Are All IQ Scores Created Equal? The Differential Costs of IQ Cutoff Scores for At-Risk Children

Tomoe Kanaya[1] and Stephen J. Ceci[2]

[1]Claremont McKenna College and [2]Cornell University

ABSTRACT—***The IQ cutoff score of 70 used for the diagnosis of mental retardation has repercussions throughout America, influencing educational, social, and legal decision making. Because of the Flynn effect and changing IQ norms, however, IQ scores are rising and falling over time, independent of actual cognitive gains or losses. These fluctuations, combined with the use of the cutoff score, result in substantial misallocations of financial, educational, and social resources for children of all IQ levels but especially for those at risk for academic failure. Such far-reaching and grave implications call into question the use of IQ cutoff scores in educational and legal policies and underscore the importance for researchers to collect data in real-world settings to understand and appreciate the issues that surround the application of developmental theory.***

KEYWORDS—*IQ; policy; at-risk children; education; Flynn effect*

IQ: THE ''ORIGINAL'' HIGH STAKES TEST

In the past few years, the issue of "high-stakes" testing, particularly as it relates to the No Child Left Behind Act, has received much attention from policymakers, the media, and psychologists (e.g., Dillon, 2005; Tuerk, 2005). But it is the standard IQ test, such as the Wechsler Intelligence Scale for Children (WISC; Wechsler, 1974, 1991), that have represented true high-stakes testing for millions of schoolchildren for more than three decades. As a consequence of the Education for All Handicapped Children Act in 1975 (Public Law 94-142), later renamed the Individuals with Disabilities Education Act (IDEA), all children with physical or mental disabilities are guaranteed free and appropriate special education services. In order to qualify for these resources, however, candidate children are required to take an IQ test to determine if they meet the diagnostic criteria for mental retardation (MR). And for many of the children being tested, a difference of 1 or 2 IQ points can determine whether or not they will receive the educational help they need.

According to numerous different federal guidelines for program eligibility (see Reschly, Myers, & Hartel, 2002, for a review), MR is characterized by "significantly subaverage general intellectual functioning, existing concurrently with deficits in adaptive behavior and manifested during the developmental period (often taken to mean by age 18, though in some definitions by age 22) that adversely affects a child's educational performance" (Code of Federal Regulations, 34 Code §300.7 © (10)). As Reschly et al. (2002) documented, all four major MD diagnostic systems in use today (American Association of Mental Retardation [AAMR], 2002; World Health Organization, 1996; Division 33 of APA [Editorial Board of the APA Division 33, 1996]; and *DSM–IV* [American Psychiatric Association, 1994]) employ an IQ cutoff score of 70 or below to identify "significantly subaverage" intellectual functioning. The primary difference among these four systems involves their

Correspondence concerning this article should be addressed to Tomoe Kanaya, Department of Psychology, Claremont McKenna College, 850 Columbia Avenue, Claremont, CA 91711; e-mail: tomoe.kanaya@claremontmckenna.edu.

view of the meaning and contributing role of adaptive behavior. In some definitions (Division 33 of the American Psychological Association and AAMR), adaptive behavior is construed as distinct from intellectual functioning, whereas in other definitions (World Health Organization), it is considered a result of deficits in intellectual functioning. The four systems also vary in whether they consider adaptive behavior to be a single factor or as having multiple factors or domains (such as social skills or daily living skills) that are hierarchically organized.

Although MR is the only special education category with a defined cutoff IQ score, the possibility of MR has to be ruled out for all the other 12 categories. For example, if a child's IQ is above the cutoff and there is no evidence of physical impairments, MR will be ruled out, but a specific learning disability (LD) diagnosis may be considered (e.g., Kavale & Forness, 1992). Therefore, in practice, most of the children tested for special education services are given an IQ test to see which side of the cutoff score they fall on. (The sole exception to this practice is Multiple Disabilities.)

THE MR CUTOFF AND THE FLYNN EFFECT

The MR cutoff score of 70 was chosen for psychometric purposes. All the major IQ tests are now normed to a mean of 100 and a standard deviation of 15 points. Therefore, in theory, a score of 70 or below, which is exactly 2 *SD* below the mean, represents the bottom 2.27% of the IQ distribution, the group that many experts feel is most in need of special assistance. In practice, however, a cutoff of 70 is quite problematic due to the Flynn effect, the steady rise in IQ scores seen throughout the world over the past half century or so (e.g., Daley, Whaley, Sigman, Espinosa, & Neumann, 2003; Flynn, 1984, 1987).

The Flynn effect occurs because as IQ test norms get older, people perform better on them, raising the mean IQ by several points within a matter of years. The reason for this rise is unclear. Given its magnitude and the speed with which it occurs, it cannot be attributed to genetics (Neisser et al., 1996). What seems the likely explanation for the Flynn effect are the many environmental changes that have coincided with it, including advancements in technology, schooling, and nutrition (for a detailed examination of the possible underlying causes of the Flynn effect, see Neisser, 1998). A model proposed by Dickens and Flynn (2001) suggests how even small environmental changes can lead to large, multiplying effects on cognitive performance among individuals who have only slight genetic proclivities for some type of cognitive performance. Ultimately, however, there is no consensus as to why these gains are occurring in comparable magnitude in both developed and undeveloped nations.

Due to the Flynn effect, as the mean performance on an IQ test increases, the MR cutoff score of 70 captures a steadily smaller percentage of the population with every passing year. Because of the mathematics of the bell curve, even small changes in the center of the distribution can result in large shifts at the tails. An increase of a few IQ points from the mean, from 100 to 103, for example, will result in far fewer scores below 70, as the distribution of raw scores slides rightward relative to earlier cohorts. A change of 1 IQ point from 69 to 70 will result in the number of persons eligible for MR decreasing from 2.68 to 2.27.

To compensate for the upward creep in IQ scores, IQ tests are renormed every 10–20 years, making the tests harder. Renormed tests reset the mean back to 100, "hiding" the previous IQ gains and resituating the cutoff score of 70 exactly 2 *SD* below the mean. Consequently, when a new version of an IQ test is introduced, it is accompanied by a sudden and significant increase in MR diagnoses because the bottom 2.27% will once again fall below 70, causing a direct and immediate impact on the lives of children, families, schoolteachers, and administrators.

In our own analyses of data on more than 10,000 special education evaluations from 10 school districts around the country, we found that on average, children in the MR and borderline MR range lost approximately 6 IQ points in the transition from the older WISC norms Wechsler Intelligence Scale for Children–Revised [WISC–R] to the newer ones Wechsler Intelligence Scale for Children, Third Edition [WISC–III]; (Kanaya, Scullin, & Ceci, 2003). More important, this precipitous drop in IQ had a significant impact on MR diagnoses over time. Not only were children of the same cognitive ability receiving different MR diagnoses due to the different test norms used but the same children also received a new diagnosis at the time of their reevaluation—independent of actual cognitive gains or losses since their previous evaluation. Further analyses of state and national data reveal a steep 12-year decline in the number of students receiving MR services prior to the introduction of the WISC–III in 1991. With the introduction of the new norms, the decline reversed, further suggesting that the Flynn effect is having a significant and pervasive impact on MR diagnoses (Scullin, 2006).

Needless to say, the yo-yo pattern in IQ created by the Flynn effect and changing norms makes the use of cutoff scores very problematic. Whether or not children in need of special education services actually receive them may depend not on their disability but on the year they are tested. Under IDEA, each child receiving special education services costs the school district approximately $9,000 (U.S. Department of Education, 2005). Therefore, the fluctuations in MR diagnoses that occur due to the natural fluctuations in the Flynn effect can result in millions of dollars in misallocated resources each year. Nevertheless, schools, parents, and policymakers continue to treat these increases and decreases as though they were the consequence of educational programs rather than simply stochastic fluctuations that occur independently of changes in schooling or in cognitive ability. In addition, given that individuals with MR

qualify for Social Security Disability Insurance[1] and are exempt from the death penalty (*Atkins v. Virginia*, 2002), MR diagnoses have far-reaching implications throughout the life course.

THE MR DIAGNOSIS: THE GAP BETWEEN THEORY AND PRACTICE

It is important to note that the AAMR (2002) and the Editorial Board of the APA Division 33 (1996) have recommended a 70–75 IQ interval, rather than a strict cutoff to account for measurement error (which on common IQ tests runs about 4 points). During our data collection, however, we found that this recommendation is rarely followed (Kanaya et al., 2003). This can be seen by comparing the MR rates of children who received WISC–R scores between 71 and 75 with those of children who received WISC–III scores between 66 and 70. Although these two groups occupy the identical locations in the IQ distributions (as noted, children score approximately 6 points lower on the WISC–III than on the WISC–R), they are on the opposite sides of the 70 cutoff score. Yet, despite the fact that these children represent the same cognitive ability *and* that they all fall within the recommended interval score for MR on this test, there was a nearly threefold increase in the percentage of children diagnosed MR on the WISC–III compared with those tested on the WISC–R. This indicates that the cutoff score of 70 is used far more often than interval scores in MR diagnoses.

Much of the reluctance to depart from a cutoff score is due to public scrutiny of how schools allocate their special education resources. In particular, there has been much criticism over the disproportionate number of minority and low-income children in special education, and particularly over the number diagnosed as MR (e.g., Losen & Orfield, 2002). Indeed, prevalence rates of MR for Black youth (16.2/1,000) are almost double that for White (9.8/1,000) and Hispanic (9.0/1,000) youth (Reschly et al., 2002). Racial/ethnic asymmetries of just this sort led to litigation in California (*Larry P. v. Wilson Riles*, 1979, 1986), which forced the state to ban the use of IQ for all MR placement decisions for African Americans, although the use was later reallowed for LD diagnoses (*Crawford v. Honig*, 1992).

Most states, of course, permit the use of IQ tests in special education diagnoses for minority children (e.g., *Parents in Action on Special Education v. Hannon*, 1980), but school psychologists, researchers, and administrators are well aware of, and concerned about, the racial and ethnic imbalances associated with IQ scores. Many of the psychologists we interviewed during our data collection talked about the pressures to use the 70 IQ cutoff score, because of its objectivity, rather than an interval score that encompassed standard errors of measurement on either side of the cutoff. Specifically, they were concerned that by designating for MR placement any children who were above the cutoff score (even if they were within the interval range), they could be subject to litigation similar to *Larry P. v. Wilson Riles*.

In addition to political pressure, there is the pressure of inadequate resources. School psychologists and counselors reported that they did not have enough funding, space, certified teacher's aides, and the like to provide for all the children in their special education programs. The services for MR are greater, and thus more expensive, than are those for LD. Therefore, the cutoff score is often used because it leads to fewer MR diagnoses, and potentially more LD diagnoses, than the interval score.

THERE IS NO "QUICK FIX"

At first glance, the solution to these problems seems simple: Because the Flynn effect is estimated to be an upward creep of approximately 3 IQ points a decade, just subtract 0.3 points for every year the current IQ norms have been in use. Unfortunately, in a follow-up study, we found that the Flynn effect was larger among younger children tested on the older WISCs than among older children (Kanaya, Ceci, & Scullin, 2005). In addition, Sanborn, Truscott, Phelps, and McDougal (2003) found different magnitude Flynn effects among LD students of different IQ levels. Because children are tested and reevaluated throughout their school lives (e.g., triennial evaluations are mandated by IDEA), any individual differences found in the Flynn effect means that a simple quick-fix solution cannot be evenly applied to children of all ages and disability categories and across all sets of norms. Rather, they will likely lead to further inaccurate measures of cognitive abilities and thus continue to lead to misdiagnoses. This situation is exacerbated by the fact that little is known about possible income or ethnic differences in the Flynn effect. Thus, it is unclear whether the IQs of poor or minority children—the population most likely to be given IQ tests—display a Flynn effect comparable to that of middle-class White children.

Due to the individual variations in the Flynn effect, another simple solution is to use the most current norms available. Indeed, the manufacturers of IQ tests (e.g., Psychological Corporation) seem to be renorming their tests more frequently than in the past (the WISC–IV was introduced in 2004, 13 years after the WISC–III, whereas 17 years had passed between the WISC–R and WISC–III). If children were tested exclusively on freshly minted norms, there would be no Flynn effect or need for adjustment.

Replacing old IQ norms with new (more accurate) norms, however, is an expensive and slow process. Faced with a cost of approximately $1,000 per testing kit and the need to purchase many such kits, school districts can adopt a new IQ norm only as quickly as their budgets allow. The speed of adoption depends

[1]MR children constitute 26% of all recipients of supplemental social security insurance, totaling to more than 1 million individuals receiving this benefit and making MR the largest diagnostic category.

as well on the willingness of testers to learn how to administer and score a new version of the test. In fact, we found that almost 2 years after the introduction of the WISC-III, WISC-R was still being used in about half of all testings. Even 4 years later, it had not been completely phased out of school systems. It thus seems exceedingly unlikely that more frequent publication of new norms would lead to their rapid adoption.

Another possible solution is to rely more heavily on measures of adaptive functioning skills than on IQ scores. Most psychologists and counselors *do*, in fact, measure both IQ and adaptive skills. However, although IQ and adaptive behavior scores are often positively correlated—to the point that adaptive behavior is regarded as virtually synonymous with IQ—there is far less agreement on the appropriate cutoff scores for adaptive behavior than there is for measures of intellectual functioning (Bruininks, Woodcock, Weatherman, & Hill, 2000; MacMillan, Gresham, & Siperstein, 1993). Of greater concern is the recent finding of Flynn (2007) regarding the Vineland Adaptive Behavior Scales, which is probably the most widely used standardized, adaptive behavior measure. Comparing the performance of children who were tested on the 1984 and 2005 Vineland norms, Flynn found that children performed better on the more recent norms. Thus, it appears that adaptive behavior, like IQ, is subject to secular changes, complicating its role in the diagnosis of MR.

The development of IQ tests that are created solely for children in the borderline MR range could also provide a better alternative to using a measure that is standardized to the general population. In other words, psychologists would use measures that are standardized specifically for those who are on the cusp of the cutoff and, therefore, could reliably distinguish between children who are separated by as little as 1 IQ point. However, given that the underlying causes of the Flynn effect are still unknown, it is unclear whether such a measure would also be subjected to the rise and fall patterns seen on the current IQ tests.

CONCLUSIONS AND FUTURE DIRECTIONS

The widespread use of cutoff scores in the diagnosis of MR is highly problematic because it assumes that IQ scores are stable over time and that a particular score indexes the same level of cognitive and adaptive functioning from one year to the next. However, because of the Flynn effect and changing test norms, IQ scores, in reality, rise and fall independently of actual cognitive gains or losses. Especially notable is the failure of IQ gains to be matched by gains in adaptive behavior, as might be expected if low intelligence were responsible for problems in everyday living (Flynn, 2007). Given the centrality of the MR cutoff score in many policies—ranging from education and social security to death penalty sentencing—the Flynn effect calls into question the validity of numerous vitally important decisions that affect individuals at all ages and IQ levels.

It may seem disconcerting that we cannot offer viable solutions to resolve this problem. However, the first step toward a solution is to raise awareness of the problem, specifically, the importance of exploring the role of the Flynn effect for children who are at risk for academic failure. Regardless of the outcome to this dilemma, it *is* clear that researchers, policymakers, and practitioners will need to work together to find viable solutions to the impact of the Flynn effect on children. To continue along the current path is to make critical life course decisions that will be detrimental to the lives of at-risk children.

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* (4th ed.). Washington, DC: Author.

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

Bruininks, R. H., Woodcock, R. W., Weatherman, R. F., & Hill, B. K. (2000). *Scales of Independent Behavior–Revised*. Itasca, IL: Riverside.

Code of Federal Regulations, Title 34, Section 300.7 © (10).

Crawford v. Honig, 37 F.3d 485 (1992).

Daley, O., Whaley, S. E., Sigman, M. D., Espinosa, M. P., & Neumann, C. (2003). IQ on the rise: The Flynn effect in rural Kenyan children. *Psychological Science*, *14*, 215–219.

Dickens, W. T., & Flynn, J. R. (2001). Heritability estimates versus large environmental effects: The IQ paradox resolved. *Psychological Review*, *108*, 346–369.

Dillon, S. (2005, July 15). Young students post solid gains in federal tests. *New York Times*, p. 1.

Editorial Board of the APA Division 33. (1996). Definition of mental retardation. In J. W. Jacobson & J. A. Mulick (Eds.), *Manual of diagnosis and professional practice in mental retardation* (pp. 13–47). Washington, DC: American Psychological Association.

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. (2007). *What is Intelligence?: Beyond the Flynn effect*. London: Cambridge University Press.

Kanaya, T., Ceci, S. J., & Scullin, M. H. (2005). Age differences in secular IQ trends: An individual growth modeling approach. *Intelligence*, *33*, 613–621.

Kanaya, T., Scullin, M. H., & Ceci, S. J. (2003). The Flynn effect and U.S. policies: The impact of rising IQ scores on American society via mental retardation diagnoses. *American Psychologist*, *58*, 1–13.

Kavale, K. A., & Forness, S. R. (1992). History, definition and diagnosis. In N. N Singh & I. L. Beale (Eds.). *Learning disabilities: Nature, theory, and treatment*. New York: Springer-Verlag.

Larry P. v. Riles, C-71–2270 FRP. Dist. Ct. Citation (1979, 1986).

Losen, D., & Orfield, G. (2002). *Racial inequality in special education*. Cambridge, MA: Harvard Education Publishing Group.

MacMillan, D., Gresham, F. M., & Siperstein, G. N. (1993) Conceptual and psychometric concerns about the 1992 AAMR

definition of mental retardation. *American Journal on Mental Retardation*, *98*, 325–335.

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

Neisser, U., Boodoo, G., Bouchard, T. J., Boykin, A. W., Brody, N., Ceci, S. J., et al. (1996). Intelligence: Knowns and unknowns. *American Psychologist*, *51*, 77–101.

Parents in Action on Special Education v. Hannon, 506 F.Supp 831 (1980).

Reschly, D. J., Myers, T. G., & Hartel, C. R. (Eds.). (2002). *Mental retardation: Determining eligibility for social security benefits*. Washington, DC: National Academies Press.

Sanborn, K. J., Truscott, S. D., Phelps, L., & McDougal, J. L. (2003). Does the Flynn effect differ by IQ level in samples of students classified as learning disabled? *Journal of Psychoeducational Assessment*, *21*, 145–159.

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.

Tuerk, P. W. (2005). Research in the high-stakes era: Achievement, resources and No Child Left Behind. *Psychological Science*, *16*, 419–425.

U.S. Department of Education. (2005). *Twenty-fifth annual report to Congress on the implementation of the Individuals with Disabilities Education Act*. Washington, DC: Author.

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

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

World Health Organization. (1996). *ICD-10 guide for mental retardation*. Geneva, Switzerland: Author.