Psychology, Public Policy, and Law
2000, Vol. 6, No. 1, 191–198

1076-8971/00/$5.00 DOI: 10.1037//1076-8971.6.1.191

THE HIDDEN HISTORY OF IQ AND SPECIAL EDUCATION
Can the Problems Be Solved?

James R. Flynn
University of Otago

Over the last 50 years, IQ gains and renorming tests have radically altered the percentage of Americans with IQs below 70. Moreover, the criterion for mental retardation was altered from 70, normed against Whites only, to 70, normed against all Americans. In fact, the proportion eligible to be classified as mentally retarded has varied from a high of 1 in 23 to a low of 1 in 213. These enormous fluctuations in the past engendered no response from practicing psychologists and no adequate response from test publishers. It must be concluded that no IQ criterion of mental retardation can be justified in terms of its correlation with impaired adaptive behavior. Because this correlation is the sole rationale of an IQ criterion, consideration should be given to the notion of abandoning IQ tests in favor of direct tests of impaired adaptive behavior. The fact that people will get quite different scores on different IQ tests can be manipulated by psychologists to suit their clients' needs.

Introduction

Evidence has put the phenomenon of massive IQ gains over time beyond doubt; data from 20 nations show not a single exception. British Raven's trends suggest that gains began no later than the onset of the Industrial Revolution. Although there are national differences in IQ gains, the most interesting differences relate to the kind of IQ test. Tests of fluid intelligence such as Raven's—tests that measure on-the-spot problem solving using patterns presumed to be recognizable across cultures—show gains of about 20 points over one generation, that is, over 30 years (Flynn, 1987, 1998a; Raven, Raven, & Court, 1993, Graph G2). Fluid tests with culture-specific items may show gains of only 10 points, but this result could be a regional difference peculiar to Scandinavia (Emanuelsson, Reuterberg, & Svensson, 1993). Wechsler performance scales show gains ranging from 9 to 20 points, and Wechsler verbal scales average gains of about 9 points. Subtest data show losses or nil gains for arithmetic; low gains for information; and negligible gains for vocabulary, at least in English-speaking countries because German-speaking countries show large vocabulary gains (Flynn, 1984, p. 46; 1987, pp. 185–186; 1990; Schallberger, 1987, p. 9; Schubert & Berlach, 1982, p. 262; Wechsler, 1992, p. 198).

IQ gains pose fascinating theoretical problems: Does the fact that these gains are largely unaccompanied by academic dividends imply that they are not true intelligence gains? Further, if they are not true intelligence gains, what does this indicate about the ability of IQ tests to compare groups for intelligence across cultural distance? However, the focus of this article is on a practical problem and

Correspondence concerning this article should be addressed to James R. Flynn, Department of Political Studies, University of Otago, P.O. Box 56, Dunedin, New Zealand. Electronic mail may be sent to jim.flynn@stonebow.otago.ac.nz.

is therefore much narrower. I contend that Wechsler IQ gains in America reveal a hidden history: They reveal (a) that American psychologists have never had a defensible IQ criterion for classifying people as mentally retarded, (b) that criteria for learning disability are suspect, (c) that criteria must be tested against certain research designs, which may discredit the whole enterprise, (d) that IQ tests must be normed frequently—perhaps too frequently to be practical—and (e) that alternatives that might appear plausible, such as predicting IQ gains or using tests normed at the same time, are no solution at all.

The Hidden History

For over half a century, the American Association on Mental Deficiency has been the dominant authority concerning the criterion for classifying people as mentally retarded. The Association states that both impaired adaptive behavior and impaired intellectual functioning should be considered. The recommended criterion of the latter has been an IQ of 70 or below, and Wechsler manuals published as recently as 1992 attest to its persistence (Wechsler, 1992, p. 8). The criterion of an IQ of 70 or below has no independent rationale—its sole justification is that it is supposed to signal the possibility of impaired adaptive behavior. Such behavior is thought to be characteristic of two or three percent of the population. Thanks to the mathematics of a normal curve, an IQ of 70 lies 2 *SD*s below the population mean of 100 and isolates the bottom 2.27% of the biologically normal population.

All of this assumes, of course, that the standardization samples used to norm IQ tests are representative of American schoolchildren. Wechsler researchers do perhaps the best job of anyone when it comes to sampling, but perfect sampling is impossible. Comparative analysis of two recent Wechsler samples suggests discrepancies of up to 2 IQ points at the mean (Flynn, 1998b, Table 3). This implies that a Wechsler IQ of 70 may isolate anything from 1.65% to 3.09% of the biologically normal population, which means that the target group could be anywhere from 27% too small to 36% too large. However, I will set this aside and assume that a score of 70 isolates the bottom 2.27% of the population on the day the test is normed.

The day after the test is normed, American IQ gains on Wechsler tests would begin to erode that percentage. Sometimes these gains are slightly greater for those test takers below 70 than for those with average IQs. Between 1947–1948, when researchers normed the Wechsler Intelligence Scale for Children (WISC), and 1972, when researchers normed the Wechsler Intelligence Scale for Children—Revised (WISC–R), children at that level (i.e., IQ = 70) gained 8.25 IQ points (Flynn, 1985, p. 240). Every year, more and more low IQ children climbed above a score of 70; indeed, by 1972, only the bottom 0.54% were eligible to qualify as mentally retarded. Then, in 1974, the WISC–R with its updated norms was published, and overnight, the percentage of those eligible to qualify as mentally retarded dramatically escalated, at which point new IQ gains began to erode the percentage eligible all over again. In 1989, researchers normed the Wechsler Intelligence Scale for Children—Third Edition (WISC–III) on a new, higher performing sample. There is insufficient data to make a precise estimate of how much, at the level of mental retardation, IQ gains had raised IQ scores by that

date. A small sample from the WISC–III manual suggests a value of 9 points (Wechsler, 1992, p. 211), which is larger than expected and sampling error may be a factor. However, using 9 points means that if an IQ of 70 isolated 2.27% in 1972, it was isolating only 0.47% in 1989. Then, the WISC–III was published, and overnight, another leap back to a higher percentage occurred.

In other words, Wechsler tests give us a score that is supposed to isolate the bottom 2.27% of the biologically normal population. Yet, at best, it does so only for an instant, after which that very same score gradually lets most of the target group escape over the course of the next 20 years. Then, thanks to a new test, it suddenly snaps back to targeting the bottom 2.27%, and the 20-year cycle starts all over again. This pattern is the hidden history of how IQ tests have been used to classify American children as mentally retarded. Actually, the pattern tells only half of the story. The other half consists of how clinical psychologists reacted during these cycles.

As far as I can determine, no clinical or school psychologist using the WISC over the relevant 25 years noticed that its criterion of mental retardation became more lenient over time. No one who began practicing late in its cycle, say in 1970, noticed that the WISC was classifying only the smallest fragment of the biologically normal population as mentally retarded. When the WISC–R appeared, scholars did administer both the WISC and WISC–R to the same participants, and they did notice, to their concern, that the old test gave inflated scores compared with the more recent one. Yet no one drew the obvious moral about psychologists in the field: They simply were not making any systematic assessment of the IQ criterion for mental retardation.

Anyone making a systematic assessment over time would have complained that a WISC score of 70 was either too harsh at the beginning of the cycle or too lenient at the end. This assumes that these IQ gains over time were not true intelligence gains; if they were intelligence gains, then the pool of those eligible to be classified as mentally retarded *should have* diminished. However, on that assumption, the publication of the WISC–R in 1974 should have caused an outcry. After all, its more demanding criterion of mental retardation suddenly plunged those who were too intelligent to be so classified back into the pool. What does this history of the silence of psychologists mean? Presumably, each of them was content to develop a purely personal trade-off between IQ scores and something else. The something else could have been the number of children they were accustomed to classify as mentally retarded, pressures from schools or parents as to which children should be classified as mentally retarded, or, one hopes, their clinical judgment on a case by case basis.

To drive the point home, I will try a bit of verbal engineering. Recall that the only rationale for using an IQ score to classify children as mentally retarded is that it is supposed to correlate with impaired adaptive behavior. And yet, as far as psychologists in the field were concerned, there simply could not have been a consistent correlation between impaired adaptive behavior and an IQ score of 70. Sadly, some may have had a blind faith in a score of 70 and compromised their clinical judgment. Probably, a larger number trusted their clinical judgment and took IQ scores with a grain of salt. In any event, one thing is certain: A collective or professional consensus about what IQ score was a valid criterion of mental retardation was not possible. Insofar as they paid any attention to IQ scores,

various psychologists must have used scores that put the pool of those eligible to be classified as mentally retarded at anything from 2.27% of the biologically normal population down to 0.5%.

These facts confer a new critical perspective on the claims the test manuals make about "evidence" for their criterion of mental retardation. In 1944, Wechsler (1944, pp. 36–48) began his career poking fun at Terman for classifying people by using numerical criteria. He noted that Terman's cutting lines all ended in zero (i.e., 70, 80, 90, etc. were used to classify those tested as feeble-minded, deficient, dull, etc.), pointed out that the odds against a statistical procedure giving that particular result are 10,000,000 to 1, and objected that Terman gave no rationale for these cutting lines rather than some other set. Wechsler then suggested a statistical criterion of mental deficiency that later evolved into the traditional criterion of 2 *SD*s below the mean. Yet he clearly saw that a statistical criterion was no less arbitrary than a numerical criterion: Putting the cutting line at a nice, neat number of *SD*s below the mean is just as absurd as wanting a number that ends in zero—unless a rationale is provided.

It was here that Wechsler appealed to evidence: He said that he had available various estimates of the incidence of mental deficiency in America, and that these gave a mean figure of about 3% of the total population. Such evidence would justify classifying about 2.27% of the biologically normal population as mentally retarded. By way of explanation, about 0.75% of the total population suffer from brain damage or chromosomal abnormality (Jensen, 1980, pp. 109–110); deducting them from Wechsler's 3% gives a value very close to 2.27%. Other evidence would have to be forthcoming, of course—for example, the 2.27% Wechsler tests isolate among the biologically normal would have to be the same 2.27% dysfunctional behavior isolates. However, when Wechsler referred to the crucial evidence for his criterion, he gave absolutely no citations. Indeed, I cannot find a cited body of evidence in any Wechsler manual from the WISC to the WISC–III.

In 1974, 30 years later, the WISC–R manual introduced a new criterion of mental retardation. Before that time, the line was drawn at 2 *SD*s below the IQ mean of *White* Americans; the WISC–R drew the line at 2 *SD*s below the IQ mean of *all* Americans, including lower-scoring minority groups. The fact that a score of 70 remained the criterion masked the fact that, on paper at least, the criterion had become less demanding by fully 4.56 IQ points; that is, a score that escapes the bottom 2.27% of all Americans falls 4.56 points below the cutting line that isolates the bottom 2.27% of White Americans—thanks to the former averaging fewer items correct than the latter. The WISC–R manual stated that its scale provides "a time-tested classification of IQ equivalents for diagnostic terms in common use" (Wechsler, 1974, p. 24). It is unstated how the same body of evidence could attest to two criteria that were 4.56 points apart. I say that the new criterion was less demanding "on paper" because the new criterion was more demanding in terms of the real world. After all, the new WISC–R norms wiped out 25 years of IQ gains. This toughened the criterion for mental retardation enormously, enough to swamp the relaxation entailed by the change from White to all-races norms. Practicing psychologists may have thought they were using a criterion 4.56 points less demanding; in fact, they were using a criterion 4.55 points *more* demanding.

The true state of play in 1974 was this: First, during the previous 25 years, no

one could have possibly been accumulating evidence for the old White American criterion of mental retardation, because it had been becoming more lenient by 8.25 IQ points. Second, Wechsler had introduced a new criterion, one he knew to be far more lenient, apparently without any evidential justification. Third, even assuming Wechsler had evidence that he did not bother to cite in favor of a more lenient criterion, the new WISC–R criterion was actually much more demanding. Indeed, it was 9.11 points higher than such evidence would have justified! Note the discrepancy of 0.86 points between the two values used in this paragraph (for an explanation, see Flynn, 1985, p. 238).

In 1991, the WISC–III manual appeared (non-U.S. editions appeared in 1992) with results for 28 mentally retarded children who took the WISC–III and for whom WISC–R scores were available. The children scored 8.9 points lower on the then-new test, and psychologists were told they should consider this "during re-evaluation of children with mental retardation who have already been assessed with the WISC–R" (Wechsler, 1992, pp. 211–212). The psychologists were not told that for every biologically normal child that they had classified as intellectually deficient the previous month, the newer norms would classify four more children as such (2.27% ÷ 0.47% = 4.8), nor were they told which assessment they should trust. To its credit, the manual was commendably honest in presenting this study, and at least empty references to a "time-tested" body of evidence had disappeared. The study itself contributes no evidence: Because the children were classified as mentally retarded partially on the basis of the WISC–R, it is hardly surprising that they have low IQs on the WISC–III.

What would a proper research design look like? The Wechsler team could identify 20 psychologists whose judgment they trust, get from each 20 children classified as mentally retarded on purely behavioral criteria, giving them a pool of 400 children to test. They would then assess the IQ scores to see if any common ceiling emerged. Ideally, something like 18 of the 20 psychologists would classify all children below a particular score as mentally retarded, and classify only the rare child who tested above that score as such. That particular WISC–III score could, at that moment in time, be recommended to school psychologists in the field as having some claim to function as a check on their clinical judgment.

However, due to the possibility of IQ gains over time, the research team would have to repeat the whole experiment after no more than 7 years. Even assuming (a doubtful assumption) that most of the 20 psychologists, using clinical judgment alone, share a common IQ ceiling at the time a test is normed, imagine what might occur over the next 7 years. Some psychologists might find their ceiling had stayed at 70, others might find that it had risen to 72, and others to 74. No one would know the magnitude of IQ gains, or whether they had persisted at all, until the next norming. Therefore, no one would know whether their ceiling for classifying children as mentally retarded had remained constant—only collective agreement at start and finish would inspire confidence. Who believes that such an outcome is at all likely, which is to say, would all of this be worth the trouble? Perhaps it would be worth the trouble if it disabused school psychologists from taking IQ criteria of mental retardation too seriously.

Predicting IQ Gains

Some might think that I am too pessimistic about predicting the magnitude of IQ gains. Why not calculate the most recent rate of gain—that prior to the current norming of a test—and use that rate to predict gains over the next decade or two, thus allowing psychologists to adjust norms and avoid obsolescence? Unfortunately, predictions simply do not work very well. When The Psychological Corporation normed the WISC–R in 1972, the rates of previous years would have predicted a 5-point gain over the next 17 years. In 1989, the norming of the WISC–III did show an average gain of about 5 points, but at the level of retardation, it suggested a gain of almost 9 points (Wechsler, 1992, pp. 198, 211). The recent norming of the Wechsler Adult Intelligence Scale (Third Edition), admittedly on adults rather than children, suggests a 3-point gain at both normal and retardate levels (Wechsler, 1997, Table 4.2). So, at present, the true value could be anywhere between 3 and 9 points. The fact must be faced: IQ gains are not caused by something analogous to a physical law but are phenomena whose causes are unknown. Sometimes gains suddenly stop, as recently in Sweden; sometimes they accelerate for prolonged periods, as in the Netherlands between 1952 and 1982 (Emanuelsson, Reuterberg, & Svensson, 1993; Flynn, 1987, p. 172).

Some psychologists assume that they can allow for the effects of IQ gains, at least for some purposes, so long as tests are equally obsolete. That is, they seek safety in using only those tests that were normed on the same sample at the same time. *They are mistaken.* So long as there are differential rates of gain on different kinds of tests, misdiagnosis is likely. Once again, the WISC–III manual illustrates the point. The WISC–III was administered to 99 children diagnosed as suffering from learning disabilities or reading disorders. These children tended to do badly on 4 subtests—arithmetic, information, coding, and digit span (AICD). The manual defines an AICD profile as present when a child receives scores on all four of these subtests equal to or less than the lowest score on any of the remaining seven subtests (mazes and symbol search are excluded). A partial AICD profile is present when three of the four designated subtests meet this criterion. The partial AICD profile occurred in 20.7% of the learning-disabled/reading-disordered sample, compared to 5.6% of the standardization sample, which was presumably representative of American children in general. The manual suggests that when an AICD profile is present, the possibility of a learning disability should be investigated (Wechsler, 1992, pp. 212–213).

The WISC–III manual also contains a table that reveals score gains on various subtests between 1972, when the WISC–R was normed, and 1989, when the WISC–III was normed. During that period, the subtests showing the lowest gains are information, at −0.3 scaled score points (a loss); digit span, at +0.1; and arithmetic, at +0.3. With the exception of vocabulary (+0.4), all of the other subtests show gains that are two to six times as great (+0.6 to +1.9). Coding is the only AICD subtest with a significant gain (+0.7), standing at sixth in terms of lowest gains (Wechsler, 1992, p. 198). In sum, if the WISC–III standardization sample had been tested on the WISC–R, its general tendency would have been to exhibit a partial AICD profile (i.e., to score lower on arithmetic, information, and digit span than on any other subtest). Suspicion of learning disability and reading

disorders would have been rife in a sample typical of American children in general. It must be emphasized that the pattern of differential gains on subtests prevalent between 1972 and 1989 may not hold between 1989 and 2006, assuming that as a likely date for norming the fourth edition of the WISC. With good luck, only two of the four AICD subtests will show the lowest gains. On the other hand, if the recent pattern of gains has persisted, it will become significant over the next few years. With bad luck, differential gains on subtests may now be occurring that are even more destructive of proper diagnosis than the old.

Summary

I end this article with the message with which I began. Behind the facade of constancy, the hidden history of IQ testing shows huge fluctuations in the IQ criterion of mental retardation and paucity of evidence for any particular criterion. It also inspires doubt about the use of differential scores on Wechsler subtests to isolate the learning disabled. If IQ tests are to play a role of minimal respectability, researchers will have to renorm them every 7 years. Furthermore, the renorming will have to be accompanied by an ambitious research design, one which may well show that no particular IQ criterion of mental retardation can claim a solid evidential foundation.

Looking back over the last 50 years, what has been the human significance of all of this? Setting aside the very real possibility of sampling error, the proportion of Americans eligible to be classified as mentally retarded has fluctuated from something like 1 in 23, circa 1949 (2.27% of White Americans meant 16.90% of Black Americans and 4.32% of all Americans), to 1 in 213, circa 1989 (0.47% of all Americans). The practice of psychologists in the field would have done much to moderate fluctuations in the numbers actually classified. Nonetheless, it is certain that over the past 50 years, literally millions of Americans evaded the label of mentally retarded designed for them by the test manuals. Whether this was good or bad depends on what one thinks of the label. Some will say millions avoided stigma. Others will say that millions missed out on needed assistance and classroom teachers were left unaided to cope with pupils for whom aid was needed. I leave this balance sheet for those who consider themselves competent bookkeepers.

Even if nothing is done, however, psychologists are now empowered. Some states require an IQ score below 70 before they will give a benefit to someone diagnosed as mentally disabled. Psychologists who want their clients to be eligible should pick the most recent test and score it, allowing for obsolescence. On the other hand, psychologists who want pupils to escape the label of mental retardation should pick the oldest test they can get away with. Depending on the tests chosen, there will be no problem in rigging IQ scores by at least 10 points. It is very unlikely that the bureaucracy will ask any questions. However, if psychologists are really interested in whether someone suffers from impaired adaptive behavior, they should forget IQ scores and use the kind of "test" described by Jensen (1981, p. 65). When children say that their favorite sport is baseball, have a chat and see whether or not they have grasped the concept of a double play.

References

Emanuelsson, I., Reuterberg, S.-E., & Svensson, A. (1993). Changing differences in intelligence? Comparisons between groups of thirteen-year-olds tested from 1960 to 1990. *Scandinavian Journal of Educational Research, 37,* 259–277.

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

Flynn, J. R. (1985). Wechsler intelligence tests: Do we really have a criterion of mental retardation? *American Journal of Mental Deficiency, 90,* 236–244.

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

Flynn, J. R. (1990). Massive IQ gains on the Scottish WISC: Evidence against Brand et al.'s hypothesis. *Irish Journal of Psychology, 11,* 41–51.

Flynn, J. R. (1998a). Israeli military IQ tests: Gender differences small; IQ gains large. *Journal of Biosocial Science, 30,* 541–553.

Flynn, J. R. (1998b). 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.

Jensen, A. R. (1980). *Bias in mental testing.* London: Methuen.

Jensen, A. R. (1981). *Straight talk about mental tests.* New York: Free Press.

Raven, J., Raven, J. C., & Court, J. H. (1993). *Manual for Raven's Progressive Matrices and Vocabulary Scales (Section 1).* Oxford, England: Oxford Psychologists Press.

Schallberger, U. (1987). HAWIK and HAWIK–R: Ein empirischer Vergleich [HAWIK and HAWIK–R: An empirical comparison]. *Diagnostica, 33,* 1–13.

Schubert, M. T., & Berlach, G. (1982). Neue Richtlinien zur interpretation des Hamburg Wechsler—Intelligenztests fur Kinder (HAWIK) [New guidelines for the interpretation of the Hamburg Wechsler Intelligence Tests for Children (HAWIK)]. *Zeitschrift fur Klinische Psychologie, 11,* 253–279.

Wechsler, D. (1944). *The measurement of adult intelligence* (3rd ed.). Baltimore: Williams & Wilkins.

Wechsler, D. (1974). *WISC–R manual.* New York: Psychological Corporation.

Wechsler, D. (1992). *WISC–III: Wechsler Intelligence Scale for Children—Third Edition: Manual* (Australian adaptation). San Antonio, TX: Psychological Corporation.

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale—Third Edition: Manual.* San Antonio, TX: Psychological Corporation.

Received March 12, 1998
Revision received August 15, 1998
Accepted August 17, 1998 ■