CHAPTER FOUR

The Flynn Effect and the Shadow of the Past: Mental Retardation and the Indefensible and Indispensable Role of IQ

James R. Flynn* and Keith F. Widaman†

Contents

Abstract

Gains in IQ over time render test norms obsolete within a decade of the publication of an intelligence test. Obsolete norms inflate IQs and drain the pool of those eligible to be classified as having mental retardation (MR). As a result, many are missing the services they need and capital offenders are being

* Department of Political Studies, University of Otago, PO Box 56, Dunedin, New Zealand
† Department of Psychology, University of California at Davis, 265 Young Hall, Davis, CA 95616, USA

International Review of Research in Mental Retardation, Volume 35
ISSN 0074-7750, DOI: 10.1016/S0074-7750(07)35004-0

executed who should qualify as mentally incompetent. The history of IQ gains and IQ criteria for MR pose deeper and more problematic questions: Can we salvage IQ tests as measures of intelligence; can we find an IQ criterion for MR that has external validity; can we find a mental test that measures moral culpability? All solutions involve an agonizing reappraisal of present practice.

1. Introduction

The Flynn effect refers to the well-documented improvement in performance on tests of intelligence that takes place over decades (Flynn, 1984, 1987, 1998). This phenomenon of massive IQ gains over time has implications for all assessment, but this chapter focuses on the diagnosis of mental retardation (MR). With that in mind, we undertake the following:

1. A description of the havoc that obsolete IQ tests and their norms have on the current classification of people as having MR.
2. Particular attention is due to capital offenders on death row. From this discussion, it will emerge that an IQ criterion of MR is indispensable if justice is not to be fatally compromised. Therefore, a way of compensating for obsolete norms will be recommended. However, that solution assumes that we have a defensible IQ criterion of MR, specifically a criterion that reflects the impaired reasoning or judgment exhibited by persons with MR.
3. Pursuing the history of IQ and MR back to 1947, we show that clinical psychology does not have and never has had a defensible criterion in terms of external validity, such as one referring to levels of reasoning.
4. That history also poses a dilemma that must be solved if IQ tests are not to be discredited as measures of intelligence. Trends on the various Wechsler Intelligence Scales for Children (WISC) subtests will suggest a solution.
5. However, when we push IQ gains back to their origin, that is, to a time no later than 1900, the dilemma will reemerge. Piagetian concepts and techniques of modern test theory will shed light on performance on tests like Similarities and Raven's and suggest workable solutions.
6. Our solution will indicate how the use of an IQ criterion of MR can be made defensible. But this cannot be done without tears.

2. The Flynn Effect and MR Diagnosis

2.1. MR becomes a matter of life and death

The best evidence of massive American IQ gains comes from the ever-improving performance by succeeding cohorts on the WISC and its successors. From the WISC (normed in 1947–1948), through the WISC-R (1972), through the WISC-III (1989), to the WISC-IV (2002), we can compare the performance of the normative samples (Flynn, 2006a, Table 1). Whenever the same group took both an older and a newer version of the test, they found it easier to exceed the norms set by the older standardization samples. A group with a mean IQ score that was only average when scored against the WISC-IV norms achieved a mean that was well above average when scored against the WISC-III norms; a group that was average on the WISC-III was above average on the WISC-R; and so forth.

Clearly, as we go back into the past, the norms become weaker, which means that representative samples of American schoolchildren set lower and lower standards of performance. Conversely, as we go from past to present, American children attained higher standards of performance. The rate at which performance improved over the 55 years between 1947 and 2002 hardly varied. For example, a group only average on the 1972 norms (with a mean of 100) would score at about 107.50 on the 1947–1948 norms, for a rate of gain of 0.3 points per year (7.50 divided by 25 years = 0.3).

This means that the IQ an individual receives is like a lottery, with the outcome dependent on when he or she was tested, what test was administered, and in what year that test was normed. Assume a group of children born in 1990 who were exactly normal, that is, representative of the U.S. population, and therefore deserved an IQ of 100. In 1996, at age 6, they take the WISC-R (the norms for which were by then 24 years obsolescent) and get a mean IQ of 107 ($24 \times 0.3 = 7.2$), a score fully 7 points more than they merit. It would be unusual not to use the WISC-III that had been published in 1991, but not unheard of. School psychologists and the administrations for which they work sometimes continue to use copies of outdated IQ test protocols that were in stock before buying new tests.

In 2000, the children, now aged 10, take the WISC-III. Its norms are by now 11 years obsolescent, so they get a mean score of 103 ($11 \times 0.3 = 3.3$). Finally, in 2003, at age 13, they take the then-new WISC-IV whose norms are only one year obsolescent and achieve a mean score of 100.3 ($1 \times 0.3 = 0.3$). So, finally, they get a score fairly close to the one they deserved all along. However, schoolteachers and administrators are dismayed by the decline in their mean IQ and are trying to find what went wrong with the school environment. And yet, this "decline" is purely an artifact. It merely reflects the time that had passed between the year they took a given test and the year in which that test was normed.

Those familiar with the mathematics of a normal curve will know that modest differences near the mean of a distribution do not have a great effect on the proportion of persons meeting or exceeding a cutoff point that is close to the mean. But, comparable score changes can make a substantial difference on the proportion of the population meeting or exceeding a cutoff point if that cutoff point is far below (or above) the mean. For example, take a population whose scores have a normal distribution with

a mean of 100 and standard deviation (SD) of 15. By definition, 50% of the population will fall below 100 and 2.27% will fall below a score of 70. If 5 points (or 0.33 SD units) were added to each person's score, the mean would now be 105 and the SD would remain at 15. The result would be that 37% of the population would now fall below 100 and 1% below 70. Note that the effect near the mean reduces those that fall below 100 by about one-fourth ($37 \div 50 = 74\%$, so 26% of those who formerly had met the cutoff score of 50 no longer do so). But the percent below 70 and therefore eligible to be classified as mentally retarded has been reduced by more than one-half ($0.99 \div 2.27 = 44\%$, so 56% of those who had formerly met the cutoff score of 70 no longer do so).

Returning to our hypothetical class of children, in 1996 at age 6, taking the WISC-R inflated their IQs by over 7 points or 0.48 SD units. At that point, only 0.66% of them were eligible to be classified as having MR. By the time they were 13 in 2003, taking the WISC-IV inflated their IQs by only 0.3 points or 0.02 SD units and 2.17% were eligible to be classified as having MR. Fully 3.3 times as many were at risk even though their actual performance on IQ tests, that is, a performance that exactly matched that of their age cohort, had not altered at all.

Flynn (2000) calculated a worst case scenario taking into account obsolescence of norms and the likelihood that gains may have been slightly greater at IQ levels near the cutoff score for diagnosing MR, plus the fact that the criterion of MR was altered by moving from norms based on white individuals only to norms based on representation by all racial and ethnic groups. He showed that between 1947 and 1999, the proportion of the population eligible to be classified as MR had fluctuated from 1 in 23 (4.35%) to 1/213 (0.47%), and this assumes that no one continued to use outdated protocols of a test whose successor had been published. As a description of the fate of those being assessed, "lottery" seems too weak a word.

The above scenarios assume that IQ gains for those at low IQ levels are as robust as gains found for those who score near the mean. Flynn (2006b, Table 2) reviewed extant data and found that this was essentially true for all forms of the WISC. Two factors rendered comparable comparisons using WAIS data more complex. First, it is more difficult to obtain optimal standardization samples of adults because a representative sample is not sitting in classrooms awaiting you, as is the case with the WISC. Second, different scoring practices at low IQ levels have been employed on various forms of the WAIS (e.g., the minimum IQ a subject can get without any right answers has varied), and these scoring variations contribute to greater score volatility at low IQ levels.

Bringing some order out of this chaos is now a matter of life and death. The U.S. Supreme Court in *Furman v. Georgia* (1972) held that the death penalty must be imposed with consistency and with due regard to the

culpability of those who suffer its consequences. Thirty years later, in Atkins v. Virginia (2002), the Court held that the Eighth Amendment to the U.S. Constitution forbids the death penalty for those who suffer from MR. Subsequently, in Walker v. True (2005), the Fourth Circuit Court of Appeals held that, in applying this standard, the "Flynn effect" must be taken into account. What they meant by this was that if a capital offender's IQ had been inflated by administering a test with obsolete norms, he should not be executed just because he had suffered from the bad luck of being given a test with out-of-date, rather than contemporary, norms. For example, the fact that someone had been given the WISC (and scored above 70) rather than the WISC-R (on which that very same person would have scored below 70) would clearly be relevant when determining whether the IQ score met criteria for a diagnosis of MR.

Whether or not scores are adjusted to take obsolescence of norms into account can determine the fate of an offender. We will illustrate this by using an actual case altered only to update it so that the tests administered correspond to those given to current defendants. John Doe was convicted of murder and sentenced to death contingent on a determination of whether or not he was mentally retarded. He was born in 1984. At age 11 in 1995, Dr. Mary Smith (the school psychologist) gave him the WISC-R. This may seem odd given that the new (at that time) WISC-III had been published in 1991. Perhaps her school had limited funds and Dr. Smith needed to exhaust her supply of the protocols of the older version of the WISC before purchasing the new edition.

Dr. Smith had been taught to assess adaptive functioning independently of IQ scores. However, she found it difficult to compartmentalize the two. Her report notes John's poor performance in reading and arithmetic despite extra tutoring. But then, she rejects a diagnosis of MR on the basis of a WISC-R IQ score of 75. Today, we know that, thanks to 23 years of obsolescence of its norms, use of the WISC-R inflated Johnny's score by 7 points and his score should have been lowered to 68, easily in the MR range. Johnny is likely to be executed simply because his school's budget did not extend to purchasing the latest test. However, he might have had bad luck anyway. If tested at age 6 in 1990, there would have been no alternative to the WISC-R and his fate would have been the same.

The consequences of misclassification due to obsolete norms extend beyond the death penalty. As Kanaya, Scullin, and Ceci (2003) pointed out, each year 2 million children are tested for special education, including MR services, and in a given school year over 600,000 actually receive MR services. Adults who are classified as having MR are eligible for social security disability benefits and are ineligible for military service. If those with inflated IQs, thanks to obsolete norms, are not classified as having MR, the government saves millions of dollars because it does not have to provide special services to these misclassified individuals. The other side of the coin

is that the misclassified persons do not get the help they surely need. As for the military, the use of obsolete norms could lead the armed forces to enlist thousands of people who, at least in their opinion, lack the level of mental ability needed to make correct decisions on the battlefield. The losses in money and lives are potentially huge.

Kanaya et al. (2003) confirmed that, when the WISC-R was replaced by the WISC-III, the scores of low-IQ children dropped by at least the amount predicted, a rate of 0.30 points per year of obsolescence. Because the norms of the latter test were set 17 years after those of the former, the predicted loss would be 5.10 IQ points, and Kanaya et al. found an average loss of 5.55 points (0.37 SDs). If IQ were the sole criterion of MR, and everyone had used the new WISC-III as soon as it became available, a score fluctuation of 0.37 SDs predicts that the number classified as MR would rise from less than 1% (0.89) in 1990 to almost 2% (1.97) in 1991; that is, there would be a doubling overnight of those who qualify for MR services. Kanaya et al. selected a large, economically and geographically diverse sample of students tested for special education. They found that school psychologists who first tested a child on the WISC-R and then retested using the WISC-III were twice as likely to submit a recommendation of MR, when compared to those who retested using the same test.

However, Kanaya et al. (2003) also found that 88% of students were still being given the WISC-R in 1991, 41% in 1992, and that even in 1995, the old test was still in use. Where a single score existed, school psychologists showed a greater reluctance to trust the lower WISC-III IQ than the higher WISC-R IQ. Only half of students with an IQ score below 70 were actually recommended for a diagnosis of MR. Note that by 1996, the WISC-III was itself seven years obsolete and inflating IQs by 2.1 points. So, someone who retested a child on the WISC-III in 1996 and compared the result to a WISC-R score obtained in say 1989 (17 years of obsolescence equals 5.1 points) would notice a discrepancy of only 3 IQ points. A difference of 3 points is well within the margin of measurement error for the WISC and WISC-R. The discrepancy would neither reveal what was happening nor change what was happening: Between 1991 and 1995, many thousands of children did not receive a diagnosis of having MR simply because of the test they took.

Scullin (2006) collected data from all 50 states plus the District of Columbia to trace trends concerning the percentage of students enrolled in MR programs. He found that a steady and general decline during the 1980s turned into an increase in the early 1990s in 43 states and Washington, DC. MR rates in 1993 were only 62% of the rate for 1981–1982, but had rebounded to 80% of that rate by 1999. Note that the weakening WISC-R norms predict a decline as the 1980s progressed and that the growing dominance of the new WISC-III predicts an upward trend beginning about 1994. This was precisely the trend found by Scullin.

By 1999, the percentage of children and adolescents receiving MR-related services should have risen to that of 1982. In 1999, the WISC-III norms were 10 years out of date, exactly the same as the 10-year-old WISC-R norms in 1982. The reason that MR diagnoses reached only 80% of their old level was that the new test was swimming against a tide. The diagnosis of learning disability was replacing the diagnosis of MR, thanks in part to the reluctance of school districts to assign the latter label, particularly to minority children. If those children who avoided the label of having MR find themselves at age 25 on death row, they will not be grateful.

Obsolete norms play havoc with diagnoses other than MR. The WISC-III manual notes that children with learning disabilities or reading disorders tend to do poorly on the four subtests of Arithmetic, Information, Coding, and Digit Span (the AICD profile; Wechsler, 1992). Getting the lowest scores on three of the four subtests constitutes a partial AICD profile. The WISC-IV technical manual states that low scores on Arithmetic, Information, Vocabulary, and Letter-Number Sequencing characterize reading disability; and that low scores on Arithmetic, Information, and Comprehension are associated with expressive language disorder (The Psychological Corporation, 2003, pp. 79–82).

Trends over time reveal that all of the above subtests, except Coding and Comprehension, have shown virtually nil gains over time. Time-related trends for Letter-Number Sequencing are unknown because it is a new subtest (Flynn, 2006a, Table 1). Huge gains on all of the remaining subtests are the cause of the massive Full-Scale IQ gains on record (Flynn, 1984, 1987). In other words, after the WISC-III's norms became obsolete, perfectly normal children started to show a partial AICD profile. If they were typical of their cohort, they tended to score closer to the old norms on Arithmetic, Information, and Digit Span than on any other subtest. And after the WISC-IVs norms become obsolete (circa 2015), we have reason to believe that normal children will tend to look as if they have reading or language disorders. They will tend to do worse on Arithmetic, Information, and Vocabulary, a bit below their average on Comprehension, with Letter-Number Sequencing unknown.

2.2. A temporary expedient

Some might argue that a simple solution exists to the problem of obsolete norms: Dispense with IQ tests and assess MR on adaptive behavior alone. Whatever the merits of this proposal, it would perpetrate a terrible injustice. America's adversarial legal system makes it highly likely that, when prosecution and defense experts interview defendants, they will reach opposite conclusions about whether MR is present. In John Doe's case, the defense psychologist noted the hesitancy and conceptual vagueness typical of

those who have MR; the prosecution psychologist found him far too alert and fluent to have MR. Therefore, IQ scores were crucial.

On the face of it, John Doe's case looked hopeless. Recall that, at age 11 in 1995, he received a Full-Scale IQ score of 75 on the WISC-R. On death row, at age 21 in 2005, both psychologists administered the WAIS-III; the defense psychologist scored him at 70, and the prosecution psychologist at 72. In fact, all of his IQ scores indicate MR. The WISC-R score should be put at 68 thanks to 23 years of obsolescence of its norms, and the WAIS-III scores should be reduced to 67 and 69 respectively thanks to 10 years of obsolescence. In addition, Flynn (2006b) reported evidence that the WAIS-III inflated IQs by 2.34 points even at the time it was normed due to a substandard normative sample. Taking this into account would lower the WAIS-III scores to a bit below 65 and 67, respectively.

In order to give justice some chance of being done, we must salvage IQ scores; and to keep IQ scores from being deceptive, they must be lowered for obsolescence. To aid jurists, Flynn (2006b) proposed a simple formula: Test Score $-$ $(I \times 0.3)$ = IQ. The letter "I" stands for the interval between when the test was normed and when the subject was tested. The test must be a Wechsler (e.g., WISC or WAIS) or Stanford–Binet test normed in America. As Flynn (2000b) showed, we cannot have the same confidence that WAIS scores have become obsolescent at 0.3 points per year that we have concerning WISC scores. But the rate of 0.3 points per year is our best (even if rough) estimate, and to make no adjustment at all would leave capital offenders at the mercy of IQ scores that are clearly inflated. As we have seen, if the Wechsler test happens to be the WAIS-III, an additional 2.34 points should be deducted—because the WAIS-III inflated IQs by that amount even at the time it was normed.

You may wish to know the true cutting line for MR at the time a test was administered, that is, what score was at 2.0 SDs below the mean. The formula then becomes: $100 + (I \times 0.3) - 30$. If the test is the WAIS-III, add an extra 2.34 points. For example, if the WAIS-III were administered in 2005: $100 + (10 \times 0.3) + 2.34 = 105.34$; and that minus $30 = 75.34$ should be used as the true cutting line.

2.3. The history of the bottom 2.27%

Adjusting for obsolescence is merely a temporary expedient. IQ gains over time pose deeper problems that must be faced if intelligence and its attendant IQ score are to be salvaged as a criterion of MR.

The criterion of an IQ of 70 or below has no intrinsic rationale. Its selection is probably best thought of as an implicit agreement between professionals and policy makers that only a relatively small percentage of persons who exhibit the most extreme impairment on tests of intelligence can and ought to receive services under the rubric of MR. Therefore, the

criterion is more a matter of sociology and social engineering than a precise indicator of the level of impaired reasoning that persons with IQs of 70 or below exhibit. It is, of course, not an entirely arbitrary choice. It is also supposed to signal the likelihood of impaired adaptive behavior.

In effect, professionals are confident that people in at least the bottom 2.27 of the population are characterized by impaired reasoning and adaptive behavior, or more accurately, at least the bottom 2.27% of the biologically normal population are so described. There is another group whose members suffer from MR because of specific genetic or other organic or biological factors, and inclusion of this group brings those who would actually score at 70 or below closer to 3% of the population. What follows simplifies by focusing on the fact that 70 is 2.0 SDs below the "normal" mean and isolates the bottom 2.27% of the "normal" population.

As we have seen, a WISC score of 70 probably does do a good job of isolating the bottom 2.27% on the dimension of general intelligence, but it does so only during the year in which the test is normed. Thanks to obsolescence of norms and other factors, a score of 70 or below isolated anything between the bottom 0.47% and the bottom 4.35% during the period between 1950 (when the WISC was published) and 1985 when the problem of obsolescence was made public (Flynn, 1985).

As far as we can determine, over the relevant 35 years, no clinical or school psychologist using the various WISC tests noticed that the percentage of the population meeting the IQ criterion of MR was fluctuating wildly over time. No one who began practicing late in WISC era, say in 1970, noticed that the test scores resulted in only a small fragment of the biologically normal population receiving an MR diagnosis. When the WISC-R appeared, scholars did administer both the WISC and WISC-R to the same subjects and they did notice, to their concern, that the old test gave inflated scores compared to the recent one. But no one drew the obvious conclusion that psychologists in the field simply were not making any systematic assessment of the accuracy of the IQ criterion for MR, that is, its accuracy in terms of isolating the correct percentage eligible to be classed as having MR.

Some scholars, such as Hamm et al. (1976, p. 7) were alarmed: "Results from the present study support Doppelt and Kaufman's conclusion that the WISC-R typically yields lower IQ scores for children who function in the EMR range. Thus, more students will be classified as 'mentally deficient' as a result of its administration. Such findings suggest that considerable care be exercised when assigning special class placement based largely upon WISC-R scores. There may be a need for reevaluation of criteria for special class placement." Excellent advice, but it did not go deep enough in its diagnosis of causes for alarm.

If clinical psychologists were making a systematic assessment of whether the IQ criterion for having MR was truly matched by impaired adaptive

behavior, they should not have had to wait for the appearance of the WISC-R (and its lower scores) to tell them something was wrong. They should have noticed either that a WISC score of 70 was too harsh at the beginning of its era (was classifying children who were not impaired) or was too lenient at the end (was failing to classify children who were clearly impaired). What does the silence of psychologists mean? It means that psychologists could not possibly have been finding a consistent relation between a particular IQ score and the level of intellectual or socially adaptive behavior it was supposed to indicate. In 1949, a child who got a WISC IQ of 70 was almost 2.0 SDs below the mean. By 1974, a child who got a WISC IQ of 70 was 2.53 SDs below the mean. So, which cutoff score criterion had external validity for indicating MR on the basis of a correlation with impaired adaptive behavior? Clearly neither did, in that psychologists were as happy with one as with the other. And yet one cutoff score isolated a pool of 2.27% of persons with a label of having MR and the other a pool of 0.57%.

That no coherent criterion was operational in the field could be interpreted in any one of several ways. First, the failure to recognize the problem might mean that the possibility of a collective or professional consensus about IQ scores and MR is impossible. This seems unlikely given the general consensus that an IQ score around 70 is a reasonable cutoff score for supporting a diagnosis of MR. Second, small numbers may have been the culprit. MR is an uncommon occurrence, affecting approximately 2 children in every 100. Unless an individual psychologist tested a very large number of children, fluctuations in the number of children identified would be difficult to detect. Suppose that 1 child in 100 was identified as having MR one year, and 3 children in 100 were identified the next year. Even though this represents a 3:1 ratio in rate of identification, such small fluctuations could easily be due to chance in the clinical experience of an individual school psychologist.

Of course, someone, particularly persons associated with the diagnostic test, should have been tracking the rate of identification of children as having MR "in the large," for example, at the state or national level. That no psychologist at any level publicized the changing rate of identification using the WISC is puzzling indeed. The answer may lie in the persistent pressure to stop labeling children as having MR, due to the negative connotations of the label. This certainly contributed to a *zeitgeist* in which falling rates of identification might go unnoticed or, at least, considered unworthy of mention.

These facts confer a new critical perspective on the claims the test manuals make about "evidence" for the criterion of MR. In 1944, Wechsler (1944, pp. 36–48) began his career by poking fun at Terman for using numerical criteria to classify subjects. He noted that Terman's cutting lines all ended in zero (70, 80, and 90 were used to classify subjects as feeble-minded, deficient, and dull) and that the odds against a statistical procedure

giving that result were 10,000,000 to 1. He also objected that Terman gave no evidential rationale for these cutting lines rather than some other set of cutoff points.

Wechsler (1944) then suggested his own statistical criterion of mental deficiency, one that later evolved into the traditional "2.0 SDs below the mean." At this point, we must defend Terman's sanity. Wechsler did not like nice, neat numbers, but what about putting cutting lines at a nice, neat number of SDs below the mean: 2.0 SDs (IQ = 70) for mildly MR, 3.0 SDs (IQ = 55) for moderately MR, and 4.0 SDs (IQ = 40) for severely MR? Certainly, wanting SDs that end in even SD units is just as absurd as wanting IQ scores that end in zero. It was a relief when Wechsler went beyond a spurious debating point and appealed to evidence. He stated he had available various estimates of the incidence of mental deficiency and these gave a mean figure of about 3% of the total population. This would justify classifying about 2.27% of the biologically normal population as mentally retarded. However, as odd as it seems, Wechsler (1944) provided absolutely no citations. No one knows what studies he had in mind or whether they actually supported his contention.

In 1974, 30 years later, the WISC-R manual introduced a new criterion of MR. Prior to 1974, the line was drawn at 2.0 SDs below the IQ mean of white Americans; now it was drawn at 2.0 SDs 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 was less demanding: One must score 4.56 IQ points higher to avoid being in the bottom 2.27% of white Americans than in the bottom 2.27% of the lower-scoring all Americans (Flynn, 1985). The manual states that the scale provides "a time-tested classification of IQ equivalents for diagnostic terms in common use" (Wechsler, 1974, p. 24). How the same body of evidence could attest to two criteria that were 4.56 points apart was unstated.

Although the new criterion was less demanding "on paper," it was more demanding in terms of the real world. After all, by 1974, the WISC norms had accumulated 26.5 years of obsolescence. When these weakened norms were swept away, the criterion for MR was enormously toughened, enough to swamp the relaxation entailed by going from white to all-races norms. Practicing psychologists may have thought they were using a criterion 4.56 points less demanding but, in fact, they were using one that was 4.55 points more demanding. At the cutting line for MR, Flynn (1985) showed that the WISC-R set norms about 9 points more demanding than the old WISC norms.

To summarize, in 1974, this was the state of affairs: First, during the previous 25 years, no one could possibly have been accumulating evidence for the old white American criterion of MR, because it was becoming more lenient by 8.25 IQ points (somewhat more than the 7.5 points of IQ

gains at the level of the mean). Second, Wechsler had introduced a new criterion, one he thought to be far more lenient, apparently without any evidential justification. Third, even assuming Wechsler had evidence he did not bother to cite in favor of a more lenient criterion, the new WISC-R criterion was actually much more demanding. Indeed, at the level of MR, it was 9.11 points higher than such evidence would have justified! The alert reader will note a discrepancy of 0.86 points between the two values used in this paragraph. As Flynn (1985, p. 238) argued, a special effort to include low-IQ subjects in the WISC normative sample raised WISC IQs and this inflation was absent in the norms of the WISC-R.

In 1991, the WISC-III manual appeared, and it reported results for 28 children with MR who took the WISC-III and for whom WISC-R scores were available. The children scored 8.9 points lower on the new test and psychologists were told they should consider this "during re-evaluation of children with MR 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 they classified as "intellectually deficient" the previous month, they would now classify four more as such (2.27% divided by 0.47% = 4.8), nor were they told which assessment they should trust. On the credit side, the test manual was commendably honest to present this study and to delete empty references to a "time-tested" body of evidence. The study itself contributes no evidence. Because the children were classified as having MR partially on the basis of the WISC-R, it is hardly surprising they had low IQs on the WISC-III as well. If they had been classified with MR on purely behavioral criteria, their comparative scores might have told us something.

In 2003, the WISC-IV technical manual appeared, and comparative data on 120 children with MR were reported. Although the criteria for selection of these children were not fully described, it appears that the children had IQ scores on "some standard test" showing that they were almost evenly divided between a group with IQs in the range of 40 or below and a group with IQs from 41 to 70 (The Psychological Corporation, 2003, pp. 79–82). Similar problems arise as with the WISC-III study: If other IQ tests played a large part in their classification as having MR, the children should certainly get low scores on the WISC-IV simply because scores from different IQ tests exhibit high intercorrelations. But this kind of intercorrelation does not confront the problem of which cutting line for MR is really valid: the tougher cutoff score in use soon after a test is normed or the easier one in use just before the test is to be replaced.

In sum, when norms are fresh rather than obsolete, a score somewhere between 60 and 75 is probably a decent criterion of MR based on impaired reasoning and low levels of adaptive behavior. But, no one has yet made a strong case for the external validity of any particular score in terms of the precise level of impaired reasoning that individuals with low IQ are likely to exhibit.

As noted above, external validity claims seem to reflect a general agreement among all concerned (e.g., psychologists, policy makers) that a particular proportion of the population should be identified as having MR. Such external validity claims certainly do not reflect any established relation between a given IQ score criterion and the particular forms of impaired reasoning that those below the cutoff score tend to exhibit. An IQ score of 60 is 2.67 SDs below the mean of the biologically normal population and would isolate the bottom 0.40%. An IQ score of 75 is 1.67 SDs below the mean and would isolate the bottom 4.75%. As professionals in this field, we should demand research that would identify where in this score range a defensible cutoff value would lie.

2.4. How many of our grandparents had MR?

Thus far, we have merely questioned whether we have ever known where to draw the line to get an IQ criterion of MR that has external validity. Now we go deeper and ask whether it makes sense to even try. IQ tests cannot help diagnose MR unless there is a plausible case that they measure intelligence. Therefore, the very magnitude of IQ gains over time poses a paradox. For example, let us take the children aged 6–16 who were used to standardize the WISC-IV as our point of reference. That gives the current generation, American children in 2002, a mean IQ of 100 by definition. Their parents would have been 4–14 in 1972 and would have been used to standardize the WISC-R. Using a rate of gain of 0.3 points per year, their mean IQ would have been 91 against recent norms. On average, the grandparents of the WISC-IV children would have been 8–18 in 1948 and used to standardize the WISC. Their mean IQ would have been 83.5 against recent norms.

If this grandparent generation really had that low a level of intelligence, 18.41% of them would have had an IQ of 70 or below and, adding in a few with specific genetic defects, the real total would have been 20%—or one person in five. Anyone who lived at that time, and taught a mainstream class, knows that this is absurd: Nothing approaching 20% of the members of our grandparents' generation exhibited seriously impaired behavior.

An obvious way out of the dilemma is to dismiss IQ gains as an artifact. However, these gains are not artifacts in any normal sense of the word. The leading case for artifactual status is to attribute IQ gains to growing test sophistication. Test sophistication has to do with feeling comfortable with the format of IQ tests, or whoever administers them, or using your time better, or trying harder in the test room. The 20th century has seen us go from people who have never taken a standardized test to people bombarded by them, and a small portion of score gains in the first half of the century was undoubtedly due to growing test sophistication.

However, since 1950, the role of test sophistication in accounting for IQ gains has been relatively modest. In the United States, gains have been steady at least since 1932 (Flynn, 1984). If gains are due to test sophistication, they should show a certain pattern. When naive subjects are first exposed to IQ tests, they gain a few points; but, after that, repeated exposures show sharply diminished returns. Gains in the United States show no such pattern and other nations show just the reverse. For example, The Netherlands shows a huge rate of gain escalating decade after decade from 1952 to 1982 (Flynn, 1987).

Perhaps IQ gains are due to "cultural bias." Here, we must distinguish cultural trends that make test content more familiar at one time than another from cultural trends that have truly raised the level of cognitive skills from one time to another. We measure the magnitude of IQ gains by the extent to which people do better on a test whose content is outmoded (e.g., the WISC) than they do on a test whose content is current (e.g., the WISC-R). Vocabulary or Information that was common when the test was constructed sometimes falls out of general use or general knowledge over time, and this is why the content of IQ tests is updated from time to time. Unfamiliarity with outmoded content should artificially deflate estimates of IQ gains by causing lower scores on the outmoded test. However, this pattern is the exact opposite of that caused by IQ gains, namely, higher scores on earlier, rather than later, tests. As for the IQ test items becoming public, this is least likely on tests rarely used in schools such as Raven's. Yet these are the tests that have shown the largest gains over the entire 20th century.

2.5. The WISC subtests to the rescue

Massive IQ gains are not an artifact and yet, if that is so, we are driven to conclusions that seem absurd. The solution to this paradox is to be found by focusing on the WISC subtest trends rather than Full-Scale IQ trends. As mentioned above, IQ gains vary considerably by subtest. Between 1947 (WISC) and 2002 (WISC-IV), the following trends occurred: Similarities showed a huge gain of 24 points (SD = 15), the five Performance subtests showed gains ranging from 12 to 21 points, Comprehension exhibited an 11-point gain, and the remaining Verbal subtests (Information, Arithmetic, and Vocabulary) showed very limited gains of 2–4 points (Flynn, 2006a, Table 1). Gains on the Raven's Progressive Matrices (RPM) test are also relevant. As Flynn (1998) demonstrated, advanced nations have made huge gains on the Raven's test. Although no good data are available for Americans, if we posit gains equal to the lowest gains found elsewhere, U.S. gains would at least match Similarities; indeed, at a 0.5 points per year, they would amount to 27.5 points gained over 55 years.

Let us analyze the cognitive skills needed to do well on the various IQ tests and subtests. The huge Raven's gains show that today's children are far

better at solving problems on the spot without a previously learned method for doing so. The WISC Performance subtests all measure this to some degree. These WISC subtests require arranging blocks so that the view from above duplicates a presented pattern, building an object out of its disassembled parts, or arranging pictures to tell a story. In contrast, most children have some prior experience at jigsaw puzzles or reading books in which pictures are the main vehicle of the story. We suspect that the fact that the on-the-spot element is diluted in Performance subtests explains why their gains, although substantial, lag behind Raven's gains. Children have been exposed to jigsaw puzzles and picture books for many generations, and this prior experience contributes to their success on Performance subtests.

We turn next to the subtests that show minimal gains. Having an adequate fund of general information, being able to do arithmetic, and having a decent vocabulary are very close to school-taught skills. These tests require much less on-the-spot reasoning or problem solving and are more a matter of exhibiting what you know: You either know that Rome is the capital of Italy or you know only of Rome, Georgia; you know what "delectable" means or you do not. Arithmetic, sometimes assumed to be just as rote as the more verbal tests, is more complex, as we shall see.

This contrast, the difference between on-the-spot reasoning (or Fluid Intelligence) and stored knowledge (or Crystallized Intelligence), is the key distinction made in the Horn–Cattell theory of Fluid and Crystallized Intelligence, to which Carroll has also made contributions (Carroll, 1993; Cattell, 1971; Horn, 1967, 1968, 1978). The Horn–Cattell theory is making inroads with standard intelligence tests, including the Stanford–Binet and Wechsler tests, as these tests have begun to provide subscores that correspond to constructs from the Horn–Cattell theory. In addition, the distinction parallels a distinction from cognitive psychology, the distinction between procedural and declarative knowledge. Procedural knowledge is knowledge "how," or knowledge about how to get things done, which shades directly into on-the-spot reasoning about how to solve a current problem based on often-implicit notions about how similar problems have been solved in the past. In contrast, declarative knowledge is knowledge "of," or knowledge of stored facts and concepts.

It is illuminating to take trends in WISC subtest scores in conjunction with trends on the National Association of Educational Progress (NAEP) tests, often called the Nation's Report Card. Between 1971 and 2002, that is, comparing the current generation of schoolchildren with its parents, young children made substantial reading gains. However, by the 12th grade, reading gains drop off to almost nothing (U.S. Department of Education, 2000, pp. 104, 110; 2003, p. 21). This is hardly surprising. Between 1972 and 2002, the WISC subtests show that schoolchildren made no gain in their stores of general information and only minimal vocabulary gains (Flynn, 2006a, Table 1). Therefore, although today's

children may learn to master preadult literature at a younger age, they are no better prepared for reading more demanding adult literature. You cannot enjoy *War and Peace* if you have to run to the dictionary or encyclopedia every other paragraph.

In other words, today's schoolchildren opened up an early lead on their grandparents by learning the mechanics of reading at an earlier age. But by age 17, their grandparents had caught up. Moreover, since current students are no better than their grandparents in terms of vocabulary and general information, the two generations at 17 are equal in their ability to read the adult literature expected of a senior in high school.

From 1973 to 2000, the Nation's Report Card shows 4th and 8th graders making mathematics gains equivalent to almost 7 IQ points. But once again, the gain falls off by the 12th grade, this time to literally nothing (U.S. Department of Education, 2000, pp. 54, 60–61; 2001, p. 24).

Once again, a WISC subtest suggests why. The Arithmetic subtest and the NAEP mathematics tests present a composite picture. An increasing percentage of young children have been mastering the computational skills that the Nation's Report Card emphasizes at those ages. However, during that very same period, children made no score gains on WISC Arithmetic. To do that subtest, you must know the mechanics of calculation plus something else. The questions are put verbally, which means the child cannot give a purely mechanical (times-table-type) answer. And some questions require you to diagnose what combination of operations (first division and then multiplication) is required to solve the problem. By the 12th grade, the lack of progress in terms of learning to think mathematically takes on significance. American schoolchildren cannot do Algebra or Geometry any better than their grandparents could. Although the older generation was slower to master computational skills, they were no worse off at graduation from high school.

In one area of cognitive skills, secondary students clearly have undergone a dramatic change. The huge gains on RPM show that today's youth are much better at problem solving in situations in which they have no previously learned or rote method of attacking the problem. It is likely that this advantage is sustained and perhaps enhanced by university study. There are a number of likely dividends. Every year America has an increased number of managerial, professional, and technical jobs to fill—jobs that often require decisions without the guidance of set rules.

We now know why recent IQ gains do not imply that our grandparents should seem to be much less intelligent than their grandchildren. Assume we hear a recent high school graduate chatting with his grandfather who also finished high school. The latter would be able to discuss novels as an equal and display an equally wide range of reading. He could discuss current affairs with as broad a vocabulary and fund of general information. The grandson would be much better in terms of on-the-spot problem solving, at least in certain contexts. Sometimes, the grandfather's "handicap" would affect social conversation, particularly because he would not think that abstract or "impractical" problems were very important. The grandfather might be more rule governed and would probably count that as a virtue.

2.6. Did almost everyone once have MR?

Our grandparents were assigned a median birth date of 1934 to get them in school in time for the WISC. But what of their parents and grandparents, what of the cohort born in 1906 that was in school in 1918 and the cohort born in 1877 that was in school in 1900? British Raven's data show massive gains beginning with the cohort born in 1877— they were actually tested at maturity of course (Raven, Raven, & Court, 1993, Graph G2). World War I military data show that U.S. gains were under way as far back as we can measure (Tuddenham, 1948). The Wechsler–Binet rate of gain (0.3 points per year) entails that the schoolchildren of 1900 would have had a mean IQ just under 70. The Raven–Similarities rate (0.5 points per year) yields a mean IQ of 50 (against current norms). Even if the latter accounts for most of the former, it will hardly do to say simply that our ancestors were bad at on-the-spot problem solving.

After all, innovative thinking is an important real-world skill. Only the worst child of the 2,200 schoolchildren used to norm the WISC-IV would have performed as low as the 1,900 average. To presume our ancestors were that lacking in innovation or problem-solving initiative would be to characterize them as virtual automatons. Moreover, there is some connection between mental acuity and the ability to learn. Jensen (1981, p. 65) related an interview with a young man with a Wechsler IQ of 75. Despite the fact that he attended baseball games frequently, he was vague about the rules, did not know how many players were on a team, could not name the teams his home team played, and could not name any of the most famous players.

3. Possible Solutions

3.1. Piagetian approach

Piaget made relevant and perhaps crucial distinctions, namely, among preoperational, concrete operational, and formal operational thinking, but others have done the fieldwork. If we assume that most people were still on the concrete level in 1900, they were handicapped most on the two IQ tests that show the largest and therefore the most embarrassing gains. We refer to the RPM test and the Similarities subtest.

A person on the concrete operational level lives in the world that confronts us in everyday life. When presented with a Similarities-type

item such as "what do dogs and rabbits have in common," Americans in 1900 would be likely to say, "You use dogs to hunt rabbits." The correct answer, that they are both mammals, assumes that the important thing about the world is to classify it in terms of the taxonic categories of science. Even if the subjects were aware of those categories, the correct answer would seem absurdly trivial. Who cares that they are both mammals? That is the least important thing about them on the concrete level. As long as you are on that level, it is not natural to detach abstractions and logic and the hypothetical from their concrete referents.

The key issue is not whether people use abstractions. People on the concrete level often use abstractions: The concept of hunting as distinct from fishing is an abstraction. They also use syllogistic logic: Basset hounds are good for hunting; therefore, if that is a Basset hound, that dog would be good at hunting. People operating at the concrete level would of course use the hypothetical; if I had two dogs rather than only one, I could catch more rabbits. Such persons do not have MR in any sense, but in terms of current norms they will appear to do so on Similarities. Today we are so familiar with the categories of science and are so imbued with the scientific worldview, that it seems obvious that the most important attribute that things have in common is that they are both members of a common category, such as both being animate, or mammals, or chemical compounds.

Beginning with its inception, what counts as a correct answer on Similarities favors the formal mode over the concrete and, by the time of the WISC-R, this was made explicit (italics added): "Pertinent *general* categorizations are given 2 points, while the naming of one or more common properties or functions of a member of a pair (a more *concrete* problem-solving approach) merits only 1 point" (Wechsler, 1974, p. 155). The preference for taxonic answers (categories that classify the world and extra credit for the vocabulary of science) is extraordinary and reaches an even higher level in the WISC-IV, where the "one point" for concrete answers is reduced to "merits no or only a partial credit" (The Psychological Corporation, 2003, p. 71). You are just not supposed to be preoccupied with how we use something or how much good it does you to possess it.

If children are on the concrete level, they can get no more than half credit on most Similarities items. In 1900, if children aged 14 were of average intelligence and were given a prehistoric version of the WISC-IV, they would have a raw score of about 11 and be 2.0 SDs below the current mean, which is a score of 70 against today's norms (The Psychological Corporation, 2003, p. 229). This was the "target" score that Full-Scale IQ gains implied when projected back to 1900. Note how the WISC manuals use the word "pertinent" to justify rewarding taxonic answers. This is just a synonym for claiming that classification is what is important about a pair of things. Imagine a rural child in 1900 being told that the most important

thing about dogs and rabbits is a name that applies to both, rather than what you use them for.

In sum, we need not infer that the huge gains on Similarities from one generation to another signal a general lack of intelligence on the part of our ancestors. Their minds were not permeated by the scientific worldview and they had not shifted from concrete to abstract, or formal operational, thinking.

RPM presents 60 patterns each of which has a piece missing. Six alternatives picture a candidate for the missing piece and the subject must select the one that fits the logic of the matrix design. The entire test demands detaching logic from a concrete referent, but even subjects unused to this can adapt to varying degrees under examination conditions. From a larger sample of 201 children, Styles (in press) selected 60 children who were typical in terms of age and initial testing. The 60 children selected were part of a five-year study of the intellectual development of children initially 10, 12, and 14 years of age (Andrich & Styles, 1994). The children took both a Piagetian test and items of the RPM ranked in order of difficulty. They were tested yearly on the former and twice yearly on the latter over a period of four years.

Five Raven's items were used to illustrate the sections of the test and therefore, were automatic correct answers. Two items were so easy for this group of children that everyone got them correct. The remaining 53 items mapped on to ascending Piagetian competence in ascending order of difficulty. Of these, the 20 most difficult RPM items required the subject to be either on the threshold of the formal level or operating on that level. Styles asserted that these items require using either a number of rules or a very complex rule to interpret the matrix pattern; and the subject must consider the logical relations between relations, rather than the factual relationship between a proposition and concrete reality.

In other words, if children aged 14 in 1900 were operating primarily on the concrete level, we would expect their raw scores to have a ceiling of about 40 correct items out of a total of 60. John Raven (2000, p. RS3 18) established norms for the United States circa 1982, and these norms show 40 items correct as the 38th percentile of 14-year olds. The age curve corresponding to a ceiling of 40 is that of 7.5-year olds. Their median is a score of 20, which is off the bottom of the curve for 14-year olds.

If most people in 1900 operated below the formal level of reasoning, this would serve to resolve the paradox of the huge Raven's gains between then and now. The gains can be as large as you wish without any presumption that most of our ancestors suffered from MR. They were quite capable of on-the-spot problem solving in the concrete situations that dominated their lives. The ingenuity of soldiers trying to say alive in the trenches of

World War I and the improvisations of mechanics trying to keep the first motorcars running is part of the historical record. Those who wish a fuller discussion of this issue should see (Flynn, 2007), What is intelligence?

The solution to our paradox does not imply that massive IQ gains over time are trivial. The scientific worldview, with its vocabulary, taxonomies, and detachment of logic and the hypothetical from concrete referents, has begun to permeate the minds of postindustrial people. This has paved the way for mass education on the university level and the emergence of an intellectual cadre without whom our present civilization would be inconceivable.

3.2. Psychometric approach based on item response theory

At least one other potential approach should be mentioned, a psychometric approach founded on modern test theory, which is also called item response theory (IRT). In contrast to IRT, the traditional classical test theory approach to developing tests involves sampling items from domains of content, evaluating the relations between items and total scores, and selecting items with optimal correlations with total scores. Much of this work can be done without examining the link between an individual's level of ability and the characteristics of the items to which the person responds. Items merely serve as the avenue to the estimation of a total score for each person, and the evaluation of an individual's overall score depends on where his or her score falls in the population distribution of total scores.

The introduction of IRT to the test development process changed all of this, and the close tie between the individual's level on a latent trait and the content of items at that level lies at the heart of the IRT approach. Item content plays a more central role in the IRT approach, because an individual's overall score on a test gains interpretability on the basis of the items that she or he is able to solve with a given probability. Technical details of the IRT approach go far beyond the present chapter, and readers are referred to several excellent publications (Embretson & Reise, 2000; McDonald, 1999). But, to reiterate, the IRT approach provides both person estimates (e.g., estimates of an individual's standing on the latent trait) and item difficulty estimates that are on the same scale, enhancing the interpretation of test scores.

The applicability of the IRT approach to the issue of diagnosing MR is direct. The traditional cutoff score for diagnosing MR is a score that falls at or below 2.0 SDs below the population mean. IRT methods, which have been the standard approach for scaling intelligence test items for two decades or more, can be used to identify items that supply a great deal of measurement information—and therefore discriminate well among individuals—near the cutoff score. The application of IRT methods to the normative

sample from an intelligence test could then provide information regarding what types of items discriminate best between those who fall above the cutoff score and those who fall below.

One key assumption of IRT is that a set of items is unidimensional, meaning that the set of items assesses a single, unitary underlying trait. Thus, application of an IRT model to data from all items on, say, the WAIS-III would be difficult to justify, as items on the various WAIS-III subtests are indicators of several different dimensions. However, as noted in an earlier section, most recent versions of tests including the Stanford Binet V, the WAIS-III, and the Woodcock–Johnson III provide IQ subscores on the standard IQ scale (i.e., mean of 100, SD of 15) for Fluid Intelligence and Crystallized Intelligence, which are conceived of as unitary dimensions of intelligence.

The conception of Fluid Intelligence as a general form of reasoning that supports on-the-spot reasoning in novel situations is similar to that of Piagetian forms of reasoning. But, rather than presuming that reasoning conforms to hierarchically ordered levels such as concrete operational and formal operational thought, levels of Fluid Intelligence are thought to vary continuously from simpler to more complex. Even if levels of Fluid Intelligence are classified more with regard to degree than to kind, it still may be possible to identify distinctive forms of fluid reasoning that can be performed by persons who exceed the criterion of MR and that cannot be performed by persons who fall below that criterion.

The second dimension is Crystallized Intelligence, which pertains to knowledge of the culture, of the meanings of cultural artifacts (including language), and of proper actions and behaviors. Lacking full and automatic knowledge of the culture could so impair one's judgment in an on-the-spot situation that low levels of Crystallized Intelligence could also be an important impediment to making fully informed judgments, limiting legal culpability for the outcomes of one's behaviors.

The dimensions of Fluid and Crystallized Intelligence are central in current theories of ability structure and in the scoring procedures for current tests. Therefore, the future may see a move away from a score of 70 or below on Full-Scale IQ to satisfy the subnormal general intellectual functioning prong of the diagnosis of MR, toward a score of 70 or below on either Fluid Intelligence, or Crystallized Intelligence, or Full-Scale IQ. This would parallel current rules for satisfying the prong related to adaptive behavior. Regardless of whether such a major change for the intelligence prong was made, the use of IRT—with its intimate tie between item difficulty and person ability—might be another reasonable way to make the crucial distinction between those who do or do not meet a diagnostic criterion for MR, thus providing an external validity criterion for this decision.

4. Concluding Remarks

4.1. Temptation to be resisted

If IQ gains do not persist into the future, it will be tempting to forget the lessons they have taught us. That temptation must be resisted. The only problem that would be "solved" would be that scores on IQ tests taken after the date of cessation would not have to be adjusted for obsolete norms. The deeper problem of whether some IQ criterion of MR can qualify for external validity would remain. This point must be stressed because there is no reason to believe that cognitive progress will go on forever.

After all, the persistence of IQ gains is dependent on the persistence of the trends that cause them. These trends likely include more people looking at the world through scientific spectacles; higher ratios of adults to children in families; and more leisure activities and jobs that are conceptually demanding. The future may see a reversal of these trends. The spread of the scientific ethos may be terminated by the powerful forces (particularly in America) that hate science. The trend toward a higher ratio of adults to children in the home may be reversed by more single-parent homes. If that occurs, children might get less parental attention. Our willingness to be challenged by more conceptually demanding leisure activities must eventually reach a limit. The multiplication of professional and managerial jobs already depends to some degree on featherbedding. Although IQ gains are still robust in America, they have stopped in Scandinavia (Flynn & Weiss, in press; Schneider, 2006). Perhaps Scandinavian societies are more advanced than ours is and their trends show what the future holds for us.

4.2. Necessary tasks

It is time to offer a summary that suggests the tasks we must perform. In 1900, most people could cope intellectually with concrete reality but a few could not. The latter, both then and now, should be rightly classified as having MR. Massive IQ gains over time signal primarily a shift from the concrete to the formal mode, but they have not altered the ratio between those who have concrete competence and those who do not. Our first task is to investigate whether all of this is true, that is: Has the ability to cope with everyday life failed to show significant gains during a period in which IQ gains have been robust? If that is verified, we must confront our second and most fundamental task: Can we develop and justify purely behavioral criteria to validate an IQ criterion of MR? If we succeed in that, that in turn confers a third task: We should alter IQ scores inflated by obsolescence. At least as long as IQ gains persist, failure to do so allows the percentage of those classified with MR to fluctuate radically over time.

IQ scores blur the difference between the inability to cope with everyday life (lack of concrete operational competence) and the inability to cope with school subjects (lack of formal operational competence). Therefore, we have a fourth task: We should supplement IQ tests with a Piagetian test or some other kind of reasoning test that distinguishes between reasoning on the preoperational, concrete operational, and formal operational levels.

As for our first task, the Vineland Adaptive Behavior Scales can now supply relevant data. For the first time, the publishers of the Vineland have compared the performances of samples used to standardize their test in two different years (Vineland, 2006). Subjects aged 7–18 who took both tests actually found the 1984 norms more difficult to meet than current norms. That is, they received an overall Adaptive Behavior Composite of only 95.0 on the old test and one of 98.4 on the new test (SD = 15). This seems to indicate that children actually lost ground in terms of adaptive behavior over the last 20 years.

However, the loss in adaptive behavior is more apparent than real because the old test has lost some of its relevance in assessing adaptive behavior. Scores on the Communication and Socialization subtests were similar on the two versions. The lost ground was almost entirely on the Daily Living Skills subtest. The 1984 version of that subtest contains obsolete skills that would deflate the scores of contemporary children (items such as "sews or hems clothes," "makes own bed," and "uses a pay telephone"). The most judicious conclusion is that American children have marked time in terms of adaptive behavior. During the same period (WISC-III to the WISC-IV), American children made IQ gains at the traditional rate of 0.3 points per year. Therefore, IQ gains over time do not mean that fewer and fewer children find it difficult to cope with everyday life.

Our second task is the most difficult. How can we identify an IQ criterion for MR that is validated by behavioral criteria? The American Association on Intellectual and Developmental Disabilities (formerly American Association on Mental Retardation) could select a panel of 20 psychologists whose judgment they trust and ask these experts to create a pool of 400 participants (20 nominated by each) classified as mentally retarded on purely behavioral criteria. Members of the panel would then administer Wechsler–Binet tests to the participants and assess the results to see if any common IQ ceiling emerged. Supplementary information could be provided by IRT estimates of the items that discriminated between persons falling above or below the common IQ cutoff score. Ideally, most children classified as mentally retarded would fall below a particular IQ and only the rare child score would score above it. At that moment in time, that particular score on that particular test could be recommended to school psychologists as a check on their clinical judgment.

Thanks to the possibility of future IQ gains over time, the whole experiment would have to be repeated after no more than seven years.

Let us hope that the consensus does not unravel over time. If the psychologists on the panel diverged, some finding that the ceiling had stayed at 70, others finding that it had risen to 72 or even to 74, we would have reason to be disconcerted. Coherence at both the start and finish of the seven years would inspire confidence.

Our third task would be to find a formula to adjust IQs during the interim on the assumption that IQ gains were still in progress. The seven-year review, or a new standardization of an IQ test, might show that we were mistaken. But better to err on the side of caution than to sentence convicted offenders to death. As the reader knows, we have offered a formula to be used for the time being. Some might complain about the lack of precision of our formula, but this formula is clearly better than having no adjustment of any kind. Proceeding without an adjustment formula during a time of enhanced normative performance ensures that persons, perhaps many persons, who deserve the label of MR will fail to receive it. In school contexts, the label of MR may be regarded as detrimental, given the discrimination associated with it. But, in criminal proceedings, the label of MR can save a life.

This brings us to our fourth and final task. Children who score as MR on the WISC should take a Piagetian test or some other form of reasoning test, such as a test of Fluid Intelligence or Crystallized Intelligence subjected to IRT analysis, to determine whether or not they are cognitively competent on the level of concrete reality. If they are competent, the label of MR should be set aside in favor of something else, perhaps school learning problems (SLPs). In other words, the diagnosis of MR should be reserved for individuals who are unable to reason adequately on the level of concrete reality in everyday situations, regardless of whether they have problems with typical school content.

To give an example of a test that might be suitable, Trevor Bond has developed Bond's Logical Operations Test (the BLOT) to distinguish whether one uses logic on a concrete or formal level (Endler and Bond, in press, p. 8). Certain items on this test explore "whether the child has the reasoning to manipulate conclusion(s) by reversing the operations of thought (i.e., reciprocity)." We would add that such items show the hypothetical slowly being freed from the ties that bind it to concrete situations. Using the BLOT to test a sample of low-IQ subjects might discriminate between WISC items that can be handled by someone competent on the concrete level and those items that cannot. If so, these results would provide a key that would make individual administration of the Piagetian test unnecessary. Eliminating the WISC items that require formal competence and seeing how much their elimination raised the IQ score of an individual would allow us to reassess performance that was putatively at the level of MR.

4.3. Remaining problems

The use of a test like the BLOT to distinguish whether a test taker can reason at the concrete level is, of course, not without its own problems. The items Bond uses are not linguistically simple, which is to say they have a clear verbal loading. If the core aim of the assessment were to measure on-the-spot reasoning that deals with novel everyday, concrete situations, the linguistic complexity of items might contaminate results. To guard against this problem, the wording of problems would have to be reduced to a rather simple level, while retaining a focus on the level of reasoning assessed.

If the nature of the reasoning assessed were unchanged by the simplification of linguistic complexity, the items from the BLOT have clear face validity in that they require reasoning in concrete situations, something that cannot be said about the RPM. Assuming simplicity, a verbal test of on-the-spot reasoning in concrete situations is advantageous. The RPM uses abstract visual stimulus patterns. None would dispute the high levels of complex reasoning required to solve many items. However, the lack of RPM items with everyday situational content would lead many to question whether ability or inability to deal or reason with highly complex spatial stimuli would have any close parallels with one's ability to reason at a concrete level in everyday situations. We have every reason to believe that scores on the RPM would correlate very highly with scores on the BLOT, but the aims of the BLOT to distinguish whether persons can reason at the concrete level do seem to have added utility for the purpose of assessing this form of reasoning.

Other objections can and should be raised regarding the use of a Piagetian test of reasoning at concrete levels. If such a test were required in addition to the current assessment requirements (an individually administered test of general intelligence and an assessment of adaptive behavior), this would be likely to be seen as a fundamental change in the definition of MR. At present, to merit the diagnosis of MR, an individual must exhibit significantly subnormal levels of general intellectual functioning and concomitant deficits in adaptive functioning. If yet another hurdle were placed in the way, namely, that the person also exhibit an inability to reason in concrete situations, research would have to be undertaken to determine whether an appropriate number of persons satisfy all the criteria of the new definition. The behavioral criterion is, after all, the most fundamental. It would be disturbing if some people tested as competent to deal with everyday life, and yet had a life history that indicated the opposite.

In addition, many might question either the fundamental nature or the psychometric properties of a Piagetian reasoning test. Standard concerns cover issues such as various forms of reliability (e.g., inter-rater, test–retest), the standard error of measurement, the degree to which a test taker can fake bad or malinger without detection, the degree to which coaching can affect

scores, and so on. Current intelligence tests have passed these hurdles reasonably well, and any Piagetian test would also have to demonstrate adequate performance on all of these dimensions.

Piaget, however, was little interested in individual differences, concentrating on the delineation of stage-like developmental advances across chronological age. Many followers of Piaget have continued with these emphases. Indeed, most studies in the Piagetian tradition use only a small number of reasoning items, and participants are classified into groups (e.g., nonoperational, concrete operational) on the basis of the consistency with which they succeed on a small number of items. But, any life-or-death decision such as that confronted by defendants in capital cases cannot and should not be based on reasoning performance across a small number of items. Instead, any Piaget-based test would have to consist of a very large number of items, so that the level of reasoning by the test taker can be identified within an acceptably narrow confidence interval. Clearly, if research using converging operations—here, the use of an IRT-based index of Fluid Intelligence and/or Crystallized Intelligence to validate the decision achieved using a cutoff score from a Piaget-based test—found strong agreement between the two forms of assessment, both the Piaget-based and IRT-based cutoff scores would be mutually validated. Still, tests of either of these forms undoubtedly would take several years to develop and validate, so acceptable tests are, most probably, not near at hand.

4.4. "Bring the tires to me"

The primary basis for the U.S. Supreme Court decision that persons with MR should not be subjected to capital punishment was the impaired judgment or reasoning that such persons exhibit. If a person with MR cannot reason clearly and fully about his or her actions, then that person is less culpable for his or her actions. Current intelligence tests in the Wechsler and Stanford–Binet tradition assess many things. Because they attempt to arrive at an estimate of general intelligence, the tests assess a nonsystematic conglomeration of Crystallized Intelligence, Fluid Intelligence, spatial ability, perceptual speed, and memory, among other mental functions. Many of these functions have little relation to the basis for the Supreme Court decision.

For example, the Court did not base its decision on how quickly a defendant can make simple perceptual judgments (e.g., Digit Symbol Substitution). Intelligence test scores have a vast array of external validities, including correlations with many different measures of school success, job success, and so forth. But, very little or no external validity has been amassed with regard to the relation between particular IQ scores (e.g., scores of 70 and below) and common forms of on-the-spot reasoning in concrete situations. Once again, the Court was interested in culpability: Is this person so suggestible that someone could easily persuade him to participate in a crime?

Current intelligence test scores, particularly a Full-Scale IQ score that reflects a complex composite of multiple functions (many unrelated to judgment or reasoning), simply fail the crucial external validity criterion of establishing benchmarks for relations between levels of IQ test performance and mature judgment in concrete situations. Without some acceptable evidence on this front, the field of psychology must push for acceptable measures of concrete reasoning that provide a clear answer that would tell the courts whether or not society should hold individuals fully responsible for their actions.

We stress this point not so much because of what we have learned from "hard" data, such as Vineland Adaptive Behavior Scales scores or Wechsler IQs, but from reading the case histories of capital offenders who are being judged as mentally competent. For example, John Doe's case history shows that he never passed the test for a driver's license or held a job that required reasonable literacy or numeracy. Family and friends testified that he tended to lose focus if sent on errands. One boyhood companion testified that when the defendant and he were both 16 years of age, he pointed out a car and said: "That is Mrs. Smith's car. She and I are friends and she said I could borrow her tires. Would you go over and take them off and bring the tires to me." John Doe obeyed. Such a level of misunderstanding of everyday situations could easily lead a person to perform actions—mistakenly, and without cogent deliberation—that might have disastrous impacts and lead to capital charges.

Whatever his IQ, John Doe was not mentally mature enough to be held responsible for his actions. We cannot reveal identities, so the reader must remain ignorant of whether or not the public prosecutor secured his execution. We can assure you that his zeal was great. Whatever the outcome in this case, there are others in which IQ scores have played the role of executioner. The fate of these defendants is an American tragedy.

4.5. Quid faciendum est?

Perhaps a panel of philosopher kings will appeal to few except those who earn their living as philosophers. The proposal of a basic reevaluation of intelligence tests will have to overcome the enormous potency of inertia. The psychological fraternity can judge for itself the case we have made. Whatever our profession all of us are moral agents. The testers are not going to give away the application of some kind of test criterion of MR. The courts are not going to stop searching for an isle of objectivity in a sea of contradictory professional opinions. Those whose lives and welfare depend on our judgment need a "score" that serves as at least one criterion of MR. They deserve one with a strong, direct, and defensible claim to external validity.

ACKNOWLEDGMENTS

We thank Cambridge University Press for permission to use and adapt material from Flynn (2007). We thank the American Psychological Association for permission to use and adapt material from Flynn (2000, 2006b). This research was supported in part by grants from the National Institute of Child Health and Human Development, the National Institute on Drug Abuse, and the National Institute of Mental Health (HD047573, HD051746, and MH051361, respectively).

REFERENCES

Andrich, D., & Styles, I. (1994). Psychometric evidence of intellectual growth spurts in early adolescence. *Journal of Early Adolescence, 14,* 328–344.

Atkins v Virginia. (2002) 534 U.S. 304.

Carroll, J. B. (1993). *Human cognitive abilities.* Cambridge: Cambridge University Press.

Cattell, R. B. (1971). *Abilities: Their structure, growth, and action.* Boston: Houghton Mifflin.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum.

Endler, L. C., & Bond, T. G. (in press). Tracking development with the Rasch Model: Empirical evidence of growth and heterogeneity. In X. Liu, & W. Boone (Eds.), *Applications of Rasch measurement in science education.* Maple Grove, MN: JAM Press.

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. (1998). IQ gains over time: Toward finding the causes. In U. Neisser (Ed.), *The rising curve: Long-term gains in IQ and related measures* (pp. 25–66). Washington, DC: American Psychological Association.

Flynn, J. R. (2000). The hidden history of IQ and special education: Can the problems be solved? *Psychology, Public Policy, and Law, 6,* 191–198.

Flynn, J. R. (2006a). Efeito Flynn: Repensando a inteligência e seus efeitos [The Flynn Effect: Rethinking intelligence and what affects it]. In C. Flores-Mendoza, & R. Colom (Eds.), *Introdução à Psicologia das Diferentas Individuais* [Introduction to the psychology of individual differences] (pp. 387–411). Porto Alegre, Brazil: ArtMed. (English trans.:jim.flynn@stonebow.otago.ac.nz

Flynn, J. R. (2006b). 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: Cambridge University Press.

Flynn, J. R., & Rossi-Case, L. (under review). Beyond skulls and genes: La Plata, Raven's, and gender equity; also new massive IQ gains.

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

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

Hamm, H., Wheeler, J., McCallum, S., Herrin, M., Hunter, D., & Catoe, C. (1976). A comparison between the WISC and WISC-R among educably mentally retarded students. *Psychology in the Schools, 13,* 4–8.

Horn, J. L. (1967). Intelligence: Why it grows, why it declines. *Transaction, 4,* 23–31.

Horn, J. L. (1968). Organization of abilities and the development of intelligence. *Psychological Review, 75,* 242–259.

Horn, J. L. (1978). Human ability systems. In P. B. Bates (Ed.), *Life-span development and behavior* (Vol. 1, pp. 211–256). New York: Academic Press.

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

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

McDonald, R. P. (1999). *Test theory.* Mahwah, NJ: Erlbaum.

Raven, J. (2000). *Raven manual research supplement 3: American norms; neuropsychological applications.* Oxford: Oxford Psychologists Press.

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

Schneider, D. (2006). Smart as we can get? *American Scientist, 94,* 311–312.

Scullin, M. H. (2006). Large state-level fluctuations in mental retardation classifications related to introduction of renormed intelligence test. *American Journal on Mental Retardation, 111,* 322–335.

Styles, I. (in press). Linking psychometric and cognitive-developmental frameworks for thinking about intellectual functioning. In J. Raven (Ed.), *Contributions to psychological and psychometric theory arising from studies with Raven's Progressive Matrices and Vocabulary Scales.*

The Psychological Corporation. (2003). *The WISC-IV technical manual.* San Antonio, TX: The Psychological Corporation.

Tuddenham, R. D. (1948). Soldier intelligence in World Wars I and II. *American Psychologist, 3,* 54–56.

U.S. Department of Education. Office of Educational Research and Improvement. National Center for Educational Statistics. (2000). *NAEP 1996 Trends in Academic Progress,* NCES 97–985r, by J. R. Campbell, K.E. Voelkl, and P.L. Donahue. Washington, DC.

U.S. Department of Education. Office of Educational Research and Improvement. National Center for Educational Statistics. (2001). *The Nation's Report Card: Mathematics 2000.* NCES 2001–517, by J.S. Braswell, A.D. Lutkus, W.S. Grigg, S.L. Santapau, B. Tay-Lim, and M. Johnson. Washington, DC.

U.S. Department of Education. Institute of Education Sciences. National Center for Educational Statistics (2003). *The Nation's Report Card: Reading 2002.* NCES 2003–521, by W.S. Grigg, M.C. Daane, Y. Jin, and J. R. Campbell. Washington, DC.

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.

Walker v. True. No. 04–016 (4th Cir. February 17, 2005).

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

Wechsler, D. (1974). *Wechsler Intelligence Scale for Children—Revised.* New York: The Psychological Corporation.

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

VOLUME THIRTY FIVE

INTERNATIONAL REVIEW OF RESEARCH IN MENTAL RETARDATION

Edited by

LARAINE MASTERS GLIDDEN
Department of Psychology
St. Mary's College of Maryland
St. Mary's City, Maryland

ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO
Academic Press is an imprint of Elsevier