Journal of Psychoeducational Assessment

http://jpa.sagepub.com/

"Apples and Oranges Are Both Round": Furthering the Discussion on the Flynn Effect

Stephen J. Ceci and Tomoe Kanaya

Journal of Psychoeducational Assessment 2010 28: 441 originally published online 14 June 2010
DOI: 10.1177/0734282910373339

The online version of this article can be found at:
http://jpa.sagepub.com/content/28/5/441

Published by:
SAGE
http://www.sagepublications.com

Journal of Psychoeducational Assessment
28(5) 441–447

DOI: 10.1177/0734282910373339
http://jpa.sagepub.com
SAGE

"Apples and Oranges Are Both Round": Furthering the Discussion on the Flynn Effect

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

Abstract

While the magnitude of the Flynn effect is well established (approximately 3 points a decade on the Wechsler scales), the causes behind it are still unknown and hotly debated. Kaufman argues that, because of the administrative and scoring changes that occurred with the introduction of the Wechsler Intelligence Scale for Children–Revised, Flynn's interpretation of the effect is not appropriate. Although agreeing that these changes account for some aspects of rising IQ, this study questions the impact of these administrative/scoring changes to account for most of the impact, given the heavy documentation of the Flynn effect on multiple IQ tests and norms over time and around the world. The authors also add to the discussion led by Zhou, Zhu, and Weiss by stressing the importance of examining the role of individual differences within the Flynn effect to understand fully the exact nuances and cause(s) of it.

Keywords

Flynn effect, IQ, WISC, subtests, individual differences

The Flynn effect (FE) has been documented in more than 29 countries around the world (Flynn, 2007), and Zhou, Zhu and Weiss (2010) add to this growing body of knowledge by providing data from the standardization samples from multiple Wechsler scales. Although few question the existence of the effect, the underlying causes and interpretation of it are still poorly understood and debated (e.g., Neisser, 1998). Kaufman (2010) challenges Flynn's own interpretation of the effect in this issue that the gains are due to societal changes, where our daily routines have become more imbued with fluid reasoning skills compared with the daily routines of our grandparents and great-grandparents. More specifically, Kaufman questions the overall finding of higher gains on tests of fluid abilities, such as the Similarities and Comprehension subtests of the Wechsler Intelligence Scale for Children (WISC) norms and the Ravens Progressive Matrices, compared with tests of crystallized abilities such as vocabulary.

[1]Cornell University, Ithaca, NY, USA
[2]Claremont McKenna College, Claremont, CA, USA

Corresponding Author:
Stephen J. Ceci, Department of Human Development, Cornell University, M. Van Rensselaer Hall, Room G80, Ithaca, NY 14853, USA
Email: stevececi@cornell.edu

Similarities and Subtest Instructions

Kaufman spends a considerable amount of time on the changes in the instructions that were introduced with the WISC–Revised (WISC-R). He argues that these changes account for a portion of the documented gains. We agree that the instructions provided to an individual can have a significant impact on his or her IQ. Indeed, one of us has had firsthand experience with this phenomenon. When testing school children on the WISC-R, it became obvious that some of the children were being penalized on Similarities because they had no idea that the conceptual answers were awarded 2 points whereas their own perceptual or thematic answers were given only 1 point. When these children were administered questions that were supposed to be discontinued (because of the requisite number of consecutive incorrect answers), many of them were not only able to answer the questions correctly, but they started providing 2-point answers.

At first, it did not seem possible that children would perform *better* on test items that they were not allowed to answer due to their previous, comparatively, poorer performance on easier items. But the reason was clear. On the earlier questions, some of these children—almost always males—leapt to their answers immediately after hearing the question, and they answered by providing the most superficial basis of the pairs' similarity (e.g., "They're both round" in response to being queried "How are apples and oranges alike?"). As the basis of items' similarity became more opaque, however, it was not as easy to leap in with a superficial answer. So, long after they failed to answer easier questions correctly, they were asked questions such as "How are scissors and a pot alike?" They could not leap to an answer. Rather, they really had to stop and think, at which point, they'd say "They're both made out of metal," a 2-point answer.

From this experience, we concluded that the scoring was almost unfair to some children, particularly those with deficits in inhibitory control and impulsivity. If these children had been told the basis of the scoring procedure, they may have gotten higher scale scores all along. Therefore, we have no doubt that the substantive changes in the administration and scoring of the WISC-R compared with the WISC carefully outlined by Kaufman could account for a significant portion of the documented gains. Furthermore, we also agree with Kaufman that the elimination of five year olds must be taken into account when interpreting the magnitude of the gains given our previous finding of age-related differences on the FE on the WISC (Kanaya, Ceci, & Scullin, 2005).

An Increasingly Fluid World

The administrative and scoring changes that occurred with the WISC-R, however, do not negate the overwhelming amount of evidence that shows higher gains on measure of fluid ability compared to measures of crystallized fluid ability. For example, the fluid subtests, including Similarities, have continued to show higher gains compared to the crystallized subtests on more recent WISC norms, including the WISC-III and WISC-IV (Flynn, 2007; Flynn & Weiss, 2007). (Although Kaufman is correct in pointing out that the gains have subsided considerably on the Similarities subtest.) Another example is Norway, which experienced an increase of almost 20 IQ points within a generation on the Ravens Progressive Matrices, whereas the estimated gains in New Zealand on the Otis Test are less than half of those in Norway (Flynn, 1987). Indeed, with little (if any), exception, measures that load higher on fluid abilities show a larger FE than measures that load higher on crystallized abilities.

Studies have consistently shown the highest gains occur on the Ravens Progressive Matrices. Kaufman (2010) expresses doubt over the magnitude of these gains, pointing out that the Ravens is more prone to practice effects than other IQ measures because (a) unlike the WISC norms, it does not need to be administered by a trained professional and (b) the increase in Ravens-like stimuli (e.g., puzzle books, placemats at restaurants, cereal boxes, etc.) that are now readily

available to everyone. It is difficult to imagine, however, that practice effects from frequent test administrations would be large enough to reduce the estimated Ravens gains to the point where they are equal to the gains found on crystallized measures. After all, Zhou et al.'s analyses reveal a very small effect size for testing order (partial $\eta^2 = 0.07$) in their counterbalanced design, suggesting that practice effects have a small, potentially negligent, role in the FE.

Through the use of computer simulation models, Carpenter, Just, and Shell (1990) found that abstract reasoning and working memory capacity skills distinguished high performers from low performers on the Ravens, as both the reasoning/working memory tasks and the Ravens test compete for limited attentional resources. Similarly, Jaeggi, Buschkuehl, Jonides, and Perrig (2008) have demonstrated that training on a demanding working memory task transferred to gains on fluid intelligence tasks. Thus, there is some evidence that practice can elevate scores on measures of fluid intelligence such as the Ravens, though the degree of practice in these studies is very demanding and unlikely to characterize the increased testings by nonclinicians or increases in brief, informal exposure to fluid tasks that children and adults encounter in their daily lives.

The increase in Ravens-like stimuli in our everyday lives can be used to support both Kaufman and Flynn. Are the gains artificially high because of practice effects from readily available abstract stimuli (Kaufman's view)? Or, does the increase in abstract stimuli reflect the ever-increasing demands in fluid intelligence that were not present in the daily lives of our grandparents and great grandparents (the Dickens and Flynn model)? There seems to be more evidence for the latter. Schooler (1998) and Greenfield (1998) make compelling arguments in support of a "cultural evolution" that has increased the level of environmental complexity and demands for fluid reasoning skills throughout the world. Williams (1998) documents many home and school related changes that have occurred over the last century, including the increase in abstract toys geared for children and the "watering down" of middle school reading textbooks, that can account for the fluid-versus-crystallized trend within the FE.

Kaufman (2010) also argues that the Dickens and Flynn model cannot account for IQ gains in young children, evidenced by the results from the preschoolers tested on the WPPSI norms in Zhou et al.'s (2010) study. The developmental psychology literature, however, has clearly and consistently shown that small environmental changes can lead to long-term cognitive growth in children even younger than the Wechsler Preschool and Primary Scale of Intelligence (WPPSI) testing age range. For example, in DeCasper and Spence's (1986) seminal study, newborns learned and remembered the sound patterns they were exposed to as fetuses when their mothers read to them in utero. Furthermore, the types of toys that are marketed to parents of newborns have increased in volume and complexity. Indeed, "educational toys," which are advertised as promoting learning and reasoning skills, have become a multibillion dollar industry and many of these toys are geared toward preschool-aged children. In other words, today's newborns are also living a life with higher fluid demands than those from a generation ago. Although we are not advocating that Flynn's interpretation of the effect is definitively correct, we *are* suggesting that Kaufman's objections may not be strong enough to invalidate it.

Individual Differences in the Flynn Effect: The "Black Box" Is Bigger Than We Think

An important aspect of the model for intelligence proposed by Dickens and Flynn (2001) is the idea of the "individual multiplier." The individual multiplier explains how "a small genetic advantage on the part of an individual captures powerful environmental forces" (Flynn, 2007, p. 88). These environmental forces, in turn, multiply and lead to "massive" IQ gains within a relatively short amount of time. The study of individual differences within the FE could help determine which characteristics serve as individual multipliers and which do not.

To date, the amount of research dedicated to uncovering individual differences in the FE is sparse relative to the number of studies that have been committed to uncovering its existence in different countries and on different measures of intelligence. As we mentioned earlier, our previous work (Kanaya et al., 2005) uncovered age-related differences, where younger children experienced larger gains on the WISC and WISC-R, but this difference disappeared on the WISC-III. Sanborn, Truscott, and Phelps (2003) found ability level differences, where children with higher IQs experienced larger gains than children with lower IQs. In both of these studies, however, the IQ data were retrieved from special education evaluations. Now, thanks to Zhou et al. (2010) these trends have been replicated within the general population. They conclude their article with suggestions for future research that will help uncover the complexity of the FE. We would like to add the following suggestions to their list.

"High Stakes" Populations

When it comes to IQ, schoolchildren are one of the most heavily tested populations in the United States. Because of the Individuals with Disabilities Education Act (Public Law 94-142, 1975), an overwhelming majority of children evaluated for special education services are administered an IQ test. If they qualify for such services, they are usually tested again at least every 3 years for the required reevaluations. These mandatory, longitudinal testings make it increasingly likely for the FE to have an impact on a child's special education diagnosis, as it increases the likelihood that (a) a child's IQ will increase over time if the same norm is used repeatedly in the reevaluations or (b) a child's IQ will fall dramatically if a new norm is used during one of the reevaluations. Kanaya, Scullin, and Ceci (2003) found that the number of children who were diagnosed with mental retardation (MR) nearly tripled on the introduction of the WISC-III as more and more children obtained an IQ of 70 points or less on the newly introduced, harder norm. Therefore, IQ plays a major role in determining the educational experiences and opportunities provided to a child (and the costs incurred by the schools to implement these special educational services) throughout his or her school years. As such, special consideration should be placed on high-stakes populations, not just the general population, within the FE research community.

A separate, but related, population that should also be explored is ethnic minorities. The Black–White test gap of approximately one standard deviation is well documented (e.g., Neisser et al., 1996) as is the overrepresentation of ethnic minorities, particularly African American boys and low-income children, in MR and other special education classrooms (Reschly, Myers, & Hartel, 2002). The former has been the subject of several high-profile legal battles, such as *Larry P vs. Wilson Riles* (1979). Although Sanborn et al. (2003) found that race was not a significant predictor of the FE, their sample size was limited and all of their participants were from the same geographical region (upstate New York). Zhou et al. also found that race was not a significant predictor, but more research must be conducted before we know, definitively, that the FE does not vary by race or ethnicity. Specifically, future research should examine if there is an interaction effect between ethnicity and ability level and/or if the FE has a different impact on special education diagnoses based on the race/ethnicity of the child.

Although special education diagnoses and services pertain to school-aged children, it is worth noting that the consequences of such diagnoses can continue well beyond the school years. The social stigma associated with receiving such a label can have a significant, lifelong, negative psychosocial impact (Mercer, 1973), and children who qualify for these diagnoses, including MR, learning disabled, and emotionally disturbed, can also qualify for social security disability benefits throughout their lives. Without question, the stakes are highest for defendants in capital murder trials. Because of *Atkins vs. Virginia* (2002), individuals who are diagnosed with MR are not

eligible for the death penalty. Therefore, in some extreme cases, the FE is literally a matter of life or death. While many have written about the implications of the FE for capital punishment cases (e.g., Flynn, 2006, Gresham, 2009, and several authors in this issue) there is still little, if any, research devoted to measuring the magnitude of the FE within this specific population of imprisoned offenders.

Across the Lifespan of the Wechsler Scales

The Wechsler norms are divided into three separate scales according to age range: (a) the WPPSI for preschoolers, (b) the WISC for children between the ages of 6 and 16 years, and (c) the Wechsler Adult Intelligence Scale (WAIS) for individuals who are 16 years or older. To date, most of the research on the FE focused on IQ trends across different norms within each scale. In other words, they have compared performance on the WISC with the WISC-R or performance on the WAIS-R with the WAIS-III. The renorming cycles of these scales, however, do not overlap, thereby potentially causing even further complications when comparing IQ scores across and within cohorts. Very few studies have examined what happens when an individual "grows into" the next scale. For example, what happens when a cohort of 15-year-olds who have been tested on the current WISC norm their whole lives are suddenly tested on the contemporary WAIS norm the following year?

Even with high measurement invariance within each scale, it is an empirical question as to whether there will be a FE across the scales when moving from a more obsolete norm on a younger scale to a newer norm on a scale for older individuals. This is a situation that sometimes occurs with capital offenders whose performance on a contemporary adult scale given to them after arrest and conviction is compared with performance on a scale for children that was administered many years earlier when they were in school. The reverse is likely, though possible—a newer norm administered during their school years is compared with an obsolete norm administered on arrest and conviction. Relatedly, little research exists comparing FEs across different tests of general intelligence, such as an individual who is tested on the Ravens in childhood, but on the WAIS as an adult.

A notable exception is Spitz (1989), who examined the WISC-R to WAIS-R transition among individuals in the MR range and found a reverse FE. But, it is unclear if this reversal was because of (a) scores from a WISC norm that is "young" within its norming cycle (and, hence, yields lower scores) were compared with the scores from an older/inflated WAIS norm or (b) an actual reversal of the FE that may have only been experienced during these specific testing years, and/or only by individuals in the MR range of the distribution. Either scenario has important practical and theoretical implications, but more research must be done before we can determine which interpretation is correct.

Personality Characteristics

The experience with the Similarities subtest described earlier also points to the need to explore the impact of personality characteristics on the FE. From its beginnings, the Wechsler scales were thought to be influenced by personality as well as cognitive processing. In our experience, impulsive children leapt to the first similar dimension of the stimuli questions, without pondering the deeper, more conceptual bases. These children are unlikely to do well in school not because they lack the conceptual knowledge (just give them a bin of items and ask them to pick out the fruits, musical instruments, metal objects, and they can do so without error), but because the same reflection and response inhibition that is rewarded in the scoring is also rewarded in traditional classrooms. But, it will also be important to determine if they experience the FE to the same magnitude as others with comparable IQ performance.

Fluid Versus Crystallized Intelligence: Not Just Verbal Versus Performance

Zhou et al. (2010) used the Verbal and Performance IQs of the Wechsler scales and interpreted them as measures of crystallized and fluid intelligence, respectively. Although it is true that the Verbal IQ loads higher on crystallized intelligence and Performance IQ on fluid intelligence, it is important to take into account that several of the Verbal subtests measure fluid abilities (e.g., Similarities) and several of the Performance subtests measure crystallized abilities (e.g., Matrix Reasoning). Therefore, as both Flynn and Kaufman have done, we advocate analyses at the subtest level when contributing to the discussion on the FE on fluid versus crystallized IQ.

Conclusions

It is clear that we still have a long way to go before we understand the exact nature and magnitude of the FE for all types of individuals, across time, and across tests. Until we do, discussions regarding its cause(s) and interpretation may be futile, as will solutions for trying to "fix" it. We have pointed out elsewhere (Kanaya & Ceci, 2007a, 2007b) that it is not appropriate to merely subtract 0.3 points for every year that a norm has aged until we know that everyone experiences the same gains on the same subtests and at the same time. The current findings by Kaufman (2010) and Zhou et al. (2010) show that we are far from such knowledge and that much work needs to be done before the "black box" is fully illuminated.

Declaration of Conflicting Interests

The author(s) declared no conflicts of interest with respect to the authorship and/or publication of this article.

Funding

The author(s) received no financial support for the research and/or authorship of this article.

References

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

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the raven Progressive Matrices Test. *Psychological Review*, *97*, 404-431. doi:10.1037/0033-295X.97.3.404

DeCasper, A. J., & Spence, M. J. (1986). Prenatal maternal speech influences newborns' perception of speech sounds. *Infant Behavior & Development*, *9*, 133-150. doi:10.1016/0163-6383(86)90025-1

Dickens, W. T., & Flynn, J. R. (2001). Heritability estimates versus large environmental effects: The IQ paradox resolved. *Psychological Review*, *108*, 346-369. doi:10.1037/0033-295X.108.2.346

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101*, 171-191. doi:10.1037/0033-2909.101.2.171

Flynn, J. R. (2006). Tethering the elephant: Capital cases, IQ, and the Flynn effect. *Psychology, Public Policy, and Law*, *12*, 170-189. doi:10.1037/1076-8971.12.2.170

Flynn, J. R. (2007). *What is intelligence? Beyond the Flynn effect*. New York, NY: Cambridge University Press.

Flynn, J. R., & Weiss, L. G. (2007). American IQ gains from 1932 to 2002: The WISC subtests and educational progress. *International Journal of Testing*, *7*, 209-224. doi:10.1080/15305050701193587

Greenfield, P. M. (1998). The cultural evolution of IQ. In U. Neisser (Ed.), *The rising curve: Long-term gains in IQ and related measures* (pp. 81-124). Washington, DC: American Psychological Association.

Gresham, F. M. (2009). Interpretation of intelligence test scores in Atkins cases: Conceptual and psychometric issues. *Applied Neuropsychology*, *16*, 91-97. doi:10.1080/09084280902864329.

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences U S A*, 105, 6829-6833.

Kanaya, T., & Ceci, S. J. (2007a). Are all IQ scores created equal? The differential costs of IQ cut-off scores for at-risk children. *Child Development Perspectives*, *1*, 52-56. doi:10.1111/j.1750-8606.2007.00010.x

Kanaya, T., & Ceci, S. J. (2007b). Mental retardation diagnosis and the Flynn effect: General intelligence, adaptive behavior, and context. *Child Development Perspectives*, *1*, 62-63. doi:10.1111/j.1750-8606.2007.00013.x

Kanaya, T., Ceci, S. J., & Scullin, M. H. (2005). Age differences with secular IQ trends: An individual growth modeling approach. *Intelligence*, *33*, 613-621. doi:10.1016/j.intell.2005.08.001

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. doi:10.1037/0003-066X.58.10.778

Kaufman, A. S. (2010). "In what way are apples and oranges alike? A critique of Flynn's interpretation of the Flynn effect. *Journal of Psychoeducational Assessment, 28*, 382-398.

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

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

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., Jr., Boykin, A. W., Brody, N., Ceci, S. J., . . . Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologists*, *51*, 77-101. doi:10.1037/0003-066X.51.2.77

Public Law 94-142. (1975). *Education for All Handicapped Children Act*. U.S. Congress.

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. (2003). Does the Flynn effect differ by IQ level in samples of students classified as learning disabled? *Journal of Psychoeducational Assessment*, *21*, 145-159. doi:10.1177/073428290302100203

Schooler, C. (1998). Environmental complexity and the Flynn effect. In U. Neisser (Ed.), *The rising curve: Long-term gains in IQ and related measures* (pp. 67-80). Washington, DC: American Psychological Association.

Spitz, H. H. (1989). Variations in Wechsler Interscale IQ disparities of different levels of IQ. *Intelligence*, *13*, 157-167. doi:10.1016/0160-2896(89)90014-7

Williams, W. M. (1998). Are we raising smarter children today? School- and home-related influences on IQ. In U. Neisser (Ed.), *The rising curve: Long-term gains in IQ and related measures* (pp. 125-154). Washington, DC: American Psychological Association.

Zhou, X., Zhou, J., & Weiss, L. G. (2010). Peeking inside the "black box" of the Flynn effect: Evidence from three Wechsler instruments. *Journal of Psychoeducational Assessment, 28*, 399-411.