ETHICS & BEHAVIOR, *13*(1), 11–17

The Difficulty of Basing Death Penalty Eligibility on IQ Cutoff Scores for Mental Retardation

Stephen J. Ceci
Department of Human Development
Cornell University

Matthew Scullin
Department of Psychology
West Virginia University

Tomoe Kanaya
Department of Human Development
Cornell University

Suppose you are told that there has been an enormous reduction in the rate of Americans classified as mentally retarded during the past 30 years—from 2.27% of the school-age population in 1974 to only 0.94% of those in 1992 (Flynn, 1998a). In addition, suppose you are told that average intelligence is growing by leaps and bounds—the average Briton gained 55 IQ points between the cohort aged 20 in 1892 and the cohort aged 20 in 1992. Good news, you would undoubtedly respond. According to recent psychometric data, something akin to this situation appears to have occurred (Raven, Raven, & Court, 1993).

For example, the Ravens Progressive Matrices is a popular test of so-called fluid intelligence that was developed and standardized in 1942. At that time, many adults were given the test, including ones between the ages of 20 and 70 years old. In 1992 the Ravens was restandardized on a sample of Scottish adults who were representative of adults in the United Kingdom. (These norms show that performance peaked by ages 35–40—similarly to the 1942 norms, which peaked somewhat

Requests for reprints should be sent to Stephen J. Ceci, Department of Human Development, Cornell University, Ithaca, NY 14853. E-mail: sjc@cornell.edu

sooner but maintained that peak performance until ages 35–40.) Overall, adults in the 1992 sample scored 27 points higher than adults in the 1942 sample. However, to get an idea of the massive gains in IQ over time, adults who were 70 years old in the 1992 sample were similar in their scores to adults who were 20 years old in the 1942 sample. Because of this, Flynn (1998b) made the reasonable assumption that therefore 70-year-olds in the 1942 sample probably resembled 20-year-olds who were in the 1892 cohort, thus enabling him to trace IQ gains on the Ravens for 100 years. He found a gain of 55 points.

THE FLYNN EFFECT

This pronounced improvement in intelligence test performance over time causes IQ test norms to become obsolete within a generation or so. Intelligence test manufacturers, of course, are aware of this upward creep in IQ performance and have responded by renorming their tests periodically. Whenever they renorm their tests, they assign higher raw scores to obtain an IQ score of 100 than was necessary for the prior cohort to achieve this same score. Thus, over the course of the past century, a higher and higher proportion of questions had to be answered correctly to obtain the identical IQ score of the prior cohort. Research has demonstrated that the use of IQ norms based on a prior cohort of test-takers progressively inflates the IQ scores of subsequent cohorts of test-takers. This phenomenon is a result of the general upward IQ score trend known as the Flynn effect, after the political scientist James Flynn, who is credited with documenting the rise.

The Flynn effect is based on data collected in 20 countries to date: Flynn found that IQ increases have been on the order of (a) approximately 20 IQ points per every 30-year generation (beginning with the cohort tested in the early 1930s) and continuing until today, (b) 10 to 20 points per generation for Wechsler Performance IQs (c) 9-point increments for Wechsler Verbal IQs, (d) approximately 20-point increments per generation for so-called fluid IQ tests such as the Ravens Matrices and various types of number series tests, and (e) 10-point increments on so-called culture-fair IQ tests—though the magnitude of this latter increment may be specific to Scandinavia, as it has not been replicated elsewhere.

As a result of the Flynn effect, the same cutoff score that captured the bottom 2.27% (the mentally retarded group) when an IQ test's norms were published in 1974 would capture only the bottom 0.94% when these norms were about to be replaced by newer norms 18 years later. So, if the 1974 IQ norms were used to score individuals in 1992, then fewer than half as many would be classified as retarded. If IQ norms were held constant, IQ scores would have risen for the past 70 years at an almost constant rate, and the results would be truly astounding: "It would be difficult to defend any estimate that the mean IQ of Britons in 1892 was above 60. Therefore, at a minimum, 84% had an IQ below 75" (Flynn, 1998a, p. 20).

Flynn (1985) further estimated that, in America, psychologists who use the Wechsler Intelligence Tests (one of the most widely used tests in this country) may be leniently misdiagnosing mentally retarded individuals by anywhere between .27 to an entire standard deviation due to this rise.

Although you might be delighted to hear the news of rising IQ scores and a concomitant reduction in prevalence of mental retardation between 1974 and 1992, you might be less pleased to discover that suddenly in 1993 the rate of classifying children as mentally retarded stopped dropping and started climbing again. The beginning of this increase in the percentage of children being classified as mentally retarded corresponds to the introduction of a renormed, harder version of the major IQ test used for evaluating eligibility for special education programs. Although the 1993 rates of diagnosing someone as mentally retarded did not suddenly return to the 1974 level, there was a clear cessation of decline and the beginnings of an upward trend. This form of IQ yo-yoing is the result of changing IQ test norms, and it occurs periodically, involving hundreds of thousands of Americans. As we argue here, it has profound implications for the entire society, particularly in the way the criminal justice system handles and classifies mentally retarded capital murder defendants.

THE RAMIFICATIONS OF CHANGING IQ NORMS FOR DEATH PENALTY ELIGIBILITY

In 2002, the Supreme Court held in *Atkins v. Virginia* that execution of the mentally retarded constituted "cruel and unusual punishment" under the Eighth Amendment. The Court's reasoning was described in Part IV of the majority opinion, which points to the aspects of mental retardation that render such persons inappropriate candidates for execution because of their poor reasoning ability (available at www.supremecourtus.gov/opinions/01pdf/00-8452.pdf).

The Court's decision left it up to individual states to determine the criteria for mental retardation. Cited in footnotes to the Atkins decision were definitions of retardation from the American Association of Mental Retardation (1992) as well as from the American Psychiatric Association's *Diagnostic and Statistical Manual of Mental Disorders* (American Psychiatric Association, 1994). Both sources stipulate that the criteria for mental retardation (MR) consist of (a) below-average IQ (typically 70 and below), (b) deficits in adaptive functioning (e.g., basic skills for work), and (c) onset before the age of 18 (i.e., someone is not regarded as retarded if his or her low IQ score or poor adaptive functioning occurred for the first time after the age 17, as, for instance, the result of a car accident). Many states incorporate these same three criteria and often use the same language. For example, Nebraska stipulates that "Mental retardation means significantly subaverage general intellectual functioning existing concurrently with deficits in adaptive

behavior. An intelligence quotient of seventy or below on a reliably administered intelligence quotient test shall be presumptive evidence of mental retardation."

In its decision in *Atkins*, however, the Supreme Court provided no guidance for states to implement a ban on the execution of the mentally retarded. For example, what if someone's IQ score was above 70, but a year later it was below 70? Or what if someone's IQ is below 70 on one test while contemporaneously above 70 on another IQ test? Or what if someone's IQ was above 70 on a test that was administered in childhood but would have scored well below 70 on a newer version of the same IQ test? Such questions may seem surprising to those outside the field of psychometrics, but to those who toil in this field, answers to such questions are by no means straightforward. For example, because of the Flynn effect it is fairly common for someone's IQ score to fluctuate above and below the cutoff of 70 that most states employ.

The major IQ tests are renormed, on average, every 15 to 20 years. During these renormings, test manufacturers typically make some changes to the questions themselves (adding a few new ones and deleting a few old ones) and then collect new data from age-appropriate groups to make sure that the entire test behaves as desired. Flynn showed that each time an IQ test is renormed, it becomes substantially harder to obtain the same score on the revised test as the score that one obtained on its predecessor version.

Recently, the three of us tested the prediction that individuals given IQ tests will have their scores affected by the specific year they are tested. If they are tested at the start of a new IQ norming cycle, then, based on the Flynn effect, we predicted their IQs would be lower than if they were tested in the middle or near the end of that norming cycle. This is because of the trend for everyone to answer more and more IQ questions correctly as a norming period progresses. For example, as the years go by, individuals become better educated, more exposed to informative cultural events (e.g., museums and educational television), more test savvy, and so on, all of which help raise IQ performance. Thus, we expected that the same child's IQ would fluctuate in significant ways, depending on the year he or she was evaluated and, specifically, where this year was located in the current norming cycle. Further, we predicted massive shifts in the numbers of individuals diagnosed by school districts as mentally retarded and learning disabled as a result of these temporal developments. Specifically, we expected that each time an IQ test has a new set of norms, there would be a generalized lowering of IQ scores because the new norms would recalibrate the average IQ to remove the increases that accumulated over the previous norming cycle. Thus, we expected an increase in MR diagnoses each time new IQ norms were published. The other side of this expectancy is that we anticipated a decrease in diagnoses of learning disabilities because this latter group's IQ would be lower and therefore the gap between their IQ and their achievement would not be as great as when their IQs were higher at the end of a norming cycle. This would reduce their chances of being diagnosed as learning disabled but make them more eligible for a retardation diagnosis.

SOME SOBERING EMPIRICISM

Using our recently collected data from school district archives across the nation, we were able to examine the stability of the major IQ tests, including the most widely used Wechsler Intelligence Scales for Children (WISC) series. We examined the Full Scale IQ scores of individuals who initially received an IQ score in the range between 71 and 85 because this is the range that lies just outside the commonly cited 70 cutoff for an MR diagnosis. The Wechsler IQ tests all have standard deviations of 15 points. Two *SD*s below the mean is typically needed for an MR classification (an IQ score of 70), and this score should capture 2.27% of the entire population, assuming that the Wechsler demographers utilized a perfectly representative norming sample, which is probably never attainable, although they do a very good job. So, if 2 *SD*s below the mean isolates the bottom 2.27% of children the year the IQ test is normed, then each subsequent year will find fewer and fewer children scoring below 70 due to the strong tendency for scores to rise, that is, until the new norms come into use.

For example, in 1948 an IQ of 70 on the WISC isolated exactly the bottom 2.27% of children, making them eligible for an MR classification. But each subsequent year the same IQ score on the WISC isolated fewer and fewer children, so that by 1972 a score of 70 isolated only 0.54% of the test-taking population (Flynn, 2000). Then, when the new (collected in 1972) norms came into play in 1974, a score of 70 overnight captured 2.27% of children, once again making them eligible for an MR classification. To continue this saga, by 1989 the same IQ performance that was associated with an IQ score of 70 in 1974 was capturing only the bottom 0.47% of children. Thus, we see that virtually overnight the rate of classifying MR could have changed dramatically, although not by 400% because other factors, not just IQ, are involved in classification.

Given that the individuals in our study were just above the threshold of being eligible for the MR label, we wondered what would happen to their IQs when they were retested on either the same set of norms (e.g., a WISC–R followed by a WISC–R) or on a renormed version of the same test (e.g., a WISC–R followed by a WISC–III).

To address this question, we reclassified IQ data from approximately 10,000 school psychologist special education assessments that had been collected from three geographically and socioeconomically diverse locations spanning the WISC to WISC–R transition (early 1970s) and from eight locations spanning the WISC–R to WISC–III transition (early 1990s). Longitudinal data, including multiple testings (typically 3-year reevaluations) of students, were gathered whenever possible. Here we present preliminary conclusions based on 1,011 cases in which students scored between an IQ of 71 and 85 on a WISC series test and were

retested on either the same test or a renormed version of it less than 48 months later.

As hypothesized, we discovered that there was a significantly greater decrement in IQ test scores when children were retested on a revised IQ test than when they were retested on an IQ test with the same norms. The median decrement between the WISC–R and WISC–III retestings was 5 IQ points, and the median decrement between the WISC and WISC–R retestings was 4.5 points. In contrast to these large changes when the student was retested on a different normed version of the IQ test, the median student gained half an IQ point when tested and retested on the same WISC–R and lost 1 point when tested and retested on the same WISC–III.

WHAT DO THESE CHANGES IMPLY FOR CLASSIFICATION OF MR?

The percentage of students whose IQ dropped below 70 was significantly greater when they were retested on a renormed test than when they were retested on the same test. The differences are often quite dramatic. For example, nearly 38% of all students who scored on the cusp of being eligible for MR diagnosis (IQ just above 70) actually dropped into the retardation category when they were retested with the newer norms! This is truly a large effect. So, the question for courts is "Do you consider such a person retarded or not?" One can imagine this state of affairs may occur with even greater frequency in death penalty cases because of the frequency of retesting of defendants' IQ.

The results of our research suggest that the probability of a person who initially scored an IQ between 71 and 85, subsequently obtaining a score of 70 or below, varies substantially depending on the version of the IQ test administered. This effect cuts both ways: Borderline mentally retarded capital murder defendants who took older versions of the WISC series as children may be penalized relative to those who took the more recently normed WISC–III because the latter's scores are far more likely to fall under the cutoff score of 70. Conversely, defendants who took the most recent version of an IQ test will be advantaged relative to those whose scores were based on older norms. This means that it is insufficient for courts to simply ask for an IQ score and assessment of adaptive behavior.

Ironically, these documented changes in IQ scores that are associated with changing norms have resulted in dramatic fluctuations in the numbers of persons deemed eligible for MR diagnosis; however, this change has occurred in the absence of any meaningful change in these individuals' intellectual ability. This would seem to have significant implications for death penalty considerations.

ACKNOWLEDGMENTS

We gratefully acknowledge the support of the Smith Richardson Foundation, without whose help the research reported here could not have been accomplished.

REFERENCES

American Association of Mental Retardation. (1992). *Mental retardation: Definition, classification, and systems of supports* (9th ed.). Annapolis, MD: Author.

Atkins v. Virginia, 260 Va. 375, 534S. E. 2d 312.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Ceci, S. J. (1991). How much does schooling influence general intelligence and its cognitive components?: A reassessment of the evidence. *Developmental Psychology, 27,* 703–722.

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. (1998a). The schools: IQ tests, labels, and the word "intelligence." In J. S. Carlson, J. Kingma, & W. Tomic (Eds.), *Advances in cognition and educational practice: Vol. 5. Conceptual issues in research on intelligence* (pp. 13–42) London: JAI.

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.

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.

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