Journal of Psychoeducational Assessment
XX(X) 1–12

DOI: 10.1177/0734282910370139
http://jpa.sagepub.com
SAGE

The Flynn Effect in the WISC Subtests Among School Children Tested for Special Education Services

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

Abstract

The Flynn effect, a secular rise in IQ seen throughout the world, was examined on the WISC-R and WISC-III subtests in a longitudinal sample of more than 2,500 school children who were tested between 1974 and 2002. Multivariate analysis of variance and multiple regression analyses revealed that all the subtests experienced significant decreases in scores on the introduction of the WISC-III, as expected because of the Flynn effect, with the exception of Information and Digit Span. (Mazes was not included in the analyses because of a limited sample size.) On Picture Arrangement and Coding, however, children who were repeatedly tested on the WISC-III also experienced significant decreases compared with children who were repeatedly tested on the WISC-R. These findings add to the growing literature comparing the magnitude of the Flynn effect on crystallized versus fluid measures. Implications for special education testing and the current WISC-IV are discussed.

Keywords

IQ, WISC subtests, Flynn effect, school children

Now documented in more than 29 countries throughout the world, the Flynn effect refers to the systematic and gradual rise in IQ scores over time (e.g., Flynn, 1984, 1987; Rodgers & Wänström, 2007; te Nijenhuis & van der Flier, 2007). Because of the Flynn effect, IQ tests become easier over time, and people are able to perform progressively better on the tests as the norms age. This rise is difficult to detect on a year-to-year basis, because it is only, on average, 0.3 IQ points. Rather, it is apparent after a decade or so, when an IQ test produces a mean score of 103 or higher rather than 100. Because of this, the IQ test companies publish new, or revised, norms. The new norm is harder and resets the (inflated) mean back to 100 points. Therefore, the Flynn effect is most prominent on the introduction of a new norm, whereon scores experience a sudden and dramatic drop.

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

Corresponding Author:
Tomoe Kanaya, Claremont McKenna College, 850 Columbia Avenue, Claremont, CA 91711, USA
Email: tkanaya@cmc.edu

An overwhelming majority of the research on the Flynn effect has focused on adults in the average range of the IQ distribution. Notably less research has focused on individuals with cognitive disabilities, and still less research has focused on children with disabilities. This is somewhat surprising because the Flynn effect would seem to have its most direct and immediate impact on school children who are being tested for special education services, one of the most heavily tested populations in the United States. Indeed, Kanaya, Scullin, and Ceci (2003) found that children at the cusp of the Mental Retardation cutoff IQ score of 70 when tested on an old norm (the Wechsler Intelligence Scale for Children–Revised [WISC-R]), subsequently lose approximately 5 IQ points when retested on a newer, harder norm (the Wechsler Intelligence Scale for Children–Third edition [WISC-III]) during their federally mandated reevaluations. This drop resulted in a significant increase in Mental Retardation (MR) diagnoses at the time of retesting on the newer norm, as many of the students, whose scores on the older norm were slightly higher than the IQ 70 cutoff for classification as MR, subsequently plummeted to less than 70. In contrast to children who were tested on the new, harder norm, children who were retested on the same older norm experienced a small, statistically insignificant rise of less than one point, and no significant changes in their MR diagnoses occurred.

Children diagnosed with Learning Disabilities (LD) also experience IQ declines on the introduction of a new norm (Truscott & Frank, 2001). An LD diagnosis, however, requires an IQ within the average range of the distribution. Therefore, the sudden drop in scores on the introduction of the newer, harder WISC-III resulted in a significant decrease in LD diagnoses and related special education services.

The Flynn effect is highest on tests of fluid abilities, which measure abstract reasoning and on-the-spot problem solving (e.g., matrices, puzzles, figural relations, similarities, and mazes) compared with tests of crystallized abilities, which measure accumulated knowledge, such as vocabulary, information, and to some degree arithmetic, which has both fluid and crystallized aspects (Flynn, 1987). For example, Norway experienced enormous gains, almost 20 IQ points within a generation on the Ravens Progressive Matrices, the paradigmatic test of fluid abilities, whereas the estimated gains in New Zealand on the Otis Test, which primarily taps crystallized ability, are less than half of those in Norway (Flynn, 2007). Between 1947 and 2002, IQ gains on the WISC norms were almost twice as high on Performance IQ (mostly but not exclusively fluid measures) compared with Verbal IQ (mostly but not exclusively crystallized measures; Flynn, 2007).

As noted above, although many of the WISC Performance subtests measure fluid ability, not all of them do. Indeed, the WISC subtests have been found to tap into at least four factors: (a) Verbal Comprehension (primarily but not entirely loading on crystallized intelligence), which consists of the Information, Similarities, Vocabulary, and Comprehension subtests; (b) Perceptual Organization (analogous to fluid intelligence), which consist of Picture Completion, Picture Arrangement, Block Design, and Object Assembly; (c) Freedom from Distraction, which consist of Arithmetic and Digit Span; and (d) Processing Speed, which consist of Coding and Symbol Search (Keith & Witta, 1997; Watkins & Kush, 2002). Likewise, not all of the WISC Verbal subtests are measures of crystallized ability. Furthermore, fluid and crystallized abilities are partners in a reciprocal developmental process in which each fosters the development of the other. Cattell (1971) theorized that fluid cognitive ability acts as a precursor to crystallized ability because the former fosters the acquisition of crystallized knowledge. Recently, evidence has accumulated that at least some fluid abilities such as memory monitoring and metacognition are fostered by the acquisition of crystallized knowledge (Ceci, Fitneva, & Williams, in press). Thus, subtests that primarily load on one type of ability are likely to affect the other. Hence, low fluid ability may lower subtest scores that load on crystallized ability and vice versa.

Few have looked at the Flynn effect at the subtest level within the WISC. Therefore, the purpose of this study is to examine test–retest scores on the WISC subtests in the population that is

most influenced by IQ, school children who are tested for special education programs. Of particular interest is whether the Flynn effect that has been amply documented with the general population will be replicated among children in special education, or whether they exhibit a different pattern of subtest fluctuations.

Method

Participants and Procedures

IQ data from more than 11,000 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) and neighborhood types (rural, urban, suburban). Data included students' gender, age, testing date, IQ scores, test–retest norms used, and special education placement recommendations. Testing dates on this data set ranged from 1960 to 2004. 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.

As a result of collecting longitudinal data on children who were tested during the targeted time frame, the data set 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 retested on the WISC-III).

The sample was divided into three groups based on their testing combination: (a) RR, children who were tested on the WISC-R both times, (b) R3, children who were tested on the WISC-R the first time and the WISC-III the second time, and (c) 33, children who were tested on the WISC-III both times (see Table 1 for descriptive information). Many of the children in the data set were tested three or more times on one of the WISC norms. In these cases, only the first (earliest) two WISC tests were included in our sample. Testing dates in this sample ranged from 1974 to 2000. Most school districts did not document the race of the child in his or her special education evaluation. Therefore, race was unknown for almost 70% of the sample and is not analyzed here.

Analyses and Results

Difference scores (*D*-scores) were calculated for each child by subtracting the scaled score at Time 2 from the scaled score at Time 1 for each subtest and Full IQ (see Table 2). First, to compare *D*-scores from our sample with those from previous studies on the Flynn effect with children in special education (Kanaya et al., 2003; Sanborn, Truscott & Phelps, 2003; Truscott & Frank, 2001) multiple regression analyses were conducted to determine the relationship between the *D*-scores for Full IQ and testing combinations. The RR group served as the reference group in these analyses.

Specifically, we expected the R3 sample to experience a significantly larger decline in IQ compared with the RR group, whereas the 33 group would not be significantly different from the RR group. The Flynn effect would predict this pattern because the rise within an individual is steady but slight and statistically insignificant. Rather, there would be a significant drop on the introduction of a new norm as an individual's performance on an outdated and highly inflated norm will be directly lowered in response to his or her performance on the new, harder norm. Therefore, children who are tested on the new norm both times will perform similarly to children who are tested on the old norm both times. On the other hand, children who are tested on an old

Table 1. Descriptive Information of the RR, R3, and 33 Groups

Group	Sex, *n* (%)				Age in Months, *M* (*SD*)		Months Between Tests *M* (*SD*)
	Male	Female	Unknown	Total	Test 1	Test 2	
RR	952 (69.2)	398 (28.9)	26 (1.9)	1,376 (100)	106.93 (22.79)	146.35 (25.48)	41.51 (47.25)
R3	451 (69.7)	192 (29.7)	4 (0.6)	647 (100)	105.49 (21.34)	149.64 (24.07)	45.96 (46.42)
33	363 (67.9)	169 (31.6)	3 (0.6)	535 (100)	107.43 (21.85)	143.73 (22.90)	38.30 (48.66)

Note: WISC-R = Wechsler Intelligence Scale for Children–Revised; WISC-III = Wechsler Intelligence Scale for Children–Third edition; RR = tested on the WISC-R both times; R3 = tested on the WISC-R first time and WISC-III second time; 33 = tested on the WISC-III both times.

norm (WISC-R) but retested on a new norm (WISC-III) will experience a significant drop in IQ compared with children who are tested on the same norm both times.

Previous research has shown that the magnitude of the Flynn effect can differ by age (Kanaya, Ceci, & Scullin, 2005) and ability level (Sanborn et al., 2003). Therefore, initial IQ and age at initial testing (in months) were included in the model as covariates. To account for practice effects, which are inversely related to the length of time between testings (e.g., Horton, 1992; Sirois et al., 2002), the time between tests (in months) was also included as a covariate. Although some may disagree with our decision to include initial IQ as a covariate in these analyses (e.g., Allison, 1990), this model choice is consistent with previous research. Furthermore, given that the current sample includes children with cognitive disabilities and who are experiencing academic difficulties (the main reason why children are tested for special education services), we felt it was necessary to control for initial IQ when analyzing the data.

The overall model was statistically significant, $F(5, 2{,}509) = 112.99$, $p < .001$, adjusted $R^2 = 0.18$. As expected, children in the R3 group experienced a significantly lower *D*-score, approximately 4.72 points lower, than children in the RR group, whereas there was no significant difference between the two groups tested with the same norms (the RR and the 33 groups) after controlling for the covariates (see Table 3). Because these results on Full IQ follow the same pattern as the aforementioned Flynn effect studies, follow-up analyses on the individual subtests were conducted.

A multivariate analysis of covariance was conducted on the *D*-scores for all the subtests with the exception of Mazes and Symbol Search. Mazes, a supplemental subtest, was excluded from the analyses because of its reduced sample size. Symbol Search was also not analyzed because it was not among the subtests included in the WISC-R. Testing Group, initial age, initial IQ, and time between the administration of the two tests were included in the model as predictor variables. Box's *M* test revealed that the covariance matrices of the dependent variables were equal across groups (Box's $M = 153.72$, $p = .12$). The overall model was statistically significant for each subtest (ps < .001 for every subtest). Multivariate tests revealed a significant main effect for testing combination, Wilkes $\Lambda = 0.78$, $F(22, 2{,}698) = 15.90$, $p < .001$, $\eta_P^2 = .12$. The test for between-subjects effects revealed this effect was significant for every subtest except Information and Digit Span.

To conduct post hoc analyses on Testing Group, follow-up regression analyses were conducted on each subtest except Information and Digit Span. Again, the RR group served as the reference group, and initial age, initial IQ, and time between testings were included as covariates. To control for multiple tests, the reduced alpha level of .006 ($\alpha = .05/9$) was used to determine statistical significance. As Table 3 reveals, all nine models were statistically significant despite using this very conservative alpha level. Furthermore, the Flynn effect, seen as a significant

Table 2. Means (Standard Deviations) of Subtest Data for the Three Testing Groups

	RR				R3				33			
Subtest	*N*	Time 1	Time 2	*D*	*n*	Time 1	Time 2	*D*	*n*	Time 1	Time 2	*D*
Verbal												
Information	1,356	6.53 (3.14)	6.33 (3.02)	−0.20 (2.30)	646	7.13 (3.19)	6.81 (3.13)	−0.32 (2.57)	535	7.49 (2.99)	7.21 (3.11)	−0.28 (2.42)
Similarities	1,355	8.09 (3.63)	7.86 (3.31)	−0.24 (2.94)	646	9.04 (3.47)	7.27 (3.10)	−1.76 (3.06)	535	7.52 (3.38)	7.68 (3.18)	0.16 (2.79)
Arithmetic	1,339	7.03 (2.77)	6.85 (2.61)	−0.17 (2.35)	644	7.51 (2.89)	6.68 (2.71)	−0.83 (2.54)	534	7.29 (3.07)	6.91 (2.55)	−0.38 (2.52)
Vocabulary	1,353	7.75 (3.14)	6.93 (3.05)	−0.82 (2.26)	646	8.35 (3.07)	6.48 (2.90)	−1.88 (2.44)	535	7.63 (3.13)	6.78 (3.07)	−0.85 (2.41)
Comprehension	1,336	8.11 (3.10)	7.51 (2.85)	−0.60 (2.60)	644	8.46 (3.13)	6.83 (3.25)	−1.64 (3.06)	535	7.89 (3.57)	7.22 (3.41)	−0.66 (2.99)
Digit Span	672	7.03 (2.90)	6.94 (2.71)	−0.08 (2.64)	423	7.45 (2.98)	7.59 (2.75)	0.14 (2.64)	300	7.89 (2.71)	7.80 (2.65)	−0.09 (2.64)
Performance												
Picture Completion	1,347	8.45 (3.15)	8.74 (3.13)	0.29 (2.82)	639	8.92 (2.92)	7.94 (3.34)	−0.98 (2.96)	535	7.83 (3.28)	8.36 (3.27)	0.53 (2.82)
Coding	1,290	7.52 (3.34)	7.47 (3.18)	−0.06 (3.02)	624	8.18 (3.26)	7.16 (3.14)	−1.02 (3.11)	533	7.97 (3.17)	7.26 (2.91)	−0.71 (2.92)
Picture Arrange	1,343	8.23 (3.96)	9.14 (3.33)	0.90 (3.10)	638	9.12 (3.54)	7.40 (3.28)	−1.71 (3.26)	535	7.58 (3.56)	7.84 (3.43)	0.25 (2.90)
Block Design	1,348	7.51 (3.30)	7.48 (3.38)	−0.03 (2.59)	639	7.92 (3.16)	7.06 (3.64)	−0.86 (2.87)	535	7.32 (3.36)	7.46 (3.76)	0.14 (2.72)
Object Assembly	1,342	8.10 (3.30)	8.42 (3.46)	0.32 (2.86)	636	8.46 (3.19)	7.65 (3.46)	−0.81 (2.95)	530	7.69 (3.16)	7.83 (3.26)	0.15 (2.98)
Symbol Search	1	11.00 (-)	9.00 (—)	−2.00 (—)	1	12.00 (—)	0.00 (—)	−12.00 (—)	183	4.58 (4.68)	4.75 (5.05)	0.17 (5.61)
Mazes	21	9.95 (2.46)	10.71 (2.90)	0.76 (2.96)	7	8.86 (3.81)	9.14 (4.53)	0.29 (3.77)	1	7.00 (—)	13.00 (—)	6.00 (—)
Full	1,367	83.87 (16.90)	83.51 (16.18)	−0.36 (8.61)	642	87.97 (15.34)	81.84 (14.80)	−6.12 (8.63)	534	84.86 (15.05)	83.79 (15.10)	−1.07 (7.55)

Note: WISC-R = Wechsler Intelligence Scale for Children–Revised; WISC-III = Wechsler Intelligence Scale for Children–Third edition; RR = tested on the WISC-R both times; R3 = tested on the WISC-R first time and WISC-III second time; 33 = tested on the WISC-III both times.

Table 3. Regression Coefficients Predicting Subtest *D*-Scores

	R3			33			Initial Age			Months Between Tests			Initial IQ		
Subtest	B	*SE*	β	B	*SE*	β	B	*SE*	β	B	*SE*	β	B	*SE*	β
Verbal															
Similarities[a]	−1.31**	0.14	−.19	.39	0.15	.52	−.00	0	−.02	−.02**	0	−.08	−.03**	0	−0.15
Arithmetic[b]	−0.49**	0.12	−.09	−.23	0.12	−.04	−.01**	0	−.06	−.01**	0	−.08	−.02**	0	−.16
Vocabulary[c]	−0.89**	0.11	−.16	−.04	0.12	−.01	.01**	0	.12	−.01**	0	−.09	−.02**	0	−0.15
Completion[d]	−0.82**	0.14	−.13	−.11	0.14	−.02	−.00	0	−.03	−.02**	0	−.12	−.03**	0	−.15
Performance															
Picture Completion[e]	−1.18**	0.14	−.18	.26	0.15	.04	−.00	0	−.02	−.01	0	−.02	−.02**	0	−.10
Coding[f]	−0.81**	0.15	−.12	−.65**	0.16	−.09	.01**	0	.07	−.01	0	−.03	−.03**	0	−.14
Picture Arrange[g]	−2.39**	0.15	−.32	−.59**	0.15	−.07	−.02**	0	−.12	−.00	0	−.02	−.05**	0	−.24
Block Design[h]	−.77**	0.13	−.12	.16	0.14	.02	.01	0	.05	−.01	0	−.04	−.01	0	−.04
Object. Assembly[i]	−.98**	0.14	−.15	−.17	0.15	−.02	−.00	0	−.02	−.01**	0	−.06	−.02**	0	−.09
Full[j]	4.72**	0.39	.24	.71	0.41	.03	.00	0.01	.00	.07**	0.01	.11		0.01	.31

a. $F(5, 2{,}488) = 45.89^{**}$, adjusted $R^2 = .08$.
b. $F(5, 2{,}470) = 21.80^{**}$, adjusted $R^2 = .04$.
c. $F(5, 2{,}499) = 50.53^{**}$, adjusted $R^2 = .09$.
d. $F(5, 2{,}467) = 30.64^{**}$, adjusted $R^2 = .06$.
e. $F(5, 2{,}483) = 26.55^{**}$, adjusted $R^2 = .05$
f. $F(5, 2{,}411) = 25.25^{**}$, adjusted $R^2 = .05$.
g. $F(5, 2{,}478) = 100.09^{**}$, adjusted $R^2 = .17$.
h. $F(5, 2{,}484) = 14.35^{**}$, adjusted $R^2 = .03$.
i. $F(5, 2{,}472) = 18.12^{**}$, adjusted $R^2 = .03$.
j. $F(5, 2{,}509) = 112.99^{**}$, adjusted $R^2 = .18$.
$^{**}p < .006$.

decrease with the R3 group coupled with an insignificant difference with the 33 group, was found in the following subtests: Similarities, Vocabulary, Comprehension, Picture Completion, Block Design, Arithmetic, and Object Assembly. Picture Arrangement and Coding, however, experienced a different pattern. Rather, the R3 *and* 33 groups experienced significant declines compared with the children in the RR group in these subtests. It is important to note that the initial *D*-scores (before the regression analyses) on Picture Arrangement followed the expected Flynn effect pattern; a slight increase in the RR and 33 groups whereas the R3 group experienced a substantial drop. The regression analyses, however, illustrate that the rise experienced in the 33 group was significantly lower than the rise experienced by the RR group, and thus, this subtest did not follow the traditional Flynn effect pattern.

Discussion

To date, a relatively small proportion of studies on the Flynn effect have examined the WISC subtests, one of the most commonly used IQ tests in the United States. Much of the attention on the Flynn effect have focused on adult data or measures that are more popular in other countries, such as the Ravens Progressive Matrices (e.g., Flynn, 1987). Although Flynn (2007; Flynn & Weiss, 2007) documented differential gains between the WISC subtests, this study analyzed these gains on a sample that is, potentially, the most affected by IQ tests: school children who are being tested for special education services. Furthermore, by using statistical analyses, we were able to measure these gains after controlling for confounding variables.

Our findings reveal that seven of the subtests within the WISC-R and WISC-III manifested the Flynn effect, which is marked by an insignificant change in scores when tested and retested on the same norm, but a significant decrease in scores when tested on an old norm and retested on a new norm. These subtests were Similarities, Vocabulary, Comprehension, Picture Completion, Block Design, Arithmetic, and Object Assembly. Information and Digit Span were not affected by changing test norms in our sample. Furthermore, although scores on Picture Arrangement and Coding experienced significant decreases when tested on the WISC-R and retested on the WISC-III, they also decreased when tested and retested on the WISC-III when compared with children who were tested and retested on the WISC-R. The patterns within the WISC-III suggest a possible reverse, or diminished, Flynn effect within these two subtests.

Taken together, the fluctuations on these seven subtests were not expected on the basis of the data from the general population (e.g., Flynn, 2007). Specifically, children in the United States have gained 24 points ($SD = 15$) on the Similarities subtest over the last 60 years (Flynn, 2007). In contrast, gains on Information and Vocabulary have been slight. This tendency, however, was not evident among students evaluated for special education eligibility. Although the Information subtest did not show any Flynn effect, replicating previous studies (e.g., Flynn & Weiss, 2007), Vocabulary, Comprehension, Block Design, and Object Assembly were associated with decreases on retesting on the newer, harder WISC-III after controlling for initial testing age, initial IQ, and time between testings. Furthermore, Arithmetic also experienced the Flynn effect in our analyses.

The starkest contrast was on the Similarities subtest, where our regression coefficient was substantially smaller than was reported by Flynn (2009). Kaufman (in press) has argued that Flynn's findings on many of the subtests, but particularly Similarities, are highly overestimated because of substantive administrative and scoring changes that occurred with the introduction of the WISC-R. And so, our findings may provide a more accurate estimate of the Flynn effect on the Similarities subtest. It is also important to note that items on the Similarities subtest such as "what do dogs and rabbits have in common?" have become increasingly easier for children over the past century and as a consequence, its loading on fluid ability has diminished. And so, these

findings could also reflect social changes that affect specific cognitive abilities (Blair, 2006; Ceci & Kanaya, in press), and perhaps the experiences that drive such social effects are unevenly distributed across special and regular education.

These findings add to the limited literature on the Flynn effect within assessments for children, including the Peabody Picture Vocabulary Test (Nettelbeck & Wilson, 2004), the Leidon Diagnostic Test (Resing & Tunteler, 2007), mathematics achievement tests (Rodgers & Wänström, 2007), the Otis–Lennon Test (Flynn, 2007), and even Piagetian tasks (Bocerean, Fischer, & Flieller, 2003; Flieller, 1999). The findings are also similar to a recent study by Must, te Nijenhuis, Must, and van Vianen (2009), which found differential gains on the individual subtests of the American National Intelligence Test on a sample of Estonian school children. They, as have other researchers who have examined the Flynn effect, concluded that scores from different cohorts are not comparable with each other. Our findings illustrate that scores from different norms are also not comparable with each other, particularly when data are gathered from the special education students across different points in their school careers. That fluid tests such as Picture Arrangement and Coding declined over the course of the WISC-III is worrisome, as these subtests increase or remain constant when assessed across different points in the school careers of regular education students. Similarly, the decline in some crystallized abilities (Vocabulary) is larger among these special education referrals than general education peers, suggesting that either their learning difficulties or aspects of their education place them at special risk for cognitive decline.

Strengths, Weaknesses, and Future Directions

There are relatively few studies that examine the Flynn effect on school children, one of the most heavily tested populations in the United States. And so, the longitudinal and cross-sectional nature of our data set provided a rare opportunity to examine IQ trends on the WISC norms. The present study also contains data from a geographically and socioeconomically diverse database of children who were tested as part of their special education evaluations around the country. Thus, the results better lend themselves to be generalized to children who are most affected by IQ scores and most likely to be subjected to multiple testings throughout their school years.

Despite these strengths, some limitations of the present study should be noted. Although school children who are tested for special education services are an important population to study, they do not represent the general population. This is reflected by the fact that the average IQs in our sample were approximately an entire standard deviation less than the population mean of 100 points, and a contraction of variance could produce the same results. It is worth noting, however, that Kanaya, Scullin, et al. (2003) found the Flynn effect estimates among children diagnosed with MR to be similar to those reported by Flynn (1987) within the general population. It is also still unknown if these trends will hold with adults. Furthermore, by focusing our analyses solely on the WISC subtests, we are neglecting many other standardized measures of crystallized and fluid ability, including other IQ and achievement tests that are commonly used in special education diagnoses such as the Stanford Binet, Kaufman, and Woodcock–Johnson scales.

The use of archived special education records, while ecologically valid, made it impossible to use a counterbalanced design. Few, if any, students in our data set were administered the WISC-III *before* the WISC-R during their first two evaluations. We tried to account for practice effects by including the amount of time between tests as a covariate in our analyses. We acknowledge, that this approach does not address and correct for practice effects as effectively as a counterbalanced design. It is important to note, however, that all three groups experienced a *decrease* in Full Scale IQ on retesting. Given the Flynn effect has been estimated as approximately 0.3 points per year and the average time between tests is more than 3 years (36 months), one would expect a one-point increase in Full IQ in the RR and 33 groups. The decrease, is

consistent with previous research on children with LD, the most prevalent diagnosis within special education (Data Accountability Center, 2007). Kaufman (1994) and Sanborn et al., (2003) have reported declines in IQ when retested on the same norm among children with LD. Therefore, it is unclear if children tested for special education services experience practice effects in the same manner as the general population and, therefore, whether any corrections for such effects are required for our sample.

Furthermore, evidence of the Flynn effect is measured by the significant drop on the introduction of the new norm (as seen in the R3 group). The fact that the R3 group experienced a significantly larger decline than the RR group, combined with the fact that the 33 group was not significantly different from the RR group, provide strong evidence of the Flynn effect that cannot be explained by practice effects or actual changes in cognitive ability. Despite this, previous research has shown that practice effects in the general population are stronger on Performance subtests and Full Scale IQ compared with the Verbal subtests (Sirois et al., 2002; Tuma & Appelbaum, 1980). Therefore, future researchers should examine the nature and magnitude of practice effects at the subtest level in this heavily tested population.

In addition, we only examined test–retest scores at 2 points within each child. By limiting the analyses to the first two tests, we reduced the risk of progressive error. Examining longitudinal patterns across multiple testings and throughout the lifespan of an individual, however, will provide more specific insight into the nature of the Flynn effect. Our data also did not allow us to examine individual differences within the IQ gains. For example, we were not able to examine if children of different ethnicities experienced differential gains. Given the long-standing documentation of the overrepresentation of ethnic minorities in special education, it is particularly important to determine if the Flynn effect varies between racial/ethnic groups (Kanaya & Ceci, 2007). Other individual factors that have been underexplored within the Flynn effect literature include sex, age (Kanaya et al., 2005; Spitz, 1989), special education/clinical diagnoses (Kanaya, Scullin, et al., 2003), IQ level (Cahan & Gejman, 1993; Sanborn et al., 2003), and income. Future research should also determine whether subtest gains have an impact on special education diagnoses. In other words, although it has been established that MR and LD diagnoses are affected by changes in Full IQ caused by the Flynn effect (Kanaya, Scullin, et al., 2003; Truscott & Frank, 2001), it will be interesting to see if the similar patterns can be predicted on the basis of changes at the subtest level.

Implications for the WISC-IV and Current School Children

To alleviate the impact of the Flynn effect on school children, the Psychological Corporation introduced the WISC-IV more quickly than it has introduced previous norms (Wechsler, 2003). Although a few studies have suggested that the Flynn effect is reduced or eliminated among adults (e.g., Teasdale & Owen, 2008) as has our findings on Picture Arrangement and Coding, this general trend has not been replicated in preliminary examinations of the current WISC-IV (Flynn & Weiss, 2007). In addition to the quicker release date, the WISC-IV has also undergone several substantive changes to increase "the contribution of fluid reasoning" (Harcourt Assessment, 2004). These changes include the introduction of new subtests, as well as the elimination of Object Assembly, Picture Arrangement, and Mazes. In addition, Arithmetic can now be used as an optional substitute for Digit Span, whereas Information can replace Comprehension and Similarities.

These substantive changes have led some researchers to question the comparability between scores from the WISC-IV and those from previous norms (e.g., Keith, Fine, Taub, Reynolds, & Kranzler, 2006). On one hand, the faster norming cycle is helpful, as it will prevent millions of children from being misdiagnosed based on inflated scores from an aging norm. On the other

hand, however, these substantive changes may be problematic when comparing scores within the same norm. For example, in our analyses, testing combination did not have a significant impact on changes in the Information subtest. And yet, on the WISC-IV Information can be substituted for Comprehension and Similarities, which *did* experience Flynn effect gains. Not only could this substitution increase the discrepancy between WISC-III and WISC-IV scores, it could also create discrepancies *within* WISC-IV scores; over time as the Flynn effect gains accumulate, an IQ derived from the original subtests may differ from one derived from the substituted subtests. It is important to note that this discrepancy could go undetected because it will not change or reduce the coefficients that are used to assess the reliability between the IQ derived from the original subtests and the IQ derived from the substituted subtests. It is clear, therefore, that researchers and practitioners must continue to explore the nature, magnitude, and impact of the Flynn effect on the current WISC-IV.

Conclusion

In conclusion, the current study adds to the growing literature on the Flynn effect. Specifically, we found strong evidence of the Flynn effect on crystallized and fluid subtests during the WISC-R to WISC-III transition. The effect, however, was not uniformly distributed among all the subtests, and there was not a clear pattern of higher gains on fluid measures compared with crystallized measures, as has been found in previous studies. Although these findings may have the most immediate impact on school children, it is important to note that IQ tests play an important role in life-altering decisions for many people. For example, individuals who are diagnosed with mild MR because of their low IQ and concomitant behavioral measures do not qualify for the death penalty (*Atkins v. Virginia*, 2002), and can also be denied organ transplants (Shapiro, 2006). Full Scale IQ scores are used to determine Social Security Disability eligibility and occupational assignments, including decisions within the military, are also based on measures that are highly correlated with traditional IQ tests and other measures that have been found to experience the Flynn effect (Kanaya, Scullin, et al., 2003). Therefore, research that provides more insight into the nature and magnitude of the Flynn effect can have significant implications in the lives of thousands, and possibly millions, of children and adults each year.

Authors' Note

Any opinions, findings and conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the Spencer Foundation or Smith Richardson Foundation.

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) disclosed receipt of the following financial support for the research and/or authorship of this article:

This research was supported, in part, by a research grant by the Spencer Foundation (Grant No. 2007-000115) awarded to the first author. This research was also supported, in part, by a research grant by the Smith Richardson Foundation awarded to the second author.

References

Allison, P. D. (1990). Change scores as dependent variables in regression analysis. *Sociological Methodology, 20*, 93-114.

Atkins v. Virginia, 536 U.S. 304 (2002).

Blair, C. (2006). How similar are fluid cognition and general intelligence? A developmental neuroscience perspective on fluid cognition as an aspect of human cognitive ability. *Behavioral and Brain Sciences, 29*, 109-160.

Bocerean, C., Fischer, J.-P., & Flieller, A. (2003). Long-term comparison (1921-2001) of numerical knowledge in three to five-and-a-half year-old children. *European Journal of Psychology of Education, 18*, 405-424.

Cahan, S., & Gejman, A. (1993). Constancy of IQ scores among gifted children. *Roeper Review, 15*, 140-143.

Cattell, R. B. (1971) Abilities: Their structure, growth, and action. Oxford, England: Houghton Mifflin.

Ceci, S. J., Fitneva, S. A., & Williams, W. M. (in press). Representational constraints on the development of memory and metamemory: A Developmental-representational theory. *Psychological Review*.

Ceci, S. J., & Kanaya, T. (in press). "Apples and oranges are both round:" Furthering the discussion on the Flynn effect. *Journal of Psychoeducational Assessment.*

Data Accountability Center. (2007). *Data tables for OSEP state reported data*. Retrieved from https://www.ideadata.org/arc_toc9.asp#partbCC

Flieller, A. (1999). Comparison of the development of formal thought in adolescent cohorts aged 10 to 15 years (1967-1996 and 1972-1993). *Developmental Psychology*, *35*, 1048-1058. doi:10.1037/0012-1649.35.4.1048

Flynn, J. R. (1984). IQ gains and the Binet decrements. *Journal of Educational Measurement, 21*, 283-290. doi:10.1111/j.1745-3984.1984.tb01035.x

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. (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

Harcourt Assessment. (2004). *WISC-IV frequently asked questions*. Retrieved from http://harcourtassessment.com/HAIWEB/Cultures/en-us/Productdetail.htm?Pid=015-8979-044&Mode=resource&Leaf=015-8979-044_FAQ

Horton, A. M. (1992). Neuropsychological practice effects x age: A brief note. *Perceptual and Motor Skills*, *75*, 257-258. doi:10.2466/PMS.75.4.257-258

Kanaya, T., & Ceci, S. J. (2007). 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., & Scullin, M. H. (2003). The rise and fall of IQ in special ed: Historical trends and their implications. *Journal of School Psychology, 41*, 453-465. doi:10.1016/j.jsp.2003.08.003

Kanaya, T., Ceci, S. J., & Scullin, M. H. (2005). Age differences in 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. (1994). *Intelligent testing with the WISC-III*. New York. NY: Wiley.

Kaufman, A. S. (in press). "In what way are apples and oranges alike?" A critique of Flynn's interpretation of the Flynn Effect. *Journal of Psychoeducational Assessment.*

Keith, T. Z., Fine, J. G., Taub, G. E., Reynolds, M. R., & Kranzler, J. H. (2006). Higher order, multisample, confirmatory factor analysis of the Wechsler Intelligence Scale for Children–Fourth edition: What does it measure? *School Psychology Review, 35*, 108-127.

Keith, T. Z., & Witta, E. L. (1997). Hierarchical and cross-age confirmatory factor analysis of the WISC-III: What does it measure? *School Psychology Quarterly, 12*, 89-107. doi:10.1037/h0088950

Must, O., te Nijenhuis, J., Must, A., & van Vianen, A. E. M. (2009). Comparability of IQ scores over time. *Intelligence, 37*, 25-33. doi:10.1016/j.intell.2008.05.002

Nettelbeck, T., & Wilson, C. (2004). The Flynn effect: Smarter not faster. *Intelligence, 32*, 85-93. doi:10.1016/S0160-2896(03)00060-6

Resing, W. C. M., & Tunteler, E. (2007). Children becoming more intelligent: Can the Flynn effect be generalized to other child intelligence tests? *International Journal of Testing, 7*, 191-208. doi:10.1080/15305050701193546

Rodgers, J. L., & Wänström, L. (2007). Identification of a Flynn Effect in the NLSY: Moving from the center to the boundaries. *Intelligence, 35*, 187-196. doi:10.1016/j.intell.2006.06.002

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

Shapiro, J. (2006). *Dispute over mental competency blocks transplant*. Retrieved from www.npr.org

Sirois, P. A., Posner, M., Stehbens, J. A., Loveland, K. A., Nichols, S., Donfield, S. M., . . . Amodei, N. (2002). Quantifying practice effects in longitudinal research with the WISC-R and WAIS-R: A study of children and adolescents with hemophilia and male siblings without hemophilia. *Journal of Pediatric Psychology, 27*, 121-131.doi:10.1093/jpepsy/27.2.121

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

Teasdale, T. W., & Owen, D. R. (2008). Secular declines in cognitive test scores: A reversal of the Flynn effect. *Intelligence*, *36*, 121-126. doi:10.1016/j.intell.2007.01.007

te Nijenhuis, J., & van der Flier, H. (2007). The secular rise in IQs in the Netherlands: Is the Flynn effect on *g*? *Personality and Individual Differences, 43*, 1259-1265. doi:10.1016/j.paid.2007.03.016

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. doi:10.1016/S0022-4405(01)00071-1

Tuma, J. M., & Appelbaum, A. S. (1980). Reliability and practice effects of WISC-R IQ estimates in a normal population. *Educational and Psychological Measurement*, *40*, 671-678. doi:10.1177/001316448004000310

Watkins, M. W., & Kush, J. C. (2002). Confirmatory factor analysis of the WISC-III for students with learning disabilities. *Journal of Psychoeducational Assessment, 20*, 4-19. doi:10.1177/073428290202000101

Wechsler, D. (2003). *The Wechsler Intelligence Scale for Children-IV Manual.* New York, NY: Psychological Corporation.