Pergamon

Journal of School Psychology, Vol. 39, No. 4, pp. 319–334, 2001
Copyright © 2001 Society for the Study of School Psychology
Printed in the USA
0022-4405/01 $–see front matter

PII S0022-4405(01)00071-1

Does the Flynn Effect Affect IQ Scores of Students Classified as LD?

Stephen D. Truscott and Alicia J. Frank
The University at Buffalo, SUNY

This research examined two samples of students classified as learning disabled (LD) for evidence of the phenomenon known as the Flynn effect (FE; Flynn, 1999). Triennial test data were collected for two samples. Sample 1 included students tested twice with the Wechsler Intelligence Scale for Children—Third Edition (WISC-III; $n = 59$). Sample 2, the primary data set, included students tested first with the Wechsler Intelligence Scale for Children—Revised (WISC-R), and then with the WISC-III ($n = 171$). A secondary analysis examined potential differences in the FE by ethnicity and/or gender. Results indicate that the FE does affect Wechsler IQ and component scores of students classified as LD. Further, the effect varies by task. There were no significant differences in the FE by race and/or gender. The study suggests that LD classification may be substantially impacted by the FE over the life of an IQ test version. © 2001 Society for the Study of School Psychology. Published by Elsevier Science Ltd

Keywords: Flynn effect, Intelligence, IQ, Assessment, LD.

Over the years, several studies have reported a decrease in students' IQ scores between different revisions of popular intelligence tests such as the Wechsler batteries (e.g., Bolen, Aichinger, Hall, & Webster, 1995; Carlton & Sapp, 1997; Gaskill & Brantley, 1996; Graf & Hinton, 1994; Lynn & Hampson, 1986; Slate & Saarnio, 1995; Spitz, 1989; Thorndike, 1975; Vance, Maddux, Fuller, & Awadh, 1996). One explanation for the observed decline in scores with successive revisions of IQ tests is a phenomenon known as the "Flynn effect" (FE; Flynn, 1984, 1987). In two extensive studies, Flynn (1984, 1987) detailed a consistent trend of substantive increases in the normative performances on intelligence tests in the United States and throughout the world. In short, for the U.S. population, Flynn (1984) reported an increase of approximately 3 IQ points per decade.

There is no consensus about the cause of the FE. However, it is a population phenomenon that has occurred over at least the last 60 years and, in the United States, at a relatively consistent pace (Flynn, 1984, 1987, 1999).

Received August 20, 2000; accepted February 12, 2001.

Address correspondence and reprint requests to Stephen D. Truscott, Graduate School of Education, Department of Counseling, Educational, and School Psychology, 406 Baldy Hall, The University at Buffalo, SUNY, Buffalo, NY 14260. Phone: (716) 645-2484; fax: (716) 645-6616; E-mail: sdt55@acsu.buffalo.edu

It influences scores on most IQ measures, although it appears to affect scores on nonverbal tests (e.g., Raven Progressive Matrices) more than scores on more verbally-loaded instruments [e.g., Wechsler Intelligence Scales for Children (WISC)]. Flynn (1999) explained this difference as a greater effect on fluid versus crystallized cognitive abilities. One result of the FE is that because the normative performances on IQ tests improve steadily, an individual's scores can change substantially between revisions of IQ tests. This change occurs because the gradual increase in the population mean performance is masked until the test is revised, which may occur only every 20 years or more. As a result, IQ scores derived from older test versions are increasingly inflated as the norms age. In contrast, a given performance on a test will score lower when compared with the more recent and appropriate normative group because the normative performance on that test has steadily increased over time.

Although current research suggests the phenomenon is widespread in normal adult populations, there is very limited information about the FE with other groups. Recently, Flynn (1999, 2000) outlined some potential issues regarding exceptional populations, but there are very few studies with these groups. This is, however, a critically important area for research. IQ is a basic component of diagnosis of most educationally related exceptionalities, especially mental retardation and learning disability (LD). Any widespread factor that fundamentally affects IQ scores for the population at large, or for these specific populations, will systematically alter classification criteria over time. Further, research on the FE for the normal population suggests that it is limited to IQ and does not substantially affect academic achievement scores (Neisser, 1998). If this is true for the learning disabled (LD) population, then there is a continual increase in the base-rate difference between IQ and academic achievement that alters the most salient LD classification criterion over time. A potentially confounding issue is that existing research suggests that the FE may be either reduced or not applicable for some exceptional populations and that it affects scores in some IQ ranges (e.g., mentally retarded and gifted) differently than scores in the average range (Spitz, 1989).

There is relatively little literature directly assessing the FE phenomenon for children, yet a review of research on the stability of WISC scores (Wechsler, 1974, 1991) for both normal and learning-disabled populations provides indirect information. Most researchers report that the Full, Verbal, and Performance scores are moderately to highly stable, especially as measured by correlations, when students in the United States are tested twice with the same version of the WISC [e.g., WISC-Revised (WISC-R; Wechsler, 1974); Naglieri & Pfeiffer, 1983; Oakman & Wilson, 1988; Truscott, Narrett, & Smith, 1994; or WISC-Third Edition (WISC-III; Wechsler, 1991); Canivez & Watkins, 1998]. Studies that compared scores for students tested with the two different versions of the Wechsler scales (WISC-R and WISC-

III) have shown drops in IQ scores of 5–8 points (e.g., Bolen et al., 1995; Carlton & Sapp, 1997; Gaskill & Brantley, 1996; Graf & Hinton, 1994; Slate & Saarnio, 1995). This drop corresponds to what might be expected, given the FE, because the two tests were normed approximately 2 decades apart (WISC-R normed in approximately 1972; WISC-III normed in approximately 1989).

Only two published research studies were identified that specifically examined the FE with exceptional populations (Graf & Hinton, 1994; Spitz, 1989). These studies did not specifically single out students classified as LD. Spitz included gifted, mentally retarded, and average IQ students in his study of the FE as exhibited with the Wechsler scales. Although he observed the FE in the normal ranges of IQ in his sample, the FE gradually diminished above and below the average range, and actually reversed in the mentally retarded range. As IQ decreased, an increasing disparity between the two versions of the tests was evident. Specifically, scores on the second test were increasingly higher than scores on the first test as IQ decreased (the reverse of what is expected, given the FE). Similarly, the FE was less evident in the high-IQ range in Spitz's study. Flynn (1985) disputed these results, and suggested that the differences seen in Spitz's study were attributable to norming differences between the WISC and WISC-R, but the dispute has not been resolved by replication.

Graf and Hinton (1994) examined triennial data for 84 special education children tested first with the WISC-R and then with the WISC-III. Overall, they found small but significant decreases between the WISC-R and WISC-III Performance and Full Scale scores, but not for the Verbal Scale. The differences were, however, substantially smaller than predicted using the FE [i.e., Full Scale IQ (FSIQ) score difference was only about 2 points rather than the 5.7 points expected using Flynn's earlier reports (Flynn, 1984)]. Graf and Hinton (1994) then divided the sample into five IQ score-range subgroups (e.g., IQ 76–90) for the Verbal IQ (VIQ), Performance IQ (PIQ), and FSIQ scores, respectively. Contrary to what one might predict given the FE, Graf and Hinton (1994) reported that WISC-III VIQ, PIQ, and FSIQ results were somewhat higher than the WISC-R for the two low IQ groups (60–75 and 76–90). Above the 90 IQ level, scores followed the FE expectation, and WISC-R scores were higher than the WISC-III. Both of these studies generally support the existence of the FE in special education populations, but suggest that it is a complex phenomenon that may differentially affect scores outside of the normal IQ range. Neither study substantively addressed the FE with a LD group.

The influence of the FE is an especially crucial concern for exceptional populations because IQ is commonly used to determine eligibility for extra services in schools (e.g., LD classification). This affects not only students currently being evaluated, but also those already in special education. Special education students are routinely administered the latest version of an intelligence test as part of mandated triennial evaluations, resulting in many students with scores from different versions of a given test and/or different IQ measures that were normed at widely different times in their assessment history. It is consequently critical to understand whether potential differences between tests and revisions are due to real individual differences, changes in an individual's cognitive abilities, or psychometric peculiarities.

The purpose of this study was to examine the existence of the FE with a LD sample as measured by versions of the WISC. The study examined the Verbal, Performance, and Full Scale scores of 171 LD students tested with the WISC-R and then the WISC-III as part of their triennial evaluations. Further analyses examined the subtest scores for this sample. A secondary purpose of this study was to provide a preliminary examination of the Verbal, Performance, and Full Scale scores for evidence of differences in the FE by ethnicity and/or gender.

METHODS

Participants

The current study used a previously existing data set consisting of archival records collected from 22 school districts in New York State (NYS). The data included information about 257 public school students who were classified as learning disabled (University of the State of New York, State Education Department, 1993). Sample 1 consisted of students who were evaluated twice with the WISC-III, approximately 3 years apart, as part of evaluations and triennial reevaluations mandated for special education. This sample was used solely to calculate adjustments for the 3-year interval between test scores. Sample 2, the primary data set, consisted of students who were administered the WISC-R once and then reevaluated with the WISC-III at the next triennial. Participants were eliminated if they obtained a FSIQ score below 80 on the first IQ test because Spitz (1989) suggested that scores below that point might not exhibit a consistent degree of the FE. Further, excluding students with FSIQ scores below 80 eliminated the possibility of including students more accurately classified as mentally retarded in NYS (IQ ≤ 77). This resulted in sample sizes of 49 and 171 for Samples 1 and 2, respectively.

In Sample 1 (n = 49) 27 (55%) of the students were male, 17 (35%) female, and 5 (10%) did not have gender recorded. Race information was not reported for 4 students (8%); those records that included race information were distributed as 71% Caucasian, 14% African American, and 6% Hispanic, Asian, or Native American. The average age at the first evaluation was 9.25 years (*SD* = 2.06). The average time between the tests was 2.73 years (*SD* = 0.65). The average initial WISC-III IQ was 92.33 (*SD* = 10.07).

Twenty-one schools contributed information from student records. Forty-one percent of the students came from urban settings, 13% came from small cities, 30% attended rural schools, and 16% came from the suburbs.

Thirty-three percent (56) of the 171 students in Sample 2 were female. The mean age at the first testing was 9.12 years (*SD* = 1.73). Eighteen percent of the sample did not report race information. Those records that included race information were distributed as 54% Caucasian, 23% African American, and 5% Hispanic, Asian, or Native American of the total sample. The average WISC-R IQ at the first testing was 94.32 (*SD* = 10.55). The average time between tests was 3.32 years (*SD* = 0.99). The 20 schools that contributed data were distributed as 52% urban, 20% small cities, 17% rural, and 11% suburban.

Race and gender sample. Only records from Sample 2 that included both race and gender information were selected for the Race × Gender analysis. The few records from races other than Caucasian and African American were eliminated. This resulted in an overall sample of 128 comprised of 90 Caucasian (63 males, 27 females) and 38 African American (12 females, 26 males) participants.

General characteristics of the samples. The majority of students in each sample were identified as primarily LD in reading. Nearly all of the students had received special education services for at least 3 years, with some of the students receiving it for as long as 6 years. Although a small number of the students had changes in their classification, all of them were classified as LD at their last evaluation.

Homogeneity of the sample. Variation in LD classification criteria is a general concern for most LD research. For this study, all students in each sample were identified as LD according to NYS regulations. NYS regulations require a significant difference between ability and achievement, but do not specify how to calculate this discrepancy. To provide some measure of the homogeneity of the primary sample (Sample 2), the researchers calculated the discrepancy between achievement and IQ for each student record that included achievement data. Most achievement data were recorded as standard scores. Those data recorded as percentiles and Normal Curre Equivalent scores were converted to standard scores. Achievement data recorded as grade equivalents were not converted. Of the 171 participants in Sample 2, 155 (90.6%) had IQ and achievement data that could be used to calculate a standard score discrepancy. Each student's IQ (VIQ, PIQ, and FSIQ) and achievement scores (reading, writing, and arithmetic) were examined and the discrepancy between the highest IQ score and lowest achievement score was calculated. Of the 155 student records with calculated discrepancy scores, 116 (74.8%) had discrepancies ≥15 points (*M* = 28.09, *SD* = 12.25).

This indicates that a substantial majority of the sample had a large discrepancy between IQ and academic achievement. Such a substantial discrepancy is an objective, common, and salient element of LD classification

Design

This research employed a longitudinal archival design. This allowed the use of authentic existing school records that mirror the actual practice of schools. The design presents certain limitations that will be discussed in a later section. The researchers asked employees of the respective school districts to select potential subjects based on two criteria: (a) the students were classified as LD at a recent triennial evaluation and (b) the students had existing data from two Wechsler scales. District employees, usually the school psychologist, created a code for each student and recorded the data on forms provided by the researchers. Information was collected on each student's date of birth, gender, race, age, and grade at testing, achievement test data (where possible), and primary academic difficulty in addition to the subtest and scale scores from the respective Wechsler scores.

Measures

The latest two revisions of the WISC [WISC-R (Wechsler, 1974) and WISC-III (Wechsler, 1991)] were used in this study. Correlation coefficients for the WISC-R and the WISC-III are high, averaging .89, .90, and .81 for the Full, Verbal, and Performance Scales, respectively (Wechsler, 1991). This is not surprising, given the high degree of similarity between the WISC-R and the WISC-III. Approximately 73% of the items from the WISC-R were carried over into the WISC-III (Edelman, 1996), resulting in only about one third of the WISC-III items that are new or modified (Kaufman, 1993). New items were added to several of the subtests to improve discrimination for exceptional students. These included additional floor and/or ceiling items for Similarities, Arithmetic, Picture Arrangement, Block Design, and Mazes. Overall, however, the tests are largely the same and the preponderance of research supports high correlation coefficients between them (e.g., Bolen et al., 1995; Carlton & Sapp, 1997; Gaskill & Brantley, 1996; Graf & Hinton, 1994). Therefore, it seems likely that the drop in scores commonly reported between the versions is unlikely to be an anomaly of these two specific WISC batteries. In addition, the use of different versions of IQ tests and the assumption of equivalence were regular elements of Flynn's previous work (1984, 1987).

Procedures

This study analyzed archival records of students classified as LD and tested approximately 3 years apart. The primary data set (Sample 2) contained records of students tested first with the WISC-R, then with the WISC-III. To determine whether the FE influenced the scores when the students were tested with the different versions of the Wechsler, two initial calculations were necessary. First, the predicted true score (PTS; Atkinson, 1991) for the first administration was calculated to statistically adjust for regression to the mean and the reliability of the initial test score. Second, the average IQ change attributable to the time difference (3 years between test administrations) was calculated using Sample 1 (estimated mean change; EMC). The obtained score for the first administration (WISC-R) for Sample 2 was adjusted for both the PTS and the EMC to derive a modified IQ (MIQ) that was compared with the obtained WISC-III score for evidence of the FE. The procedures for calculating these scores are described below.

PTS. The PTS was calculated for both samples using each student's first test administration and the following procedure: "a) Multiply the obtained score by the reliability of the test; b) multiply the mean of the test by one minus the reliability; and c) add the results of (a) and (b)" (Atkinson, 1991, p. 137). This method accounts for statistical regression to the mean and the reliability of the test to produce a modified score that is more suitable for comparison than simple obtained scores.

EMC. The 3-year time period between administrations of the test was expected to produce changes in IQ unrelated to the FE. There is some controversy about what changes can be expected in IQ scores for the LD population between test administrations (e.g., Canivez & Watkins, 1998; Graf & Hinton, 1994; Stavrou, 1990; Truscott et al., 1994). This study sought to control for the variety of reported Wechsler score changes over 3 years by collecting WISC-III to WISC-III data from the same schools at the same time, using the same criteria that were used for the WISC-R to WISC-III sample. Because the two samples are similar on primary variables (e.g., the representative proportions of different subgroups, mean IQ, and age at first testing), sampling errors were likely minimized.

The average expected change in IQ over 3 years was calculated using Sample 1 (WISC-III to WISC-III). Sample 1 was used because the norms for the WISC-III did not change and, therefore, the scores were not influenced by the FE. The mean changes in the scale and subtest scores were calculated for the sample. These mean changes were determined by subtracting the score of the second test from the PTS of the first test for each of the participants, and obtaining the EMC in IQ.

MIQ. The MIQ was derived for each student by calculating the PTS from the initial WISC-R score and adjusting the PTS by the EMC. MIQ values were then compared with the obtained WISC-III score. Comparisons of the

Wechsler scale and subtest scores were made with *t*-tests with a Bonferroni correction.

RESULTS

Primary Analysis

The mean changes in IQ and component scores for Sample 1 were calculated by subtracting the PTS of the first administration from the second administration of the WISC-III. This results in an estimation of the average change in score that can be expected over the course of about 3 years for this population. The EMC values derived from this calculation are reported in Table 1. For example, the EMC values for the Full, Verbal, and Performance Scale scores were −2.39, −3.49, and −0.83, respectively. This is consistent with previous reports that most IQ and component scores

Table 1
Expected Mean Change Calculations from Sample 1 (WISC-III to WISC-III)

WISC-III Component	n	Range of Obtained Scores	Obtained first WISC-III Score (*SD*)	PTS	Range of Obtained Scores	Obtained second WISC-III Score (*SD*)	EMC
Full	49	80–118	92.33 (9.66)	92.63	68–108	90.24 (9.34)	−2.39
Verbal	49	72–124	91.76 (10.50)	92.17	66–113	88.67 (9.94)	−3.49
Performance	49	70–121	94.49 (11.57)	94.99	72–127	94.16 (11.78)	−0.83
Information	49	4–15	8.14 (3.03)	8.44	2–13	7.90 (2.31)	−0.54
Similarities	49	4–16	8.88 (2.31)	9.09	4–14	8.55 (2.27)	−0.54
Arithmetic	49	4–15	7.78 (2.57)	8.26	4–13	7.29 (2.03)	−0.98
Vocabulary	49	2–14	8.53 (2.46)	8.72	4–15	7.59 (2.53)	−1.13
Comprehension	49	2–15	9.24 (2.70)	9.42	2–14	8.20 (2.97)	−1.21
Picture Completion	49	1–14	9.08 (2.44)	9.29	5–17	9.35 (2.18)	−0.05
Picture Arrangement	49	1–17	9.41 (3.14)	9.55	3–15	9.35 (2.42)	−0.20
Block Design	49	2–15	9.04 (2.60)	9.17	2–15	9.22 (2.69)	−0.06
Object Assembly	49	3–15	9.04 (2.92)	9.34	2–17	9.02 (3.00)	−0.32
Coding	49	2–19	8.47 (3.62)	8.79	2–17	8.12 (2.90)	−0.67

WISC-III = Wechsler Intelligence Scale for Children—III, PTS = predicted true score, EMC = estimated mean change.

drop between triennial evaluations for the LD population, even when the test does not change (e.g., Kaufman, 1994; Truscott et al., 1994).

Results for the overall FE analyses are reported in Table 2. The MIQ score was calculated by adjusting the obtained score from the WISC-R for regression to the mean and error (identified as PTS) and subtracting the expected change over time (EMC). The MIQ score was then subtracted from the WISC-III to obtain a residual score that contained the FE. *t*-Tests were calculated for differences between the MIQ and obtained WISC-III. Because the WISC-R and WISC-III were normed approximately 2 decades apart, in the United States, one would expect an average residual of approximately 6 points (Flynn, 1999) for the FSIQ, VIQ, and PIQ, assuming the effect was distributed evenly across the IQ scores.

The FSIQ, VIQ, and PIQ WISC-III scores are all significantly different from the MIQ WISC-R score. This suggests that the FE did affect the IQ

Table 2
Comparison of Modified WISC-R to WISC-III Scores

WISC Component	*n*	Range of Observed Scores	Obtained WISC-R IQ (*SD*)	PTS	EMC	MIQ	Range of Observed Scores	Obtained WISC-III IQ (*SD*)	Change (*ES*)[a]
Full	171	80–142	94.32 (10.55)	94.56	−2.39	92.16	56–126	87.39 (11.08)	−4.77* (−.45)
Verbal	171	66–137	92.25 (11.75)	92.72	−3.49	89.23	59–117	86.29 (11.22)	−2.93* (−.25)
Performance	171	72–138	98.04 (12.42)	98.23	−0.83	97.41	60–131	90.73 (13.26)	−6.68* (−.54)
Information	170	2–15	7.99 (2.42)	8.25	−0.54	7.76	1–15	7.39 (2.52)	−0.36 (−.15)
Similarities	171	1–18	9.87 (2.74)	9.89	−0.54	9.35	1–14	8.15 (2.51)	−1.20* (−.44)
Arithmetic	155	2–17	7.94 (2.41)	8.38	−0.98	7.43	1–15	7.17 (2.40)	−0.26 (−.11)
Vocabulary	171	1–17	8.58 (2.64)	8.78	−1.13	7.65	1–14	7.14 (2.38)	−0.51* (−.19)
Comprehension	170	3–17	9.53 (2.67)	9.64	−1.21	8.43	1–16	7.72 (2.91)	−.71* (−.27)
Picture Completion	171	4–17	10.18 (2.58)	10.14	−0.05	10.19	1–18	9.30 (2.92)	−0.89* (−.34)
Picture Arrangement	170	1–18	10.50 (2.64)	10.27	−0.20	10.07	1–18	8.60 (3.04)	−1.48* (−.56)
Block Design	171	2–17	9.37 (2.85)	9.47	−0.06	9.53	1–19	8.41 (3.19)	−1.12* (−.39)
Object Assembly	167	3–17	9.94 (2.84)	9.88	−0.32	9.56	2–16	8.73 (2.90)	−0.83* (−.29)
Coding	168	1–17	8.76 (2.98)	9.46	−0.67	8.80	1–14	7.50 (2.95)	−1.30* (−.44)

Note. WISC = Wechsler Intelligence Scale for Children, WISC-R = Wechsler Intelligence Scale for Children—Revised, PTS = predicted true score, EMC = estimated mean change, MIQ = modified IQ, WISC-III, Wechsler Intelligence Scale for children—III.

[a]ES = effect size, calculated as change divided by the obtained WISC-R standard deviation; * $p < .05$ or better.

scores of this LD population. However, there are differences between the respective IQ scores (change in FSIQ = −4.77, *ES* = −.45; change in VIQ = −2.93, *ES* = −.25; and change in PIQ = −6.68, *ES* = −.54), suggesting that the FE does not result in changes that are evenly distributed across cognitive abilities.

Comparison of the subtest scores also suggests that the FE is unevenly distributed. Two subtests scores, Information and Arithmetic, were not significantly different (.05 level) between versions of the Wechsler. The remaining comparisons were significantly different, although not evenly affected by the FE. For example, the mean difference attributable to the FE for Block Design (−1.12, *ES* = −.39) was about twice as large as that obtained for Vocabulary (−0.51, *ES* = −.19).

Gender and ethnicity analysis. The FE calculations were subjected to a repeated measures multiple analysis of variance using ethnicity and gender to determine whether there were any main or interaction effects in the FE for those groups. Results of the between-group differences are not reported because they are not pertinent to the question. Results for the within-group differences are reported in Table 3. No significant differences were evident by ethnicity, gender, or the interaction of ethnicity and gender. These results are presented solely as a preliminary investigation of gender and ethnicity differences because some of the groups in this sample were quite small (e.g., *n* = 12 for African American females).

DISCUSSION

The results of this research clearly support the existence of the FE in the LD population. Most Wechsler scores changed significantly between versions of the test (WISC-R and WISC-III), even after the initial WISC-R

Table 3
Flynn Effect on Wechsler VIQ, PIQ, and FSIQ by Ethnicity, Gender, and the Interaction for Sample 2 (WISC-R to WISC-III)

	FSIQ		VIQ		PIQ	
Source	*df*	*F*	*df*	*F*	*df*	*F*
Overall	1	49.56*	1	11.57*	1	66.03*
Gender	1	0.004	1	1.282	1	1.375
Ethnicity	1	0.038	1	0.003	1	0.003
Gender × Ethnicity	1	0.000	1	0.051	1	0.247
Error	124	(61.38)	124	(86.88)	124	(94.56)

Note. Each main effect was tested while holding other main effects constant; * $p < .001$.

scores were adjusted for regression to the mean, measurement error, and expected 3-year changes. This is an important finding because IQ is a primary determinant for the LD classification, and this study documents that a student's IQ score is influenced substantially by a factor external to the child. It is also important to note that for this sample, the calculated adjustments (PTS and EMC) functionally reduced the difference between the comparison scores. Without such adjustments, the differences would have been greater and the change attributed to the FE would have been greater. Thus, the reported differences are relatively conservative.

Differences in distribution of the FE between broad cognitive abilities were also evident. Scores on the PIQ dropped about twice as much as those on the VIQ (−6.68, *ES* = −.54 vs. −2.93, *ES* = −.25, respectively). This is consistent with previous research that reports substantial differences between the FE on verbally-loaded IQ tests such as the WISC and nonverbal IQ tests such as the Ravens (e.g., Flynn, 1999; Graf & Hinton, 1994).

However, Flynn's (1999) suggestion that the differences exhibited on different tests are attributable to varied effects on fluid versus crystallized cognitive ability was not conclusively supported by the research reported here. When the subtests that are reported to measure these abilities (Kaufman, 1994) were examined, there were no clear differences. For example, fluid ability is reported to be a common ability for Similarities, Arithmetic, Picture Arrangement, Block Design, and Object Assembly (Kaufman, 1994). In this research, Arithmetic did not exhibit a significant drop across versions of the Wechsler. The other reported fluid ability subtests changed only about 1 point, with the exception of Picture Arrangement. For the crystallized ability subtests, Information did not change significantly from test to test. Similarities changed about 1 point (−.44 *ES*), Comprehension changed about 0.7 of a point (−.27 *ES*), Picture Arrangement changed almost 1 1/2 points (−.56 *ES*), and Vocabulary only about 0.5 of a point (−.19 *ES*). One potential confound is that the subtests are not pure measures of the fluid/crystallized construct. For example, Picture Arrangement, which was the subtest exhibiting the greatest decrease between versions of the Wechsler, reportedly measures both fluid and crystallized abilities. This finding is consistent with previous reports that the Wechsler subtests are inadequate measures of fluid intelligence (McGrew & Flanagan, 1998).

Another possible interpretation of why the FE affects some scores more than others does not rely on fluid versus crystallized cognitive theories. It is possible that scores for school-based learning tasks (such as the arithmetic and information subtests) are not as influenced by the FE. This interpretation is consistent with information about the lack of the FE for achievement tests (Neisser, 1998) and avoids reliance on fluid versus crystallized ability theory, which is not universally accepted. J. R. Flynn has recently modified his position about fluid versus crystallized abilities, and posits that the difference in effect may be that tasks that are "school relevant" show less FE (personal communication, August 26, 1999).

The lack of a significant difference in the FE between African Americans and Caucasians in the second analysis is somewhat surprising. Flynn (1987) reported substantial differences in the phenomenon between different nations (e.g., 8 IQ points per decade in Japan vs. 3 points in the United States). Those findings suggest that differences in ethnicity and/or cultural groups within countries are also quite possible. In contrast, the relatively consistent and persistent difference in IQ between African Americans and Caucasian Americans suggests that the FE may affect scores for both groups to about the same degree over time (Flynn, 1998a). The lack of a significant difference reported here is, however, a preliminary finding. This sample contained only two groups and a relatively small number of African Americans ($n = 38$). Larger and more diverse samples might produce different results. Variations in the FE for different cultural groups in the United States remains a distinct possibility. The lack of differences between gender is less surprising because it is consistent with Flynn's (1998b) report of significant but small differences between males and females with a much larger sample of Israeli adults.

This research clearly supports Flynn's (1999) contention that the FE has significant ramifications for the identification of students for special education, particularly regarding the LD classification. The results show that the FE variably affects component scores of the Wechsler including the VIQ, PIQ, and subtests over time. This is an important finding because it serves as another reason that practitioners should be wary of subtest analysis. The cause of the FE is not known (Flynn, 1999) and it is entirely possible, as Flynn suggests, that it is simply an artifact that has little bearing on real world performance. If future research results are similar to this study, then students' subtest score patterns may change between test versions on broad-based IQ measures as a result of the FE, rather than as a true measure of the students' abilities. Similarly, the finding that the FE influences students' VIQ and PIQ scores differently suggests that the likelihood of finding Verbal–Performance discrepancies changes over the life of a test version and between test versions.

A critical finding of this study is that the FE probably contributes to misdiagnosis of LD. If this research is combined with previous reports that academic achievement may be unaffected by the FE (Neisser, 1998) it strongly suggests that, over the life of a test version, IQ-achievement discrepancies, the most salient LD criterion, are exaggerated. One potential result of such an exaggeration of IQ-achievement discrepancies would be that, as test norms aged, fewer students would score in the mentally retarded range (Flynn, 2000) and more students would qualify for LD based on inflated severe discrepancies. The variable subtest scores, changes in Verbal–Performance differences, and exaggerated IQ-achievement discrepancies attributable to the FE could all have substantive influence on the classification of LD.

Implications

This research has clear implications for practitioners. LD classification that relies on Wechsler IQ scores is increasingly influenced by the FE over the life of a test version. This influence appears to change the VIQ, PIQ, and subtest scores in ways that may not be related to true cognitive ability. The result may be increasing numbers of students who have higher IQ scores, larger Verbal–Performance differences, and altered subtest score patterns. If previous reports stating that academic achievement is unaffected by the FE are true (e.g., Neisser, 1998) then IQ-achievement discrepancies will also be exaggerated. All of these increase the likelihood of false-positive classification as LD. That is, the FE makes it more likely that students will meet requirements for LD classification as test norms age. When a test is revised, however, all of these influences of the FE will also immediately disappear. Such differences between test versions create confusion for parents, teachers, and multidisciplinary teams, as students may no longer meet the criteria for LD simply because the IQ test norms were updated. Further, no simple adjustment to the IQ score as test norms age can rectify the problem. This and other research suggests that the FE is not uniformly distributed by task or over time, and its effect for other potentially important variables (e.g., ethnicity and age) are unknown. These confounding variables and unknowns make simple adjustments impossible. However, it is clear that using IQ tests with current and recent norms is much preferable to using tests with either obsolete or aging norms. This *strongly* suggests that test publishers should renorm cognitive measures much more frequently than they have in the past.

The implications for research are also substantial. Aging test norms make comparisons to previous research based on IQ less accurate because the metric (IQ) changes over time. Also, because the FE appears to influence scores differently according to task and over time, and the effect for other variables is unknown, comparisons between groups becomes perilous as test norms age.

One clear implication of this study is that much more research needs to be done. This research clearly indicates that the FE variably influences Wechsler scores for students classified as LD, but the exact nature of that influence is not entirely clear. For example, more research with larger sample sizes would allow for comparison across age ranges, which some previous related research suggests are important (Thorndike, 1975). The potential influences of ethnicity differences also should be explored further with larger and more diverse samples. The effect of the phenomenon on other tests is also not clear, and Flynn's assertion of differential influence on fluid and crystallized abilities could be better examined with other instruments.

Limitations

There are several limitations to this study. Although the sample is representative of students classified as LD in NYS, it suffers from the lack of a clear uniform definition for LD, as does much of the previous research on this population. This research did not make distinctions between various possible subtypes of LD. For example, this research did not differentiate between the normal IQ/low achievement group and the low IQ/low achievement group. However, reporting the number of students who exhibited a substantial discrepancy between IQ and achievement helped to establish some idea about the homogeneity of the group. Using triennial evaluation information for comparison mirrors actual practice; however, it makes an adjustment for the 3-year time span necessary. Using existing information from triennial evaluations also means that the students' histories were not controlled. A better method would be to administer both the WISC-R and WISC-III at approximately the same time. The use of different versions of the WISC is also, however, a limitation. Although the WISC-R and WISC-III are very similar, it seems plausible that at least some of the variation between the scores is due to differences between the tests rather than differences between the normative performances. An additional limitation is that although the sample size used here is adequate for this investigation, a much larger sample size from a larger geographic region would give greater confidence that the results apply to the U.S. LD population as a whole.

ACKNOWLEDGMENTS

The authors wish to thank the school districts that provided data for this study, James McDougal for helping collect the data for the previous study, and John Rosenbach, who influenced our decision to research the Flynn effect. We also thank J. R. Flynn for reviewing an earlier version of the manuscript.

REFERENCES

Atkinson, L. (1991). Three standard errors of measurement and the Wechsler Memory Scale-Revised. *Psychological Assessment: Journal of Consulting and Clinical Psychology, 3,* 136–138.

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

Canivez, G. L., & Watkins, M. W. (1998). Long-term stability of the Wechsler Intelligence Scale for Children-Third Edition. *Psychological Assessment, 10,* 285–291.

Carlton, M., & Sapp, G. L. (1997). Comparison of WISC-R and WISC-III scores of urban exceptional students. *Psychological Reports, 80*, 755–760.

Edelman, S. (1996). A review of the Wechsler Intelligence Scale for Children-Third Edition (WISC-III). *Measurement & Evaluation in Counseling & Development, 28*, 219–224.

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

Flynn, J. R. (1985). Wechsler intelligence tests: Do we really have a criterion of mental retardation? *American Journal of Mental Deficiency, 90*, 236–244.

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

Flynn, J. R. (1998a). IQ gains over time: Toward finding the causes. In U. Neisser (Ed.), *The rising curve: Long-term gains in IQ and related measures.* Washington, DC: American Psychological Association.

Flynn, J. R. ((1998b). Israeli military IQ tests: Gender differences small, IQ gains large. *Journal of Biosocial Science, 30*, 541–553.

Flynn, J. R. (1999). Searching for justice: The discovery of IQ gains over time. *American Psychologist, 54*, 5–20.

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

Gaskill III, F. W., & Brantley, J. C. (1996). Changes in ability and achievement scores over time: Implications for children classified as learning disabled. *Journal of Psychoeducational Assessment, 14*, 220–228.

Graf, M. H., & Hinton, R. N. (1994). A 3-year comparison study of WISC-R and WISC-III IQ scores for a sample of special education students. *Educational and Psychological Measurement, 54*, 128–133.

Kaufman, A. S. (1993). King WISC the Third assumes the throne. *Journal of School Psychology, 31*, 345–354.

Kaufman, A. S. (1994). *Intelligent testing with the WISC-III.* New York: Wiley.

Lynn, R., & Hampson, S. (1986). The rise of national intelligence: Evidence from Britain, Japan and the U.S.A. *Personality and Individual Differences, 7*, 23–32.

McGrew, K. S., & Flanagan, D. P. (1998). *The intelligence test desk reference (ITDR): Gf-Gc cross-battery assessment.* Boston: Allyn & Bacon.

Naglieri, J. A., & Pfeiffer, S. I. (1983). Reliability and stability of the WISC-R for children with below average IQs. *Educational and Psychological Research, 3*, 203–208.

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

Oakman, S., & Wilson, B. (1988). Stability of WISC-R intelligence scores: Implications for 3-year reevaluations of learning disabled students. *Psychology in the Schools, 25*, 118–120.

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

Spitz, H. H. (1989). Variations in Wechsler interscale IQ disparities at different levels of IQ. *Intelligence, 13*, 157–167.

Stavrou, E. (1990). The long-term stability of WISC-R scores in mildly retarded and learning-disabled children. *Psychology in the Schools, 27*, 101–110.

Thorndike, R. L. (1975). Mr. Binet's test 70 years later. *Educational Researcher, 4*, 3–7.

Truscott, S. D., Narrett, C. M., & Smith, S. E. (1994). WISC-R subtest reliability over time: Implications for practice and research. *Psychological Reports, 74*, 147–156.

University of the State of New York, State Education Department. (1993). *Regulations of the Commissioner of Education. Subchapter P. Part 200.* Albany, NY: SUNY.

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

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

Wechsler, D. (1991). *Manual for the Wechsler Intelligence Scale for Children—Third Edition.* New York: The Psychological Corporation.