The Flynn Effect and U.S. Policies

The Impact of Rising IQ Scores on American Society Via Mental Retardation Diagnoses

Tomoe Kanaya *Cornell University*
Matthew H. Scullin *West Virginia University*
Stephen J. Ceci *Cornell University*

Over the last century, IQ scores have been steadily rising, a phenomenon dubbed the Flynn effect. *Because of the Flynn effect, IQ tests are periodically renormed, making them harder. Given that eligibility for mental retardation (MR) services relies heavily on IQ scores, renormed tests could have a significant impact on MR placements. In longitudinal IQ records from 9 sites around the country, students in the borderline and mild MR range lost an average of 5.6 points when retested on a renormed test and were more likely to be classified MR compared with peers retested on the same test. The magnitude of the effect is large and affects national policies on education, social security, the death penalty, and the military.*

Ever since the introduction of standardized IQ tests in the early 20th century, there has been a systematic and pervasive rise in IQ scores all over the world, including the United States. Known as the *Flynn effect* after James Flynn, the political scientist who has extensively documented this rise, the Flynn effect causes IQ test norms to become obsolete over time (Flynn, 1984, 1987, 1998). In other words, as time passes and IQ test norms get older, people perform better and better on the test, raising the mean IQ by several points within a matter of years. Once a test is renormed, which typically happens every 15–20 years, the mean is reset to 100, making the test harder and "hiding" the previous gains in IQ scores.

Although the Flynn effect is well documented within the average range of the IQ distribution, less is known about its impact on those who score well below the mean, such as those in the mental retardation (MR) and borderline intellectual functioning (borderline) IQ range. Depending on the size of the Flynn effect within this lower IQ range, meeting the IQ criteria for a diagnosis of MR might depend heavily on the IQ test norms being used the year an individual is tested (Flynn, 2000). If true, this could have enormous social implications beyond the incidence of MR because of its impact on various U.S. national policies, including special education financing and eligibility for social security benefits, the death penalty, and military service. In this article, we focus on MR diagnoses among school-age children as a window into related policies that are based on IQ scores. Using archived special education testing records from nine sites around the United States, we explored the impact of the Flynn effect on the Wechsler Intelligence Scale for Children—Revised (WISC–R; Wechsler, 1974) and Wechsler Intelligence Scale for Children—Third Revision (WISC–III; Wechsler, 1991) Full Scale IQ scores of children in the MR and borderline range, as well as its impact on psychologists' special education placement recommendations.

How Is Mental Retardation Defined?

According to the American Association of Mental Retardation (AAMR; 2002), mental retardation is characterized by significantly subaverage intellectual functioning, existing concurrently with limitations in conceptual, social, and practical adaptive skills (e.g., communication, social functioning, activities of daily living). Additionally, the onset of MR must occur by the age of 18 (hence ruling out instances of brain damage that resulted during adulthood to previously nonretarded individuals). The most commonly used measure of intellectual functioning to determine subaverage functioning is an IQ score of 70, two standard deviations below the mean of 100 used by the standard IQ tests (AAMR, 2002). Allowing for measurement error, AAMR raised the ceiling for its recommended IQ criteria to 75 in 1992.

The MR criteria from the American Psychiatric Association's (1994) *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.; *DSM–IV*) are similar to AAMR's, with a recommended ceiling of an IQ score of 70 for mild MR. The Social Security Administration (SSA; 2002) also requires subaverage intellectual functioning along with deficits in adaptive functioning, which must be demonstrated before the age of 22, to qualify for disability benefits for intellectual impairment. If an individual's Full Scale IQ score is below 60, however, that individual auto-

Tomoe Kanaya and Stephen J. Ceci, Department of Human Development, Cornell University; Matthew H. Scullin, Department of Psychology, West Virginia University.

This research was supported by a grant from the Smith Richardson Foundation to Stephen J. Ceci.

Correspondence concerning this article should be addressed to Tomoe Kanaya, Department of Human Development, Cornell University, Ithaca, NY 14853. Email: tk74@cornell.edu

Vol. 58, No. 10, 778–790 DOI: 10.1037/0003-066X.58.10.778

Tomoe Kanaya

matically qualifies for disability, regardless of adaptive functioning.

Thus, the "official" definitions vary slightly, although all of them stipulate an IQ of 70 or below as being an important criterion for classification as MR. However, whether a child who meets the IQ score criteria for MR is actually labeled MR can vary substantially between school districts and agencies and even between psychologists within an agency or a school district (Reschly, 1981; Reschly & Ward, 1991). Part of this variability is due to the ambiguity in the definition of *limited adaptive skills*. For example, some psychologists may use subjective judgments to determine limited adaptive behaviors and social skills, whereas others may use a specific cutoff score from a standardized measure, such as the Vineland Adaptive Behavior Scales (Sparrow, Balla, & Cicchetti, 1984). Moreover, adaptive behavioral expectations vary by age, so that it is possible to be subaverage during one epoch of life but not during another.

In addition to variance related to differing definitions and changes in adaptive behavior, psychologists vary in the amount of emphasis they actually place on adaptive functioning measures for the diagnosis of MR. The degree to which adaptive functioning assessments may affect MR placement decisions is unclear due to the fact that IQ scores and scores from adaptive behavior measures (such as the Vineland) are positively correlated in the .4 to .6 range (Bruininks, Woodcock, Weatherman, & Hill, 2000; Kamphaus, 1987). Therefore, although most psychologists agree that children need to be classified MR on the basis of deficits in multiple domains (intellectual and adaptive), in reality an MR diagnosis depends heavily on an IQ score.

Fluctuating IQ Scores Over Time: The Flynn Effect

Using data from 20 nations, Flynn discovered there have been IQ gains ranging from 5 to 25 points in a single generation (Flynn, 1984, 1987, 1998). Flynn's (1984, 1987) analyses have shown that the Flynn effect is stronger on tests of fluid intelligence (intelligence needed for on-the-spot reasoning, abstraction, and problem solving) rather than on tests of crystallized intelligence (intelligence centered around accumulated knowledge such as vocabulary, arithmetic, and general information).

The most dramatic findings of the Flynn effect—a gain of 21 IQ points in 30 years—has been found using the Ravens Progressive Matrices, a test of fluid intelligence (Flynn, 1987). Because the most commonly used IQ tests in the United States (e.g., the Wechsler and Stanford–Binet series) measure both crystallized and fluid intelligence, IQ gains on these tests in the United States are not as strong as the rise seen on the Ravens Progressive Matrices, but they are still dramatic—approximately 0.311 points a year for a total gain of 14.31 points within just 45 years (Flynn, 1984, 1987). More specifically, when comparing the WISC with the WISC–R, Flynn discovered a 10- to 20-point increase in the Wechsler Performance IQs (more heavily loaded on fluid abilities) and a 9-point increase in the Wechsler Verbal IQs (Flynn, 1984, 1987). In other words, an individual tested on the WISC–R must answer more questions correctly, or must answer harder questions, to obtain the same score as on the WISC. Because the Flynn effect takes effect immediately on the introduction of a new IQ test, the norms are most valid at the time the norms are released. Although there is not a consensus among professionals as to why these gains are occurring or what these gains actually mean (e.g., are we really getting smarter?), all are in agreement that the gains occur and that they hold great theoretical and practical importance (for a review, see Neisser, 1998).

Although the nature of increasing IQ scores and the impact of new norms on changes in IQ have been studied, the repercussions such changes may have on American social policies, particularly special education policy, have rarely been examined (but see Flynn, 2000). Most relevant to the aims of the present study, in a reanalysis of 26 studies compiled by Zimmerman and Woo-Sam (1997), Flynn estimated the mean difference in Full Scale IQ scores between the WISC–R and the WISC–III to be 5.3 points (Flynn, 1998). In practical terms, this means that someone who received a score of 105 on the WISC–R would on average receive a score of 100 on the WISC–III. Assuming that this estimated difference in mean scores between the WISC–R and the WISC–III holds true at other points of the IQ distribution, the same IQ cutoff score of 70 that captured the bottom 2.27% (two standard deviations below the mean) in 1974 under the brand new norms of the WISC–R would only capture the bottom 0.94% 17 years later, right before the newly normed WISC–III was introduced in 1991. Therefore, if the 1974 norms were used to score individuals in 1992, fewer than half as many would be

Matthew H. Scullin

considered as having an IQ score below 70, the usual cutoff for diagnosing someone with MR.

A few studies have noted a decline in IQ scores among children classified as MR once the WISC–III replaced the WISC–R. In a study reported in the WISC–III manual, the mean WISC–R Full Scale IQ in a sample of MR children was 8.9 points greater than the mean WISC–III Full Scale IQ (Wechsler, 1991). Given there is far less variability in performance of MR students compared with students who fall within the average IQ range, this 8.9-point difference is actually larger than the MR populations' entire standard deviation of 7.8 points. This general finding within MR samples has been replicated by other researchers, although the exact difference in Full Scale IQ scores between the WISC–R and the WISC–III varies between 5 and 9 points (Bolen, Aichinger, Hall, & Webster, 1995; Slate & Saarnio, 1995; Vance, Maddux, Fuller, & Awadh, 1996). This suggests that the Flynn effect and changing IQ norms not only affect individuals in the average range of the IQ distribution but also individuals with MR, and that the magnitude of the effect may be even larger among MR children than among children at the mean.

As we address in greater detail later, knowing that the Flynn effect exists among individuals evaluated for MR calls to question the use of fixed IQ cutoff scores to determine eligibility for an MR diagnosis. This, in turn, can have a large impact on many U.S. policies, including educational financing, social security disability benefits, eligibility for the death penalty, and occupations within the military. In other words, fluctuations in IQ scores as a result of aging IQ norms being replaced by new, harder norms could have unexpectedly large public policy implications. We briefly describe some of the potential policy implications below so that readers can judge for themselves the importance of this issue.

Educational Implications: Who Receives MR Services

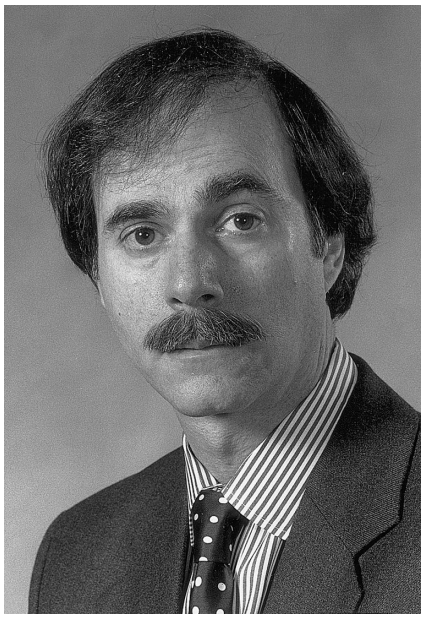
Although the exact services a child diagnosed with MR receives vary greatly between school districts, some of the more common services offered to MR students include modified regular classroom assignments (usually making the assignments shorter and/or easier) and direct instructions that explicitly teach the skills necessary to complete assignments such as organizing materials for the student or showing the student exactly where the necessary information is in the text (Burns, 2003). For many MR students, special education services may include more extensive educational interventions, such as removal from regular classrooms for all or part of the day to receive instruction from special education specialists, paid aides, and volunteers (Singer, Butler, Palfrey, & Walker, 1986; U.S. Department of Education, 2002). As IQ norms age, fewer students receive MR services, but when a newly normed test is introduced, the number of students eligible for these services will suddenly increase. Although there are positive benefits to receiving individualized MR services, there are also some negative psychosocial consequences associated with the label *mentally retarded* (e.g., Baroff, 1999; Mercer, 1973).

Financial Implications: The Costs of MR

The federal Individuals With Disabilities Education Act (IDEA) and its predecessor, P.L. 94-142, mandate that individuals with MR receive educational and other services for their disability (U.S. Department of Education, 2002). Consequently, local, state, and federal governments collectively spend billions of dollars on special education, and drastic fluctuations in the number of students eligible for MR programs due to changing test norms could have dramatic financial implications. These financial implications also hold true for individuals who qualify for social security disability benefits based on MR status.

Legal Implications: The Flynn Effect and the Law—A Matter of Life or Death

Ever since *Buck v. Bell* (1927), in which the Supreme Court upheld eugenics sterilization laws and Justice Oliver Wendell Holmes Jr., writing for the majority, famously stated that "three generations of imbeciles are enough," MR has had a long and complex relationship with the law. Most recently, the Supreme Court ruled in *Atkins v. Virginia* (2002) that the execution of people who are mentally retarded constituted "cruel and unusual punishment" under the Eighth Amendment. Justice Stevens explained that because of their "disabilities in areas of reasoning, judgment, and control of their impulses [the mentally retarded] do not act with the level of moral culpability that characterizes the most serious adult criminal conduct" (p. 1). Therefore, a diagnosis (or nondiagnosis) of MR could, literally, be the difference between life or death for some prison inmates.



Stephen J. Ceci
Photo by Charles Harrington

Occupational Implications: The Flynn Effect and the Military

All individuals interested in joining military services must take the Armed Services Vocational Aptitude Battery (ASVAB), an eight-subtest, multiaptitude, multiple-choice exam. The ASVAB is looked on as a measure of trainability and a predictor of job performance within the armed forces. Four of these subtests comprise the Armed Forces Qualification Test (AFQT), which bears resemblance to a truncated form of the Wechsler series and is known to be highly *g*-loaded (Ceci, 1996). Two other ASVAB subtests are then added to the AFQT to determine an individual's Military Career Score. It is this score that determines one's eligibility for various military occupations (and their subsequent pay raises and benefits), as most of these occupations require a minimum score (U.S. Department of Defense, 2001). In peacetime, the United States military imposes a limit on the percentage of low-IQ recruits formerly referred to as "Category 4s," who score in the 10th–30th percentile on the AFQT, it accepts. There is a long history documenting the reasons for this limit (Stitch, 1991). Given the similarity between the AFQT and the Wechsler tests as well as the similarity between "Category 4s" and individuals with borderline intellectual functioning, the impact of the Flynn effect on borderline children will have a significant relevance to occupational policy within the military.

Exploring the Flynn Effect's Impact on MR Diagnoses

In summary, the role of the Flynn effect on MR diagnoses is understudied, yet appears to have important implications for U.S. national policies. Using a database of the longitudinal testing history of 8,944 students from nine geographically and demographically diverse school districts from around the United States, we sought to address several hypotheses about the impact of the Flynn effect and changing IQ norms. We looked specifically at school children tested and in some cases retested for special education programs on the WISC series who fell in the borderline range, encompassing the standard deviation above the typical IQ 70 cutoff for MR services (i.e., IQ 71–85 on the Wechsler series IQ tests), and in the mild MR range, encompassing the standard deviation below and including the IQ 70 cutoff for MR services (i.e., IQ 55–70). Using this dataset, we examined three hypotheses.

Hypothesis 1: The Size of the Flynn Effect in the Mild MR to Borderline Range Will Be Large Enough to Have a Significant Effect on Placement Decisions

As mentioned in the introduction of this article, several researchers have already noted that MR children scored lower on the WISC–III than on its predecessor, the WISC–R. All of these previously cited studies had relatively small sample sizes and did not study children who fell in the region critical for our hypotheses, namely, individuals who scored just above the 70 cutoff for MR and who were not placed into an MR program. On the basis of the estimates from these other studies, we anticipated that children in our dataset who are just above and below the 70 cutoff point for MR would exhibit a 5–9 point drop in IQ scores when they were retested on the WISC–III, thus resulting in significant elevations in MR diagnoses—despite controlling for actual cognitive ability. We make this expectation explicit in our second hypothesis.

Hypothesis 2: The Introduction of a Renormed IQ Test Will Strongly Affect Children in the Borderline Range During Their Triennial Reevaluations

Children who score just above 70 typically fall into a gray area in special education programs. Their IQ scores may be considered too high for MR programs but may also be too low for learning disabled (LD) programs. Some children who score just above 70 will be classified as MR because they have had a history of receiving MR services, have poor adaptive behavior, or have the measurement error for their IQ score taken into consideration. Others may receive special education services for behavior disorders, LD, or other reasons. However, by law, all students receiving special education services must be reevaluated at least every three years to determine if they are still eligible for continuing services. Often, a new IQ test is given, and special education placements are changed when deemed necessary. We hypothesized that more children who initially tested above 70 on a WISC–R would be recommended for MR services after being retested on a WISC–III than would be recommended for MR services after being tested and retested on a WISC–R or tested and retested on a WISC–III. On the basis of our initial estimates, we

expected that the magnitude of this difference would be quite large.

We further hypothesized that because of normal fluctuations in retest scores and cognitive losses by some special education children, a certain percentage of borderline students who initially score above the cutoff score of 70 will see their scores drop below 70 when tested and retested on the same WISC–R norms, but that the percentage of students who will drop below 70 will be much larger when children are originally tested on the WISC–R but retested on the newer WISC–III norms. However, we expected that the percentage of children who are reclassified as MR will be smaller than the percentage of children whose IQ scores drop below 70. This would occur because psychologists will take children's earlier IQ scores and placement decisions into account when making a determination on whether to classify a child as MR or not. Thus, the impact of the Flynn effect will be somewhat mitigated because the psychologist has some discretion over how much weight to give a low test score.

Hypothesis 3: The Probability of an MR Diagnosis for Borderline Children Will Increase on the Introduction of New IQ Test Norms

Although on theoretical and measurement grounds we were led to this prediction at the outset to this study, we felt that this was our most controversial expectation. We examined this hypothesis by creating IQ score ranges of children who were administered the WISC–R and the WISC–III that reflected a common metric based on our estimate of the size of the Flynn effect in the MR and borderline range. In other words, an IQ range of children who had been administered the WISC–III were paired with an equivalent IQ range of children who had been given the obsolete WISC–R a mere one to three years earlier. In our comparison of greatest interest, we grouped data from the range of children who scored just above the threshold for MR on the WISC–R with the equivalent range of children who scored just below the threshold for MR on the WISC–III. We predicted that in this comparison, children who were tested on the WISC–III will be much more likely to be classified MR than the equivalent IQ range of children who were tested on the WISC–R. As will be seen, testing this hypothesis requires certain statistical adjustments to equate the children's cognitive ability across IQ test norms.

Method

Sample

IQ data from 8,944 school psychologist special education assessments were collected from nine different school districts across the United States representing a diverse sample of geographical regions (Midwest, Southeast, West, South), neighborhood types (rural, urban, suburban), and participants' socioeconomic status. Data included students' gender, age, testing date, IQ scores, test–testing norms used, and special education placement recommendations (17 mutually exclusive categories as well as a dichotomous MR or not MR category). Information regarding evaluation type (e.g., initial evaluation, retest, triennial reevaluation) was obtained for all testings. Students' ages at the time of the most recent evaluation ranged from 6 to 17 years, and targeted testing dates spanned from 1989 to 1995, covering the WISC–R to WISC–III transition. If children were tested multiple times, all IQ test data available in the children's files were collected, including test data from before and after the target test dates. Data were gathered by traveling to each school district and recording all necessary information from each student's psychological testing file. This usually entailed two people working full time for five to seven days to complete the transcriptions for a single school district.

As a result of collecting longitudinal data on children who were tested during the targeted time frame, the dataset includes students who were repeatedly tested, typically for a required triennial reevaluation. Some were repeatedly tested on the same test (e.g., repeatedly tested either on the WISC–R or on the WISC–III), and some were retested on a different test (e.g., initially tested on the WISC–R but then retested on the WISC–III).

Table 1 shows the exact breakdown of WISC–Rs and WISC–IIIs given per year during this time frame. For samples in which we examined test–retest data, analyses were restricted to records in which students scored between plus or minus one standard deviation from the MR cutoff score—that is, between 55 and 85 on their Time 1 WISC series test—and then were subsequently retested on a WISC series test that was administered less than 48 months

Table 1
Number of Students With IQ Scores Between 55 and 85 Tested on the WISC–R and WISC–III Between 1989 and 1995

Test type	1989	1990	1991	1992	1993	1994	1995
WISC–R	353 (100%)	484 (100%)	498 (88%)	144 (41%)	47 (17%)	27 (10%)	8 (4%)
WISC–III	0 (0%)	0 (0%)	70 (12%)	210 (59%)	227 (83%)	238 (90%)	217 (96%)
Total	353	484	568	354	274	265	225

Note. WISC–R = Wechsler Intelligence Scale for Children—Revised; WISC–III = Wechsler Intelligence Scale for Children—Third Revision.

later. For samples in which we examined single test data, analyses were restricted to children who scored between 55 and 85 on a test, in which case one test per child was randomly selected for inclusion. As will be made clear below, the sample sizes vary depending on the type of analyses conducted. Finally, we note that many of the children in our sample were tested for LD, behavior disorder, or other special education services and had IQs above the range we examined in this study.

Results

Hypothesis 1: The Size of the Flynn Effect in the Mild MR to Borderline Range

We initially split the analyses into two different ranges to assess whether the Flynn effect had a similar impact for borderline children who scored between 71 and 85 and for mild MR children who scored between 55 and 70. These ranges approximate the second and third standard deviations below the mean IQ score of 100 on the Wechsler series. Table 2 shows the descriptive variables of the borderline children who initially scored between 71 and 85 when they were both tested and retested on the WISC–R (the WR/WR group), tested on the WISC–R and retested on the WISC–III (the WR/W3 group), and tested and retested on the WISC–III (the W3/W3 group). We calculated difference scores (*d* scores) between Time 2 testings and Time 1 testings and, as expected, found that the one-way analysis of variance (ANOVA) for *d* scores by WISC group was significant, $F(2, 523) = 27.19$, $p < .001$, with a priori post hoc comparisons confirming that children in the WR/W3 group had significantly greater magnitude *d* scores ($M = -4.48$) than either the WR/WR group or the W3/W3 group (*M*s = 1.17 and −0.44, respectively), who also did not significantly differ from each other.

Similar results are shown in Table 3 when we examined mild MR children who initially scored between 55 and 70 on the WISC series tests. Once again, the one-way ANOVA for *d* scores was significant, $F(2, 214) = 12.5$, $p < .001$. The planned follow-up comparisons revealed a similar pattern of results with *d* scores for the WR/W3 group being significantly more negative ($M = -5.32$) than for the WR/WR and W3/W3 groups (*M*s = .12 and .81, respectively), with the WR/WR and W3/W3 groups again not significantly differing from each other. Thus far, the data are in line with our predictions.

To control for some variables known to affect IQ scores, in Table 4 we conducted a regression analysis in which age at Time 1, number of years between initial testing and retesting, and IQ at Time 1 were all included as covariates and the WR/WR group *d* scores were used as the reference group. WR/WR *d* scores were used as the reference group because in most of the school districts we visited, special education policies were changed in the mid-1990s so that if a child's IQ had remained stable for a couple of testings, it was no longer considered necessary to continue administering full IQ tests unless a change in placement was being considered. With these covariates, children who were initially tested with the WISC–R and subsequently retested with the newer WISC–III norms (WR/W3 group) dropped 5.6 IQ points when retested. In contrast, children who were tested and retested on the WISC–III (the W3/W3 group) did not differ significantly from children in the WR/WR group. In sum, the all-important WR/W3 IQ difference was precisely along the lines predicted.

In all three groups, IQ scores in the 55–85 range remained relatively stable from testing to retesting considering the extreme restriction of range of the sample, with test–retest correlations ranging only from .72 to .76. This illustrates the point that major differences in group mean scores are not necessarily associated with changes in correlations if rank ordering remains similar.

Thus, on the basis of the means, medians, and regression estimates of the size of the WR/W3 difference for children in the mild MR and borderline groups, it would

Table 2
IQ Scores for Children (N = 526) Who Initially Scored Between 71 and 85 When Retested on the Same and Revised WISC Tests

		M T1		*M* T2			
T1 test to T2 retest	*N*	Age (months)	IQ score	Age (months)	IQ score	*M* T2–T1 IQ *d* score	Median T2–T1 IQ *d* score
WISC–R to WISC–R	192	117.6 (25.0)	79.0 (4.6)	151.3 (25.2)	80.2 (8.9)	1.2[a] (7.0)	1.0
WISC–R to WISC–III	157	113.9 (25.7)	78.4 (4.4)	148.5 (25.6)	73.9 (8.8)	−4.5[b] (7.4)	−5.0
WISC–III to WISC–III	177	117.6 (24.8)	78.5 (4.2)	150.7 (25.0)	78.1 (8.4)	−0.4[a] (7.15)	0.0

Note. Standard deviations are in parentheses. T1 = Time 1; T2 = Time 2; WISC–R = Wechsler Intelligence Scale for Children—Revised; WISC–III = Wechsler Intelligence Scale for Children—Third Revision.
[a,b] *d* scores with different subscripts differ from each other at the $\alpha < .05$ level after a Bonferonni correction.

Table 3
IQ Scores for Children (N = 217) Who Initially Scored Between 55 and 70 When Retested on the Same and Revised WISC Tests

T1 test to T2 retest	N	M T1		M T2		M T2–T1 IQ *d* score	Median T2–T1 IQ *d* score
		Age (months)	IQ score	Age (months)	IQ score		
WISC–R to WISC–R	81	119.3 (24.1)	64.1 (5.2)	154.5 (24.6)	64.2 (8.1)	0.12[a] (7.4)	0.0
WISC–R to WISC–III	53	122.7 (27.5)	64.2 (4.9)	157.8 (27.9)	58.9 (7.9)	−5.3[b] (6.8)	−6.0
WISC–III to WISC–III	83	123.0 (25.8)	63.5 (4.5)	155.7 (24.5)	64.3 (8.2)	0.81[a] (7.8)	1.0

Note. Standard deviations are in parentheses. T1 = Time 1; T2 = Time 2; WISC—R = Wechsler Intelligence Scale for Children—Revised; WISC–III = Wechsler Intelligence Scale for Children—Third Revision.
[a,b] *d* scores with different subscripts differ from each other at the $\alpha < .05$ level after a Bonferonni correction.

appear that the size of Flynn effect in these groups is very close to Flynn's (1998) estimate of a 5.3-point difference between the average scores of the older WISC–R and the average scores of the newer WISC–III norms. So, our best estimate is that the Flynn effect falls between 5 and 6 IQ points in the mild MR and borderline ranges, almost exactly the same magnitude that Flynn found in the middle of the IQ distribution.

Hypothesis 2: Changes in MR Classification Because of the Flynn Effect

Figure 1 depicts the percentage of borderline children who initially scored just above the MR cutoff of 70 (between 71 and 85) on a Wechsler series test but who received a recommendation from the psychologist for an MR classification. Initial classification as MR means that the child was recommended to be classified MR by the school psychologist even though the child scored above the cutoff score of 70. The percentage of children receiving an MR classification was consistently around 9% across groups, $\chi^2(2, N = 526) = 0.29$, *ns*. Retest classification as MR reflects the percentage of children who were recommended to be classified as MR after being retested. Whereas only 9.6% of WR/WR children and 8.3% of W3/W3 children were classified as MR on retesting, 19.8% of WR/W3 children were classified as MR, $\chi^2(2, N = 526) = 12.16$, $p < .01$, approximately doubling of the percentage of MR recommendations. Hence, in contrast to students who were retested on the same version of the IQ test on which they had been originally tested, being retested on newer, harder WISC–III norms posed a significantly greater likelihood of MR classification for those originally tested with the older WISC–R norms—as predicted.

There is also evidence that psychologists were reluctant to classify everyone who retested under 70 as MR once the WISC–III was introduced. The "Retest Under 70" bars in Figure 1 reflect the actual percentage of children who scored under 70 when retested. Among WR/WR children, only about 13.0% received Full Scale IQ scores of 70 or lower when retested. This contrasts sharply with the 34.4% of WR/W3 children who received scores of 70 or lower when retested. Among W3/W3 children, 20.3% scored below 70 when retested. The difference in counts between groups was highly significant, $\chi^2(2, N = 526) = 23.46$, $p < .01$. It is unclear whether the increase in the percentage of children who scored below 70 in the W3/W3 group vis-à-vis the WR/WR group was due to the WISC–III being a more difficult test or due to changes in educational policy in the mid-1990s that reduced the number of children who were administered the WISC–III if their IQ had been stable for a couple of prior testings. Overall, our results suggest that psychologists did not immediately classify all students who dropped below 70 as MR. However, the change in test norms did result in a near doubling of the number of children classified as MR for the children in the WR/W3 group. As will be seen in the next section, in which we analyze single test from children, psychologists often did not classify children as MR even when they tested well below 70.

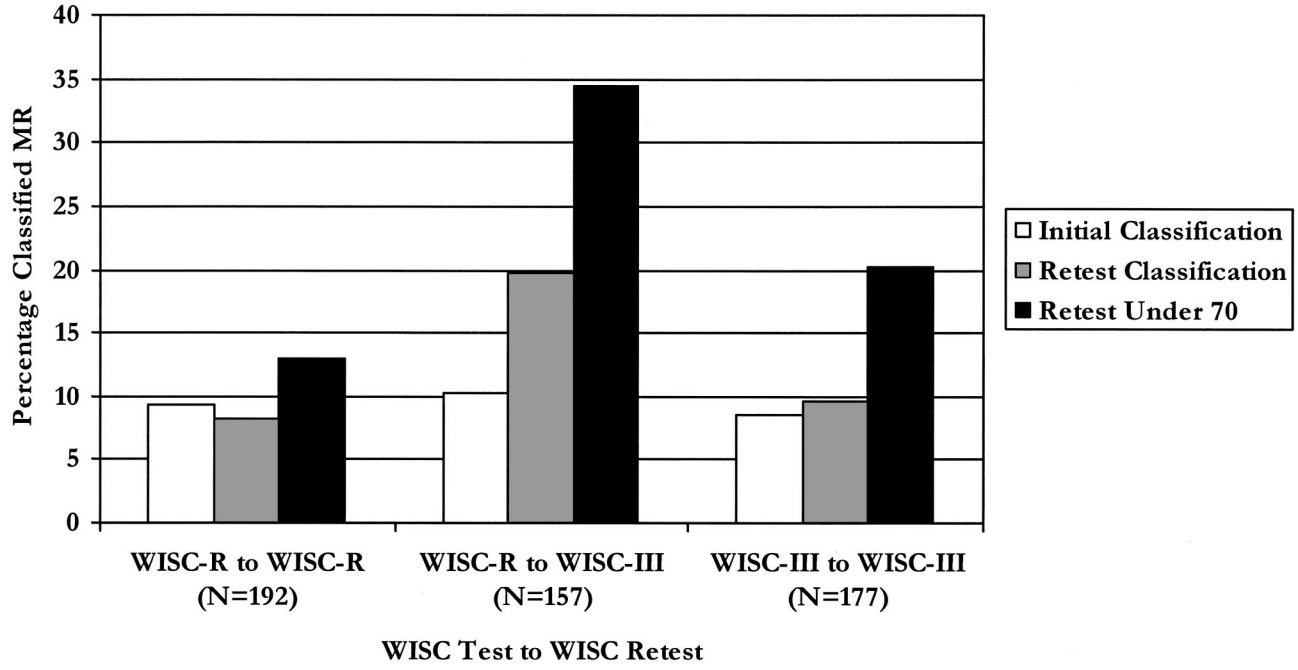
Table 4
Summary of Regression Analysis for Variables Predicting IQ Difference Score (N = 743)

Variable	*B*	*SE B*
Age in months at first testing	−0.02	.01
Time between testings	−0.05	.03
IQ at Time 1	−0.01	.03
WISC–III test/WISC–III retest	−0.94	.63
WISC–R test/WISC–III retest	−5.55*	.67

Note. $R^2 = .10$. WISC–R test/WISC–R retest is the reference group. WISC–R = Wechsler Intelligence Scale for Children—Revised; WISC–III = Wechsler Intelligence Scale for Children—Third Revision.
* $p < .01$.

Figure 1

Percentage of Children Classified in the Mental Retardation (MR) Range Who Initially Scored Between 71 and 85 on a WISC Series IQ Test When Tested and Retested on the Same Versus Revised WISC Tests



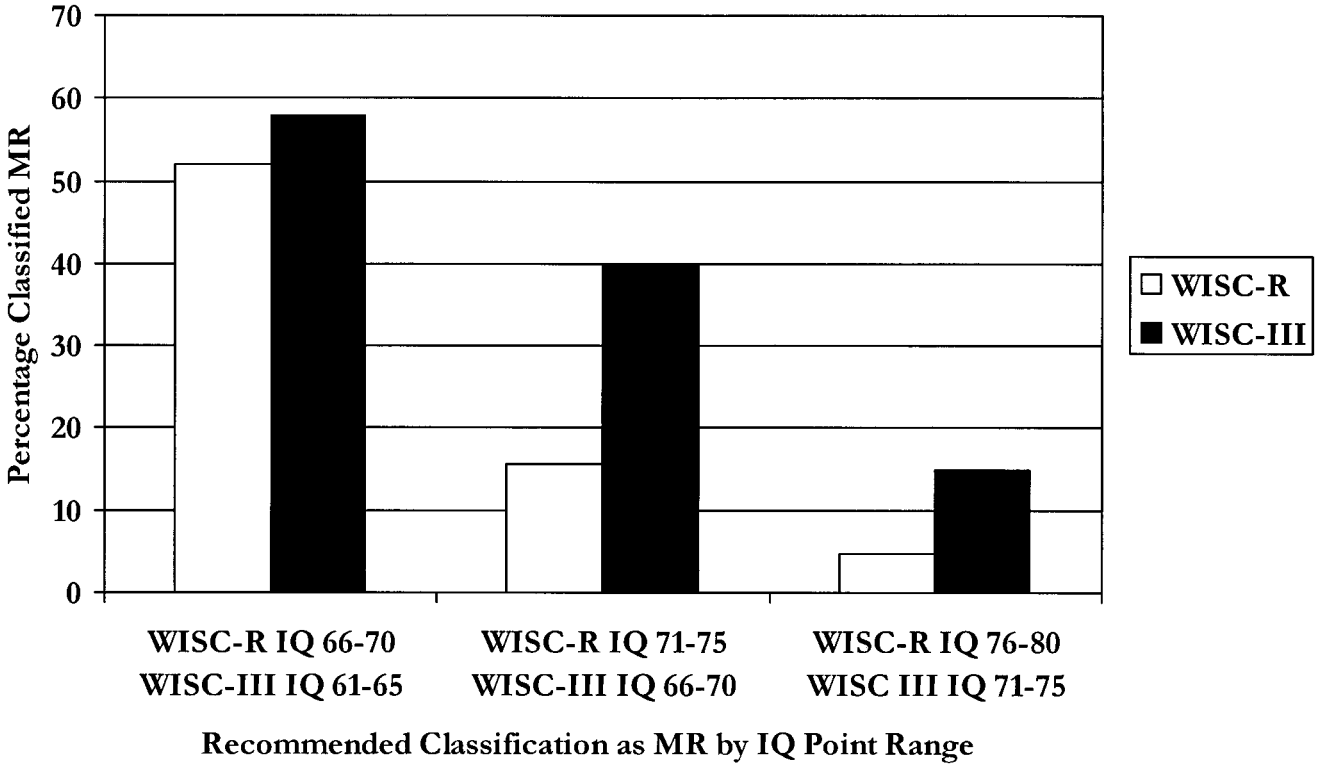
Note. Difference for counts classified MR between retest types, $\chi^2(2, N = 526) = 12.16$, $p < .01$; difference for counts of 70 and below versus 71 and above between retest types, $\chi^2(2, N = 526) = 23.46$, $p < .001$. WISC–R = Wechsler Intelligence Scale for Children—Revised; WISC–III = Wechsler Intelligence Scale for Children—Third Revision.

Hypothesis 3: Impact of the Flynn Effect on Borderline IQ Ranges

For the following analyses, we used a conservative estimate of the size of the Flynn effect from the WISC–R to the WISC–III among children in the mild MR and borderline range—5 IQ points. We examined the impact on children who would have been on the cusp of qualifying for MR services on the older test using a single test per child (thus including children who were tested once and did not receive special education services as well as children who were tested multiple times). In Figure 2, adjacent bars represent 5 IQ point ranges that could be considered equivalent given the Flynn effect (e.g., a range of 71–75 on the WISC–R is equivalent to a range of 66–70 on the WISC–III). We chose to examine IQ in 5-point ranges because if WISC IQ scores fall in a perfect normal distribution, the number of children in 5-point ranges above IQ 60 increases at a nearly geometric rate. For example, although only 0.4% of children would be expected to have IQ scores below 60, 0.6% would be expected to have scores from 60 to 65, 1.3% from 65 to 70, 2.5% from 70 to 75, and 4.3% from 75 to 80. Thus, among all children who have IQ scores of 75 or lower, mathematically more than half of these children would be expected to have IQ scores between 70 and 75 (2.50% of all children would score 70–75, 2.28% would score 70 or lower). As can be seen by the *N*s for Figure 2, the number of children tested in each IQ range does approximate the geometric progression that would be expected given a normal curve. Note that in the case of the WISC–R, a falloff of IQ test administration in Range 3 (IQ 76–80) would be expected because these children typically do not qualify for special education services.

Of greatest interest to us are the children in Range 2 who had IQ scores of 71–75 on the WISC–R and IQ scores of 66–70 on the WISC–III. There is a nearly threefold increase in the percentage of children in this IQ range classified as MR on the WISC–III when compared with the percentage of children who scored in the equivalent range on the WISC–R. As might be expected given our earlier analyses of children who were tested and retested, the percentage of children who tested between 66 and 70 on the WISC–III who were classified MR was smaller than the percentage of children who tested between 66 and 70 on the WISC–R. In the next IQ range of children who scored 71–75 on the WISC–III compared with children who scored 76–80 on the WISC–R, there continued to be a threefold percentage increase of children classified as MR on the WISC–III when compared with the WISC–R (14.8% vs. 4.8%).

Figure 2
Percentage of Children Classified in the Mental Retardation (MR) Range by 5 IQ-Point Range on the Wechsler Intelligence Scale for Children—Revised (WISC–R) and the Wechsler Intelligence Scale for Children—Third Revision (WISC–III)



Note. *N*s for Range 1: WISC–R = 52, WISC–III = 67; *N*s for Range 2: WISC–R = 90, WISC–III = 95; *N*s for Range 3: WISC–R = 127, WISC–III = 169.

It is interesting to note that only 104 out of 214 children who received either a WISC–R or WISC–III score of 70 or below were classified as MR by the psychologist administering the test. Of the 110 children receiving other special education placement recommendations, nearly half (53) were recommended for LD services. This usually occurred when there was a significant discrepancy between the Verbal and Performance subtests of the WISC. For example, a child with a Verbal IQ score of 64 and a Performance IQ score of 79 could obtain a Full Scale IQ score of 69 on the WISC–III. Even though the Full Scale IQ is within the MR range, the psychologist might consider the Full Scale IQ score to be invalid and recommend LD services. LD services were sometimes also recommended for children with Full Scale IQs in the MR range when a child received an IQ score substantially lower than in prior testings that had placed the child into LD services in the first place. Among children not receiving MR or LD recommendations, an additional 26 were not recommended for special education. Fewer than 10 children were in each of the additional categories of behavior disordered, emotionally disturbed, otherwise health impaired, speech services, and nonspecific special education.

In summary, although psychologists compensated (to some extent) for the harder norms of the WISC–III by not making as many MR recommendations for children who scored just under 71 on the WISC–III as they did for children who scored just under 71 on the WISC–R, the harder norms of the WISC–III nonetheless resulted in a marked increase in the number of children eligible for MR services among children who would have scored between 71 and 75 if they had been administered the WISC–R. Because of this, children of equivalent cognitive abilities were diagnosed differently with regards to MR. And, as predicted, the children who were at the cusp of the IQ cutoff score of 70 were most vulnerable to differential diagnoses.

Discussion

According to most criteria, the diagnosis of mental retardation consists of a subaverage intellect, usually specified as an IQ score of 70 or below, in addition to evidence of limited adaptive life skills and an onset before adulthood. Because of the systematic increase in IQ scores seen throughout the last 80 years (the Flynn effect), there is reason to believe that many students are diagnosed as MR

based on the year in which they are tested and test norms used rather than on their cognitive ability. More specifically, as norms age, fewer children are diagnosed MR as more children's IQ scores rise above the 70-point cutoff. With the introduction of newer norms, suddenly more children score below the 70-point cutoff. As shown in this report, identical levels of cognitive performance on the Wechsler IQ test led to large disparities in MR classification rates.

Despite its obvious policy relevance, this is the first study to investigate the size of the rise in IQ scores within the population of children who score in the mild MR and borderline range. Although past studies have examined students who were already diagnosed MR on the WISC–R, the present study examined how the Flynn effect influences actual diagnoses of children who are being considered for either new or continuing placement in special education programs (i.e., students who both were and were not diagnosed MR). Using a large, geographically and economically diverse sample of students tested for special education, the present study explored the nature of the Flynn effect within the mild MR and borderline IQ population.

When we examined the mean differences in test–retest scores of MR and borderline students who were tested and retested on the same IQ norms (e.g., tested and retested on either the WISC–R or the WISC–III), we found that IQ scores on average changed by around 1 IQ point or less. In contrast, the mean difference in test–retest scores in which individuals were tested initially on the WISC–R and subsequently retested on the WISC–III was 4.5 points for children in the 71–85 IQ range and 5.3 for children in the 55–70 IQ range. Our regression model estimate for the full 55–85 IQ range after we controlled for factors known to influence performance showed a decline of 5.6 points, which is approximately the same difference found among average intellect individuals reported by Flynn (1998). However, given the smaller standard deviations found within individuals in the tails of the distribution, this difference is over two thirds (72%) of the standard deviation of MR IQ scores. Hence, a 5.6-point mean difference reflects a far larger magnitude effect than it might if it were to occur near the middle of the IQ distribution.

We examined whether a higher percentage of borderline children were classified as MR on retesting when they were initially tested on the WISC–R and retested on the WISC–III (the WR/W3 group), compared with children who were tested and retested on the same test (the WR/WR group and the W3/W3 group). We found that nearly 20% of children in the WR/W3 group were classified as MR on retesting. This was more than double that of children in the WR/WR group or the W3/W3 group. More than a third of the children in the WR/W3 group who had originally tested in the borderline range later scored below 70 on retesting, markedly exceeding the percentage of children in the other two groups who tested below 70.

Our results also show the Flynn effect has an impact on which individuals are diagnosed MR and which are not, regardless of their actual cognitive ability. We used the average difference between the WISC–R and WISC–III in our sample of 5 points to examine the extent to which an MR diagnosis was dependent on an IQ score. For this analysis, we equated scores on the earlier WISC–R norms with lower scores on the subsequent, harder WISC–III norms. As hypothesized, we found that students' probability for an MR diagnosis changed significantly when IQ scores were adjusted to a common metric by comparing ranges of scores on the WISC–R and the WISC–III that could be considered equivalent given the Flynn effect. There was a tripling in the percentage of MR placement recommendations for children who fell in the upper reaches of the mild MR range (IQ 66–70) during the first five years of the WISC–III when compared with MR diagnoses for children who fell in the cognitively equivalent WISC–R IQ range (71–75) during the last five years of the WISC–R.

In addition, simple descriptive statistics showed that a majority (88%) of the students in our sample were given the WISC–R in 1991, the first year of the WISC–III. In other words, 88% of the students tested in our sample were given a test that had already begun to be outdated, whereas 12% were given a harder, newly normed test. In 1992, 59% were tested on the WISC–III, whereas 41% were tested, and thereby diagnosed and classified, on its outdated, easier predecessor. This overlap of the use of different tests often occurs because the purchase of revised versions of the test depends on school district budgets. Not all psychologists and school districts replace their IQ tests immediately after a newly normed test is released because IQ tests are expensive (approximately $1,000 per set including a supply of test record forms), and school budgets are often not able to accommodate providing new test kits to all of a school district's testers at the same time. Also, some psychologists and districts may prefer not to use a newly normed test until all of the older test record forms are used up, so it may take many years before an older IQ test is completely phased out of a school system. In our experience, before an old test is completely phased out, different children may be tested on different norms in the same year—even within the same school district. These IQ test scores are still compared with one another, regardless of the fact that different norms were used, and diagnoses are assigned accordingly. Even by 1995, the WISC–R had not been completely phased out of the school systems, which means that some students may effectively have different IQ cutoff scores for MR diagnoses than others.

For all of the preceding reasons, the introduction of a newly normed IQ test can create havoc among MR diagnoses for several years after a new test is introduced, and two children in the same classroom with the same cognitive ability could be diagnosed differently simply because different test norms were used for each child. The present study provides for the first time evidence that this may occur. Parents of children will undoubtedly want assurance that their children are receiving services based on their cognitive needs rather than an IQ score that may be inflated or deflated by virtue of the vicissitudes of the Flynn effect, progressively making scores go up until new norms arrive when they plummet back toward baseline.

In the introduction we asserted that the impact of the Flynn effect in the lower ranges of IQ appears to be quite significant and may affect many domains. In the following sections, we reexamine these domains and make the numbers more explicit to drive home the pivotal role the Flynn effect has in the educational, social, legal, and military policy domains of the United States.

Educational Consequences

Each year, nearly two million students are tested for special education services, including MR services. In the 1999–2000 school year, over 600,000 students were receiving MR services (U.S. Department of Education, 2002). Not all of these testings are initial diagnoses, however, as each child receiving special education services must undergo a reevaluation at least every three years to determine if he or she is still eligible for such services. Often, a new IQ test is given during such reevaluations. Therefore, the introduction of new IQ norms could dramatically affect a student's education at the initial diagnosis as well as any subsequent reevaluations.

Some students who would be eligible for MR services under new IQ norms will fail to receive them because the older norms of the IQ test they were given allowed them to score above the cutoff. In addition, students who would not have qualified for MR services had they been tested a year earlier will now do so if they are given an IQ test with newer, harder norms. A very conservative estimate would be that tens of thousands of children tested or retested for MR may be affected by these IQ trends over the course of their school years.

There are also social consequences for children who have been diagnosed as MR. Much research has been conducted on the stigma of classroom labeling (labeling a student at a particular cognitive level such as MR or "low ability") on academic performance and later life outcomes (MacMillan, Gresham, Siperstein, & Bocian, 1996; Mercer, 1973). Many individuals with MR go to great lengths to hide the label and attempt to pass as normal in the community (Mercer, 1973), and many parents are concerned about the negative impact this label will have on their children (Baroff, 1999; Edgerton, 1967). Indeed, the fact that the MR label carries with it an inherent negative stigma is no better illustrated than by the fact that a former label is continually supplanted by newer ones over time. For example, terms such as *imbecile* and *feeble-minded* were considered scientific and acceptable in the first quarter of the 20th century but were replaced after time with successive euphemisms. Even now, some prefer terms such as *general learning disordered*, *developmentally delayed*, or *mentally handicapped* instead of *mentally retarded* or *mentally deficient* to reduce the stigma of the MR label (e.g., Baroff, 1999). Whether escaping the negative consequences of being labeled MR outweighs the benefits of receiving special education MR resources is an interesting empirical question.

Financial Consequences

The financial implications of the Flynn effect on MR diagnoses go beyond mere total dollars spent but also raise questions about whether these resources are properly allocated. If MR students are underdiagnosed as IQ norms age, then students who would have qualified and benefited from these expensive services under newer norms will be shortchanged. The other side of this assertion is that because the new norms are harder and result in lower IQs, one could argue that the resulting increase in diagnoses of MR that accompany these new norms represents an overestimate, with some individuals inappropriately receiving MR services. The present results imply that millions of taxpayers' educational dollars may be misallocated because students are being misdiagnosed every year that an IQ test norm ages.

It is also important to note that many children requiring special education services other than MR have eligibility at least partially determined by an IQ score. LD, for example, is the most common special education diagnosis and is characterized by whether a student's IQ score is sufficiently discrepant from an achievement test score (American Psychiatric Association, 1994). In the case of LD, the introduction of a newly normed test may make it less likely that a child will receive special education because a depressed IQ score will reduce the chance that a significant discrepancy will be found between a child's IQ test score and his or her achievement test scores (e.g., Truscott & Frank, 2001).

The financial impact of a diagnosis or nondiagnosis of MR extends well beyond educational dollars. As mentioned in the introduction, federal social security disability benefits are available for those diagnosed with MR. Those who receive test scores in the borderline range just prior to the introduction of a newly normed IQ test may be denied these benefits.

Legal Consequences

Our results imply that the year that a capital murder defendant was tested can determine whether she or he is sentenced to die as opposed to life imprisonment. This raises concerns regarding inmates on death row who tested above the 70–75 IQ cutoff on a test that was near the end of its norming cycle (when scores are highly inflated) as well as an inmate who tested in the MR range during the earliest years of a new norm (when the test is hardest). Should the State be permitted to insist on new testing to see if the score can be elevated to avoid the MR diagnosis? It is worth noting that Daryl Renard Atkins (of *Atkins v. Virginia*) was tested on the Wechsler Adult Intelligence Scale—III the year it was released. Because of his low IQ score of 59, he was diagnosed MR without the necessity of having to have poor adaptive behavior under SSA rules and was ultimately found to be ineligible for the death penalty. If he had been tested on an IQ test with older norms, however, it is interesting to speculate on his fate as well as the fate of the Supreme Court decision.

Although it is unknown exactly how many MR inmates are on death row and, more importantly, how many have been found to have borderline intelligence on the cusp of MR eligibility, there are currently over 3,000 inmates on death row in the 38 states that have the death penalty (24 of them also allow the death penalty for juvenile offenders). According to Amnesty International (2002), 12 out of the 350 people executed since 1990 were known to have IQ scores of 70 or below. Thus, a potentially important implication of the Flynn effect is that some borderline death row inmates or capital murder defendants who were not classified as MR in childhood because they were administered an older version of an IQ test will qualify as MR if they are administered a more recent test. Given the magnitude of the effect (nearly a full standard deviation decrease in IQ is associated with changing norms since the first edition of the WISC was phased out in the early 1970s), the shifts in eligibility for death row inmates could be significant. Once aware of this effect, attorneys on both sides could be expected to "game" the system by ordering reevaluations, locating archival records that coincided with times when the norms would be most favorable to their case, or adjusting IQ scores to compensate for obsolete norms.

Military Occupational Consequences

For our purposes, the question arises as to whether the Flynn effect is relevant to the AFQT. If its norms become obsolete over time (and given how closely related it is to the Wechsler tests, there is every reason to believe they do), then similar issues would apply to the selection of recruits and their occupations. Specifically, depending on the norms used, a recruit might be eligible or not for military service, and if deemed eligible to enlist whether they will be permitted to enter certain occupations within the armed services and be availed a particular level of pay. This is especially true of those individuals who score in the borderline region, just above or below the cutoff for selection or career placement. Thus, the year that one is tested within AFQT's norming cycle could affect who gets enlisted, selected for training, or deemed to be eligible for certain ranks.

Caveats and Conclusions

In closing, we bring to readers' attention some strengths and weaknesses of the present analyses. As a result of the fairly large sample size of the present study, we were able to systematically estimate the size of the Flynn effect among mild MR and borderline individuals using both initial testings and later reevaluations. As previously mentioned, Flynn himself did not include the MR (or the gifted) in his large, systematic analyses, and all of the published research done on MR individuals mentioned in the introduction (e.g., Bolen et al., 1995; Slate & Saarnio, 1995; Vance et al., 1996) has used comparatively small sample sizes or solely used children who were already classified as MR as opposed to those who were not. In addition, our data represent a wide array of geographic locations, neighborhood compositions, and average socioeconomic status, making our results more generalizable than results gathered from previous research.

Against these strengths, however, we frequently only had access to the psychologists' recommendations for each student's placement. Although in most cases psychologists' recommendations are congruent with the final placement recommendations for a student, at times the placement committee may override the psychologist's recommendation or parents or guardians may place pressure onto the committee so that their child does or does not receive services. Such instances, however, occur rarely and would not change the overall trends found within our analyses. In most court cases, such as in *Atkins v. Virginia*, or in a clinical setting, a psychologist's recommendation is sufficient.

Finally, we do not know whether these trends in MR diagnoses are due solely to the Flynn effect and changing IQ norms as opposed to changing special education policies due to amendments in IDEA, funding fluctuations, or other cohort effects. Notwithstanding this caveat, however, the patterns that were found here are exactly what one would predict assuming a yo-yo trend in IQ scores. Fortunately, regardless of these issues, the actual IQ score differences that were observed are unaffected by any of the above potential variables.

Clearly, the main conclusion that can be drawn from these results is that caution should be used when basing important financial, social, or legal decisions on an IQ score. Perhaps the most important times to be particularly cautious are when a test is either at the beginning or at the tail end of its norming cycle. Although test scores are most valid at the beginning of a norming cycle, they run the greatest risk of being compared to highly inflated scores from the waning years of the previous norming cycle. Of course, a score is least valid when taken from an IQ test at the tail end of its norming cycle. These cautions are especially germane when comparing scores between the same tests at different points of the norming cycle (e.g., comparing an individual tested on the WISC–R in 1972 with another individual tested on the WISC–R in 1980). Nowhere are the consequences of IQ score fluctuations due to the Flynn effect more critical than in the determination of whether a death row inmate (or a capital murder defendant) can be considered mentally retarded. There are approximately 3,525 convicts awaiting execution (NAACP Legal Defense and Educational Fund, 2003). With as many as 10% testing in the MR range and many others in the borderline range, psychologists and other practitioners involved in forensic evaluations must exhibit the highest standard and duty of care.

It may also be important to consider the differential impact of the Flynn effect on African Americans facing the death penalty. Although African Americans constitute 12.8% of the population of the United States, they make up 43% of the inmates on death row (Kane, 2003) and 35% of all executed inmates (Death Penalty Information Center, 2003). In addition, African Americans are overrepresented in special education programs (U.S. Department of Education, 2002). Because African Americans score on average

around 10–15 points lower on IQ tests than Whites (see Jencks & Phillips, 1998, for a review), the percentage of African Americans who fall in the borderline and mild MR IQ range is higher than the percentage of Whites who fall within this range.

Flynn (2000) suggested that a team of qualified psychologists gather a representative sample of MR children based on behavioral criteria and renorm IQ tests every seven years. Although this is a sensible recommendation for the manufacturers of these tests, our concern is more for policymakers and practitioners. Specifically, when psychologists are asked whether someone is MR on the basis of his or her IQ scores (even assuming deficits in adaptive behavior), it may not be sufficient to simply look to see whether the IQ score is below some cutoff point.

REFERENCES

American Association of Mental Retardation. (2002). *Mental retardation: Definition, classification, and systems of supports* (10th ed.). Annapolis, MD: Author.

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

Amnesty International. (2002). *US Supreme Court decision in Atkins vs. Virginia to bring US in line with international standards of decency.* Retrieved August 8, 2002, from http://www.amnestyusa.org/news/2002/usa06202002.html

Atkins v. Virginia, 534 U.S. 1122 (2002).

Baroff, G. S. (1999). General learning disorder: A new designation for mental retardation. *Mental Retardation, 37,* 68–70.

Bolen, L. M., Aichinger, K. S., Hall, C. W., & Webster, R. E. (1995). A comparison of the performance of cognitively disabled children on the WISC–R and WISC–III. *Journal of Clinical Psychology, 51,* 89–94.

Bruininks, R. H., Woodcock, R. W., Weatherman, R. F., & Hill, B. K. (2000). *Scales of independent behavior—revised.* Itasca, IL: Riverside.

Buck v. Bell, 274 U.S. 200 (1927).

Burns, E. (2003). *A handbook for supplementary aids and services: A best practice and IDEA guide "To Enable Children With Disabilities to be Educated With Nondisabled Children to the Maximum Extent Appropriate."* Springfield, IL: Charles C Thomas.

Ceci, S. J. (1996). *On intelligence: A bioecological treatise on intellectual development.* Cambridge, MA: Harvard University Press.

Death Penalty Information Center. (2003). *Race of death row inmates.* Retrieved June 8, 2003, from http://www.deathpenaltyinfo.org/article.php?scid=5&did=184#inmaterace

Edgerton, R. B. (1967). *The cloak of competence: Stigma in the lives of the mentally retarded.* Berkeley: University of California Press.

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

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

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

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

Jencks, C., & Phillips, M. (1998). *The Black–White test score gap.* Washington, DC: Brookings Institution Press.

Kamphaus, R. W. (1987). Conceptual and psychometric issues in the assessment of adaptive behavior. *Journal of Special Education, 21,* 27–35.

Kane, H. (2003). Straight talk about IQ and the death penalty. *Ethics & Behavior, 13,* 27–33.

MacMillan, D. L., Gresham, F. M., Siperstein, G. N., & Bocian, K. (1996). The labyrinth of IDEA: School decisions on referred students with subaverage general intelligence. *American Journal on Mental Retardation, 100,* 161–174.

Mercer, J. R. (1973). *Labeling the mentally retarded.* Berkeley: University of California Press.

NAACP Legal Defense and Education Fund. (2003). *Death row USA, April 1.* Washington, DC: Author.

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

Reschly, D. J. (1981). Psychological testing in educational classification and placement. *American Psychologist, 36,* 1094–1102.

Reschly, D. J., & Ward, S. M. (1991). Use of adaptive behavior measures and overrepresentation of Black students in programs for students with mild mental retardation. *American Journal of Mental Retardation, 96,* 257–268.

Singer, J. D., Butler, J. A., Palfrey, J. S., & Walker, D. K. (1986). Characteristics of special education placements: Findings from probability samples in five metropolitan school districts. *Journal of Special Education, 20,* 319–337.

Slate, J. R., & Saarnio, D. A. (1995). Differences between WISC–III and WISC–R IQs: A preliminary investigation. *Journal of Psychoeducational Assessment, 13,* 340–346.

Social Security Administration. (2002). *Disability evaluation under Social Security.* Washington, DC: U.S. Government Printing Office.

Sparrow, S. S., Balla, D. A., & Cicchetti, D. (1984). *Vineland Adaptive Behavior Scales.* Circle Pines, MN: AGS Publishing.

Stitch, T. G. (1991). Military testing and public policy: Selected studies of lower aptitude personnel. In B. R. Clifford & L. Wing (Eds.), *Test policy in defense: Lessons from the military for education, training, and employment* (pp. 1–76). Boston: Kluwer.

Truscott, S. D., & Frank, A. J. (2001). Does the Flynn effect affect IQ scores of students classified as LD? *Journal of School Psychology, 39,* 319–334.

U.S. Department of Defense. (2001). *Military careers: A guide to military occupations and selected military career paths 1998–2001.* Washington, DC: Author.

U.S. Department of Education. (2002). *Twenty-third annual report to Congress on the implementation of the Individuals With Disabilities Education Act.* Washington, DC: Author.

Vance, H., Maddux, C. D., Fuller, G. B., & Awadh, A. M. (1996). A longitudinal comparison of WISC–III and WISC–R scores of special education students. *Psychology in the Schools, 33,* 113–118.

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

Wechsler, D. (1991). *The Wechsler Intelligence Scale for Children—III manual.* New York: Psychological Corporation.

Zimmerman, I. L., & Woo-Sam, J. M. (1997). Review of the criterion-related validity of the WISC–III: The first five years. *Perceptual and Motor Skills, 85,* 531–546.