Stalking the Roving IQ Score Cutoff: A Commentary on Kanaya and Ceci (2007)

Keith F. Widaman

University of California at Davis

ABSTRACT—***T. Kanaya and S. J. Ceci (2007) reviewed research on perturbations in rates of mental retardation diagnoses associated with the Flynn effect. They argued that their results call into question the use of IQ cutoff scores in high-stakes decisions but acknowledged that no viable alternatives to cutoff scores could be strongly endorsed. Further, Kanaya and Ceci argued that Flynn-effect adjustments to IQ scores were problematic, a position that should be open to debate. Further research and clear recommendations for practice are certainly needed on this important problem.***

KEYWORDS—*mental retardation; intelligence; IQ; Flynn effect; high-stakes decisions; cutoff scores*

Determining whether an individual meets diagnostic criteria for any particular condition can be fraught with difficulties. These difficulties are probably no more contentious or momentous than in the diagnosis of mental retardation (MR) in children during the elementary or middle school years, as Kanaya and Ceci (2007) outlined. When deciding to identify a child as having MR, the school psychologist must confirm that the child meets the diagnostic criteria of (a) significantly subaverage general intellectual functioning, (b) concurrent deficits in adaptive behavior, and (c) manifestation of these behavioral markers prior to the age of 18 (American Association on Mental Retardation, 2002; American Psychiatric Association, 2000) or 22 years (American Psychological Association; see Jacobson & Mulick, 1996; see Reschly, Myers, & Hartel, 2002). The psychologist may also weigh hoped-for benefits (e.g., more appropriate instructional modalities and testing conditions) against likely costs (e.g., the inevitable social distancing and taunts in classroom and schoolyard) that accompany such a label, although these considerations should not affect the diagnostic process or outcome.

Correspondence concerning this article should be addressed to Keith F. Widaman, Department of Psychology, University of California, One Shields Avenue, Davis, CA 95616-8686; e-mail: kfwidaman@ucdavis.edu.

As Kanaya and Ceci (2007) pointed out, this diagnostic task is made substantially more difficult by the well-known Flynn effect (Flynn, 1984, 1987) the steady increase in mean IQ performance of approximately 3 points per decade by populations around the world. Kanaya, Scullin, and Ceci (2003) provided a major service to the field by assembling a large sample (N = 8,944) of special education assessments to investigate possible Flynn effects near the traditional cutoff score of 70. Kanaya et al. reported that children near the cutoff tended to score lower on a newer test than on an older test (e.g., lower on the Wechsler Intelligence Scale for Children, Third Edition [WISC–III] than on the Wechsler Intelligence Scale for Children–Revised [WISC–R]) at just about the rate expected by Flynn-effect estimates of mean changes. Importantly, the higher scores associated with the outdated norms led to a substantial decrease in the proportion of students who obtained IQ scores of 70 or below and therefore could be diagnosed as having MR. And, as predicted, introduction of a newly normed test led to a much higher proportion of students obtaining scores of 70 or below.

One might quibble about certain statistics cited by Kanaya and Ceci (2007). For example, they stated several times that about 2.27% of the population falls at or below 2.0 *SD* below the mean. This statement is correct for the theoretical standard normal distribution, which is both normal and continuous. But the IQ distribution is neither; rather, it is nonnormal (with a "bump" in the lower tail; cf. Burack, 1990; Zigler, 1967) and discrete. It is easy to show that the population proportion with scores of 70 or lower in the distribution of IQ scores, even excluding the bump in the lower tail, is approximately 2.46% of the population, not 2.27%.[1] This difference in percentages occurs even though an IQ score of 70 and a standard score of −2.0 fall exactly 2.0 *SD* below the means of their respective distributions. Although the difference between 2.27% and 2.46% seems rather small, over 8% more individuals (because $2.46 \div 2.27 = 1.084$) meet the "70 or below" criterion on the IQ distribution than meet the "−2.0 or below" criterion on the standard normal distribution. Moreover, this increase is due solely to rounding continuous scores from the normal distribution into discrete scores in the IQ distribution. If the proportion falling in the lower tail of a distribution can be affected in nontrivial fashion merely by rounding of scores, we should not be surprised to find that the proportion of students falling at or below the IQ cutoff score of 70 would be affected greatly if the Flynn effect influenced scores near the cutoff, and Kanaya et al. (2003) demonstrated clear and sizable effects of this sort.

The simplest way to adapt to the secular trend in IQ scores is to multiply the number of years since a test was normed by 0.3 and then to subtract this value from the obtained IQ score. The resulting "IQ score adjusted for the Flynn effect" could then be treated as the best estimate of an individual's IQ. Kanaya and Ceci (2007) rejected this approach because their research (Kanaya, Ceci, & Scullin, 2005) and research by others (Sanborn, Truscott, Phelps, & McDougal, 2003) found interaction, or differential, effects reflecting Flynn-effect fluctuations that varied across age or IQ levels. But a careful examination of Kanaya et al. (2005) suggests the need for a more cautious evaluation. First, although sample sizes appeared large (mostly between 400 and 600), inclusion of 11 or 12 age level means that an average of only 40–50 children per age level were included in each analysis, sample sizes that are likely too small to justify major policy decisions. Second, the only significant interactions arose in comparisons involving the WISC–R, an outdated test. Neither comparison using the WISC–III had a significant interaction (interaction effects were approximately zero), suggesting that newer and more well-normed tests may not fall prey to the problematic interaction, even though they continue to exhibit Flynn-effect fluctuations. Finally, even if an interaction were present, it may be better and fairer for all concerned to employ an adjustment that is presumably a little too much for some and a little too little for others than to employ no adjustment at all.

In rejecting statistical approaches to adjusting IQ scores based on outdated norms, Kanaya and Ceci (2007) also argued that potential differential Flynn effects as a function of ethnic or income status might make a single "one-size-fits-all" adjustment inequitable to members of certain groups. Then, acknowledging the time and monetary expenses involved in adoption of new test norms, Kanaya and Ceci argued that the use of tests with "freshly minted norms" would ensure that no Flynn effect would be present, hence no need for statistical adjustments. However, normed IQ test scores are not adjusted for income or ethnic differences, so any differential Flynn effects associated with ethnic or income status would be submerged in the norming process. The use of a newly normed test is clearly preferable to the use of statistical adjustments of IQ scores based on outdated norms. But both of these approaches represent one-size-fits-all adjustments to scoring algorithms and would likely have comparably inequitable effects on groups that have Flynn effects differing from those exhibited by middle-class White children.

Some issues discussed by Kanaya and Ceci (2007) are notable largely as spurs to additional action, not as criticisms of IQ score cutoffs. If parents, school personnel, and others wrongly interpret score fluctuations due to differently aged norms as "true" changes, we experts in the field must do a better job of informing stakeholders that their interpretations are incorrect. We should also encourage school psychologists to use the best and most recently normed tests—despite the continuing expense—so that Flynn-effect adjustments need not be made to IQ scores. Kanaya and Ceci's suggestion of developing special intelligence tests that are geared for maximal assessment precision around the cutoff score of 70 is an interesting one, although precision to "reliably distinguish between children who are separated by as little as 1 IQ point" is probably beyond reach for the foreseeable future.

Other issues appear to be tangential or problematic. Kanaya and Ceci (2007) argued that MR diagnosis might better "rely more heavily on measures of adaptive functioning skills than on IQ scores." Let us not forget that MR is, first and foremost, a phenomenon involving cognitive abilities, manifested by a lower rate of development of these abilities. Intelligence tests assess these cognitive abilities, adaptive skill measures do not, and the mere presence of positive correlations between the two types of measures does not mean they are "virtually synonymous," as Kanaya and Ceci implied many professionals believe to be the case.

Some jobs are intrinsically difficult and inexact, a description that certainly applies to diagnosing MR when the person being assessed has a true IQ score near 70. The standard errors of measurement (SEMs) of our best intelligence tests range between 2.5 and 4 points, and the inevitable nonzero SEM ensures that a person with a true IQ near 70 will obtain IQ scores that sometimes fall above the cutoff of 70 and sometimes fall below. The Kanaya and Ceci (2007) research does not call into question the use of cutoff scores. Rather, their research calls into question the use of IQ scores based on outdated norms because we know—thanks to Kanaya et al. (2003)—that the upward creep of scores due to the Flynn effect will lead to ever fewer proportions of students meeting the cutoff of 70. If Flynn-effect adjustments can dampen improper score fluctuations due to aging norms and thereby smooth out the proportions of students receiving IQs of 70 or below, I think such

[1]Interested readers should contact the author to obtain a description of the basis for differing proportions of the population falling at or below 2.0 *SD* below the mean in the standard normal distribution and the distribution of IQ scores.

adjustments should be used; in fact, it would be inappropriate not to do so. This "quick fix" is admittedly imperfect and should be monitored by continuing research to ensure that it has no unintended negative consequences. Nonetheless, use of Flynn-effect adjustments, however imperfect, is likely to have fewer negative consequences than would the failure to use such adjustments. Support for this position in no way undermines the call by Kanaya and Ceci for further concerted research efforts on the vexing problem of roving cutoff scores.

Acknowledgments—This research was partially supported by grants from the National Institute of Child Health and Human Development, the National Institute on Drug Abuse, and the National Institute of Mental Health (HD047573, HD051746, and MH051361).

REFERENCES

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

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., Text Revision). Washington, DC: Author.

Burack, J. A. (1990). Differentiating mental retardation: The two-group approach and beyond. In R. M. Hodapp, J. A. Burack, & E. Zigler (Eds.), *Issues in the developmental approach to mental retardation* (pp. 27–48). Cambridge, UK: Cambridge University 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.

Jacobson, J. W., & Mulick, J. A. (Eds). (1996). *Manual of diagnosis and professional practice in mental retardation*. Washington, DC: American Psychological Association.

Kanaya, T., & Ceci, S. J. (2007). Are all IQ scores created equal? The differential costs of IQ cutoff scores for at-risk children. *Child Development Perspectives*, *1*(1), 52–56.

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.

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.

Zigler, E. (1967). Familial mental retardation: A continuing dilemma. *Science*, *155*, 292–298.