-III IQ.
3. Difference 102.76-99.39 = 3.37 as the gain over 11 years.
4. The actual WAIS-III mean (Wechsler, 2008, p. 102.9) is 102.9, so we have come gratifyingly close! Since that is an actual mean and our WAIS-IV estimate is eccentric in carrying over WISC-III subtests (and scoring vs. the WAIS-III tables), we will subtract the difference from the WAIS-III mean to get our simulated later performance: 102.90-3.37 = 99.53 as "WAIS-IV" mean.

The rates of gain over more than a half-century are similar: the WISC shows a rate of 0.310 to 0.325; the WAIS shows a rate of 0.287. Flynn (2009, p. 181) explains why we must give a range of estimates for the WISC after 1989.

The reader can now appreciate why 0.30 points per year is a good estimate of the rate of obsolescence of the norms of Wechsler tests in America. But there is some unfinished business:

First, are IQ gains the same at the crucial level of mental retardation, that is, for scores from about 55 to 80? Flynn (2009, pp. 134–137) shows that this is certainly true for all versions of the WISC. The pattern on the older versions of the WAIS were confused by changes over time concerning the bottom threshold of scores (Flynn, 2006). Fortunately, this has been put right. The rate of gain from the WAIS-III to WAIS-IV is the same at all IQ levels (Wechsler, 2008, p. 77).

Second, does the WAIS-III inflate IQ scores over and beyond obsolescence because they happened to get a substandard standardization sample? I have hypothesized that they did and suggested that WAIS-III scores be adjusted as follows: (1) Deduct 0.3 points per year for obsolescence. If the test were administered today this would amount to 14 years (1995 to 2009) and equal 4.2 IQ points. (2) Deduct another 2.34 points because its substandard norms inflated IQs by that amount even at the time it was standardized.

Number (2) must be revisited. When calculating WAIS-R to WAIS-III gains some ten years ago, I took my estimate from a table in which I gave Full Scale IQs for the two tests (Flynn, 1998). As Larry Weiss of the Wechsler organization and I discovered some years later (Flynn and Weiss, 2007), if you want to estimate the rate of obsolescence for a test, you should keep its basket of subtests unaltered—which is the method I have followed throughout this paper (see bottom of Table 2). This raised the rate of gain from the WAIS-R to WAIS-III to a more respectable level (2.9. to 4.2 points). In addition, my analysis of the WAIS-III norms was limited to the data in Table 1 prior to adding the new results from the WAIS-IV comparisons.

As the bottom of Table 1 shows, with those new comparisons and a revised value for the WAIS-R to WAIS-III comparison, my analysis cuts its atypical inflation of IQ scores to 1.65 points. Moreover, when compared to the WAIS-R and WAIS-IV, the WAIS-III is only 0.49 points out of line ($0.90 + 0.07 = 0.97/2 = 0.49$). And another consideration: the Wechsler organization was at pains to ensure that the WAIS-III sample included a sufficient number of subjects at low IQ levels, and this could make a difference of as much as 0.554 points (Flynn, 1998, pp. 12341235). In passing, I have not added those points on to my WAIS-R to WAIS-III estimate because one cannot be certain that this is appropriate. It is difficult enough to convince nonspecialists of the basics of allowing for obsolete norms without making choices that might appear to inflate estimates of IQ gains.

Just as I was about to exonerate the WAIS-III from the charge that its standardization sample was substandard, I received a copy of Floyd, Clark, and Shadish (2008). A group of 148 college undergraduates scored 8.64 points higher (adjusted for dates of standardization) on the WAIS-III than on the Woodcock-Johnson III, and a group of 99 subjects scored 6.77 points higher (adjusted) than on the Kaufman Adolescent and Adult Intelligence Scale. These results are very unsettling because cases are being heard where WAIS-III IQs are on record. I strongly recommend simply setting the WAIS-III scores aside. In every such case, the subject should be tested anew on both the WAIS-IV and the Stanford-Binet 5. Even if given now, the results of both will, of course, have to be adjusted for obsolescence: WAIS-IV scores lowered by 0.9 points to cover three years and SB-5 scores lowered by 2.4 points to cover eight years.

Third, the *Dauburt* memo echoes a point made by virtually every prosecution brief. It notes that while scholars use the formula of 0.3 points per year to adjust Wechsler IQs in America, they are studying groups and not individuals. Well, that is what scholars do, because when you want to make generalizations, a sample size of 1 is rather too small to be reliable. As for clinical psychologists, they deal with individuals but rely primarily on their clinical judgment and are not swayed much by the whether an IQ test gives 67 or 73. In capital cases, as we shall see, clinical assessments are largely nullified by the conflicting testimony of prosecution and defense experts. Therefore, whether an IQ is 67 or 73 can determine whether the death penalty is upheld or set aside.

The argument that adjusting an individual's IQ is some sort of leap into the unknown is based on a total lack of understanding of what an IQ is. No individual ever got an IQ score except by comparison with the performance of a group, namely, a standardization sample. If he performs at the cutting line for the bottom 2.23% of that group, he gets an IQ of 70. If the sample is not representative, it is biased and gives bad IQs. It gives bad IQs to everyone—individuals, herds, groups, flocks, and the local barbershop quartet.

It makes no difference whether the sample underperformed (and inflated IQs) because it had no college graduates or whether it underperformed because it was peopled by the lower-performing Americans of the past. In either event, if you want to salvage an individual's IQ score, you must allow for the inflation occasioned by the substandard sample. In the first case, you should compare the scores with those based on standardization samples that included all educational levels. In the

second case, you should compare the scores with those based on a representative sample of Americans today, which is to say you should adjust them for obsolescence. It is just that simple.

A final point that should be underlined a hundred times is that adjusting the IQ of an individual is no less or more accurate than adjusting the mean IQ of a group. How could it be? You use the same rate of obsolescence for both: you deduct 0.3 points for every year between the time of norming and the time of testing. If the rate is accurate, both adjustments are accurate. The original IQ of an individual may be less reliable. The day before the test, a woman may find that her husband has run off with the baby sitter. Within a large group, few will have suffered that fate, but what caused the original IQ to be unreliable has nothing to do with the effects of adjustment.

THE LAW AND THE REAL WORLD

The law must not lose touch with reality by yearning after the unattainable. Ideally, it might be best to just junk IQ scores as a criterion for mental retardation and depend entirely on clinical assessments. But we must not do this because of the nature of our adversarial system of justice. I have never seen a case in which experts for the defense found a convicted murderer mentally competent: their speech was halting and simplistic. I have never seen a case in which experts for the prosecution found a convicted murderer mentally retarded: their speech was fluent and to the point.

We all know how this game is played: "Professor Flynn, I am phoning you because we do not want anything in an e-mail until we have acquainted you with the case." You bet they don't. Before they use me, they want to find out whether I will interpret the IQ record so as to support a reprieve from execution. Often I disappoint them and am not retained. The prosecution does exactly the same. They know who has a track record of finding defendants competent and feel them out. We must give great weight to the IQ scores, because as rough as they are, they at least put a weight in the scales that is a matter of record. If judges will only understand how to read their messages, they will have at least one piece of evidence that can provide an antidote to the excesses of our adversarial system.

Ideally, we would know exactly how to adjust IQ scores for obsolescence. It is quite possible that the rate of gain on Wechsler tests is 0.275 or 0.325 points per year. But one thing we know for certain: IQ gains have not been nil. Unadjusted IQs presume that fiction. All of the evidence suggests that a rate of 0.30 is about right, and varying it from case to case lacks any rationale.

Ideally, a mechanical application of the usual rules of evidence would promote equity. But in reality, *Daubert* memos, if successful, are certain to perpetuate a monstrous injustice by making the death penalty a lottery. Justice Stewart (*Furman v. Georgia*, 1972) deserves the last word: "These death sentences are cruel and unusual in the same way that being stuck by lightning is cruel and unusual." He did not want to see them wantonly and freakishly imposed.

REFERENCES

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.

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. (2009). *What is intelligence? Beyond the Flynn Effect* [Enlarged Paperback Edition]. New York: Cambridge University Press.

Flynn, J. R., & Weiss, L. G. (2007). American IQ gains from 1932 to 2002: The significance of the WISC subtests. *Journal of International Testing, 7*, 209–224.

Furman v. Georgia. (1972). 408 U.S. 238.

Vineland. (2006). Pre-publication data from the Vineland-II manual, courtesy of S. Sparrow, Ph.D., Professor Emerita and Senior Research Scientist, Yale Child Study Center.

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

Wechsler, D. (1997a). *Wechsler Adult Intelligence Scale: Administration and scoring manual* (3rd ed.). San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1997b). *Wechsler Adult Intelligence Scale: Technical and interpretive manual* (3rd ed.). San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale: Technical and interpretive manual* (4th ed.). San Antonio, TX: Pearson.