Journal of Psychoeducational Assessment
2003, *21*, 145-159

DOES THE FLYNN EFFECT DIFFER BY IQ LEVEL IN SAMPLES OF STUDENTS CLASSIFIED AS LEARNING DISABLED?

Kathryn J. Sanborn
Stephen D. Truscott
LeAdelle Phelps
The University at Buffalo, SUNY

James L. McDougal
Oswego State University, SUNY

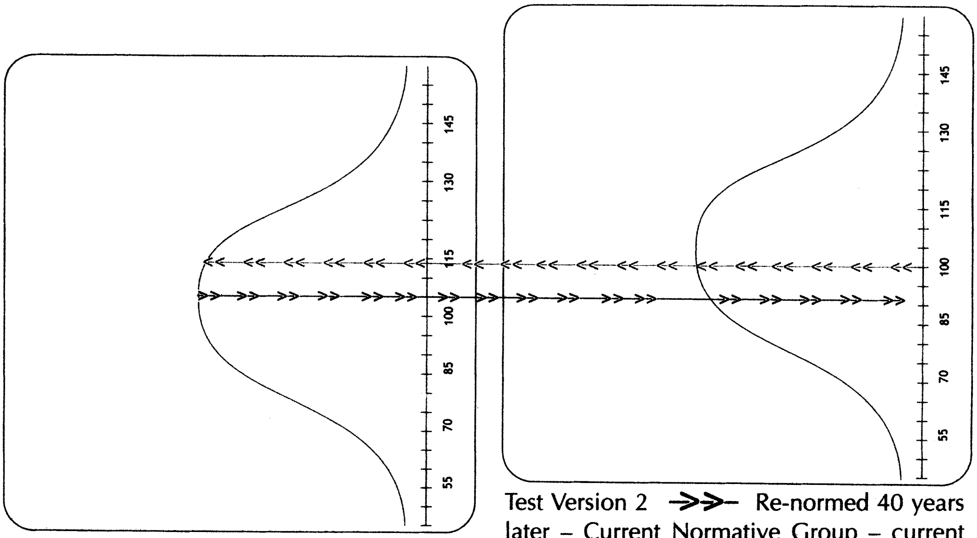
This research examined scores from two learning-disabled (LD) samples for evidence of the Flynn Effect (FE; Flynn, 1999) to determine (a) whether the FE was evident, and (b) if the magnitude of the FE differed on IQ scores at varying levels of intelligence. Sample 1 consisted of 40 children who were administered the Woodcock-Johnson Tests of Cognitive Ability–Revised (WJ-R Cog) and the Woodcock-Johnson Tests of Cognitive Ability–III (WJ-III Cog) approximately 8 weeks apart. Sample 2 consisted of triennial test data for 169 students who were tested once with the Wechsler Intelligence Scale for Children–Revised (WISC-R) and 3 years later with the Wechsler Intelligence Scale for Children–Third Edition (WISC-III). After an initial analysis for the overall FE, Samples 1 and 2 were each divided into IQ level groups (e.g., 91-105) and analyzed accordingly. Results for Sample 1 (W-J) indicated that the 2+ point observed IQ change was not significant for this relatively small sample, although it appeared to be consistent with the direction and degree predicted by the FE (Flynn, 1984). Results for Sample 2 (WISC) indicated a significant IQ change consistent with the FE. Neither sample exhibited statistically significant differences for FE by IQ level. However, consistent with previous research, observed differences increased from lower to higher IQ levels. Consequently, this research cannot rule out the possibility that a child's IQ level influences the degree to which the FE is apparent.

In the majority of public schools, IQ is an integral component in the diagnosis of learning disabled (LD) and other special education classifications. Over the past 40 years, an increasing number of researchers have documented a general decrease in participants' IQ scores as various IQ tests are revised and those people are compared to a new normative group (Bolen, Aichinger, Hall, & Webster, 1995; Carlton & Sapp, 1997; Gaskill & Brantley, 1996; Graf & Hinton, 1994; Lynn & Hampson, 1986; Slate & Saarino, 1995; Spitz, 1989; Thorndike, 1975; Truscott & Frank, 2001; Truscott, Narrett, & Smith, 1994; Vance, Maddux, Fuller, & Awadh, 1996). These decreases in IQ scores have been reported on a variety of cognitive measures such as the Raven Progressive

Correspondence concerning this article should be addressed to Stephen D. Truscott, Department of Counseling and Educational Psychology, The University at Buffalo, SUNY, Buffalo, NY 14260; e-mail: sdt55@acsu.buffalo.edu.

Matrices and the ever-popular Wechsler batteries.

James Flynn (1984, 1987) offered the most widely accepted description of this phenomenon, referred to as the "Flynn Effect." He detailed a consistent and worldwide increase in the normative performance on intelligence tests for at least the last 60 years. This consistent improvement in the normative performance on IQ measures results in a counterintuitive effect on an individual's IQ score. If IQ is increasing, why are individual IQ scores decreasing on more recently normed tests? The reason is that the Flynn Effect (FE) affects the normative performance of the standardization samples used to norm IQ tests. Standardization samples are representative of the population during the time period in which the sample was selected and tested. Therefore, if the normative IQ is increasing, the standardization samples will reflect this improved performance. Any given individual's absolute performance on the tasks that comprise the test will score (a) higher when compared to older norms that reflect a lower normative performance and (b) lower when compared to a current normative performance that reflects greater population-wide abilities. This means that if the same group of people were given two IQ tests, one normed in 1970 and the other normed in 1990, they would receive higher scores on the test normed in 1970. Thus, their individual scores between the two tests would drop from the 1970 version to the 1990 version because they are being compared against a "smarter" sample in the 1990 version. In other words, the representative samples for norming IQ tests consistently establish norms of a higher



Test Version 1 <<— Original Normative Group – original average performance

Test Version 2 —>>— Re-normed 40 years later – Current Normative Group – current average performance

FIGURE 1. Visual representation of the Flynn Effect using two versions of a test, normed 40 years apart.

standard as compared to previous samples. This relationship is depicted visually in Figure 1. Tests are restandardized specifically because their norms become obsolete and therefore fail to adequately represent the current population. Because intelligence tests are regularly renormed, reestablishing the standardization score to a higher level can mask a long-term increase in population IQ. This higher mean population-wide performance causes a decrease in any individual's score between administrations of two IQ tests normed at different times, independent of any innate cognitive abilities. Without attention to the FE, the individual's observed drop in scores can be attributed incorrectly to a cognitive failure when, in actuality, it may simply reflect differences in the normative performance on the test over time.

Several explanations have been offered for the phenomenon but none is widely accepted. These possible explanations range from theories related to increasing test sophistication in the population to improved nutrition and increasing participation in formal education. Until recently, Flynn (1998) had refuted most of the explanations offered to date. However, Flynn (2001) now postulates that IQ gains can be divided into pre-1950 and post-1950 causes. He has stated that pre-1950 gains in IQ can be attributed primarily to growth in years and quality of formal schooling. Post-1950 gains are caused by a myriad of social trends, which include reduced family size, more leisure time, and new leisure activities among several other factors. Together these social trends have given rise to the "social multiplier" effect in which the population mean for IQ has significantly increased. According to Flynn, the rising mean itself has become the primary causal factor contributing to IQ gains since 1950.

Although determining the cause of the FE is important, a more salient objective for school-focused researchers is to understand how the FE might affect the scores of children who are tested for special education eligibility. Surprisingly, there is little research to date that addresses this critical concern. There are only a handful of studies that have directly addressed how the FE might change students' scores on crucial special education classification criteria. Some research has established that, for children labeled as LD, the FE likely affects special education classification because it alters IQ scores that are central to the primary classification criteria (Truscott & Frank, 2001). These researchers examined triennial data of 171 LD students who were initially tested with the WISC-R and then with the WISC-III and found that WISC-III Full Scale IQ (FSIQ), Verbal IQ (VIQ), and Performance IQ (PIQ) scores were all significantly lower than those same scores derived from the WISC-R. However, scores on tests that relied heavily on school-based learning were generally less affected than those that tested more abstract abilities.

Two studies that included children focused on how the FE may affect IQ scores at varying levels of intelligence. Spitz (1989) examined a variety of data from Wechsler measures and found that "for subjects who had IQ's in the average and borderline range, the mean WAIS IQs were, respectively, 9.09 and 4.75 points higher than the mean WAIS-R IQ" (p. 160). However, he reported that the FE diminished as IQ fell below the average/borderline range. Graf and Hinton (1994) examined a sample of 84 regular and special education students. This sample was originally administered a WISC-R and then, 3 years later, administered a WISC-III. In the groups whose IQ levels were average (91-105) or above average (106-120+), the FE was evident because WISC-R scores were consistently, but not always statistically significantly, higher than WISC-III scores. For the groups whose IQ fell in the below-average range (60-90), this phenomenon actually reversed where WISC-III scores were higher than WISC-R scores. Thus, two research studies have reported that the FE is either diminished or nonexistent at the lower ranges of IQ. Flynn (personal communication, August 26, 1999) disputed Spitz's (1989) findings, attributing most of the inconsistency noted by Spitz to changes in the normative samples used for the WAIS and WAIS-R. Flynn has not addressed the similar finding from Graf and Hinton (1994).

The finding of potentially different FE at varying levels of IQ has also been evident in some large-scale studies with normal populations. In archival research of Danish Draft Board data regarding male IQ scores from two distinct time periods, Teasdale and Owen (1989) found that the gains in IQ scores over a 30-year period were greatest in the lower IQ levels. This finding is in opposition to the findings of both Spitz (1989) and Graf and Hinton (1994). Lynn and Hampson (1996) found evidence similar to Teasdale and Owen (1989), supporting the FE in groups with IQs in the lower ability ranges from British samples.

It is important for researchers to investigate the FE phenomena. The current research focuses on the magnitude of the FE on IQ scores at varying levels of intelligence commonly found in the LD identified population. This is particularly important for this population given that an IQ/Achievement discrepancy is often the most weighted criterion used by school districts to decide on LD classification and services. For the current study, two samples were utilized. Sample 1 consisted of children given two differently normed versions of the Woodcock-Johnson Tests of Cognitive Abilities (WJ-COG). Sample 2 consisted of children administered two differently normed versions of the Wechsler Intelligence Scale for Children (WISC). In a previously published study (Truscott & Frank, 2001), scores in Sample 2 had been identified to significantly demonstrate the FE. Thus, the first purpose of this study is to determine whether the FE is present in Sample 1. The second purpose is to add to the paucity of studies to help determine whether the FE variably affects scores at different IQ levels. In other words, is a child with a higher IQ or lower IQ more susceptible to the FE (gains in IQ)? This objective will be tested in both Sample 1 and Sample 2.

METHODS

Participants

Sample 1: *Woodcock Johnson-Cognitive.* Initial participants in Sample 1 consisted of 47 children and adolescents who attended six school districts located in western New York State. The districts were located in suburban (4), city (1), and rural (1) communities. From the initial sample of 47, the scores for 40 students whose Woodcock Johnson-Revised Cognitive (WJ-R Cog) IQ fell between 75 and 105 were selected for analysis in this research.

Fourteen (35%) of the students were female. The average age and grade placement at the first evaluation was 12.40 (*SD* = 2.03) and 6.44 (*SD* = 2.05), respectively. Nine of the students attended high schools, 15 were in middle schools, and 16 attended elementary schools. Prior to the onset of this study, all students in Sample 1 were classified as learning disabled (LD) by their school districts according to the New York State regulations. To decrease the effect of differences in classification practices between school districts, the researchers imposed the additional requirement that each participant must have had at least a 15-point standard score difference between ability and achievement on his or her most recent previous evaluation. Records from the 40 students with available achievement test scores reveal that all of the students had IQ-achievement discrepancies in a literacy-related area: Reading only = 3 (7%); Written Language only = 2 (5%); Reading and Written Language = 12 (30%); Reading and Math = 3 (7%); Math and Written Language = 5 (13%); Reading, Math, and Written Language = 15 (38%). Using the greatest discrepancy between IQ and achievement, the average discrepancy was 29.16 points (*SD* = 11.53).

The participants were tested with the Woodcock-Johnson Tests of Cognitive Ability–Revised (WJ-R Cog) and the Woodcock-Johnson Tests of Cognitive Ability–III (WJ-III Cog) approximately 6 to 8 weeks apart. Using data collected from this study, the average initial WJ-R Cog IQ and WJ-III Cog IQ was 89.32 (*SD* = 6.98) and 87.27 (*SD* = 11.27), respectively.

Sample 2: Wechsler Intelligence Scale for Children. Sample 2 consisted of archival records of 169 students from 20 school districts in New York State. Consistent with Sample 1, all students were previously classified as learning disabled (LD) by their corresponding school districts. Participants in Sample 2 were administered the WISC-R once and then reevaluated with the WISC-III approximately 3 years later. Of the 169 students in Sample 2, 33 % (55) were female. The mean age and grade at the first testing was 9.21 years (*SD* = 1.73) and 2.8 (*SD* = 1.59), respectively. The average WISC-R IQ at the first testing was 94.90 (*SD* = 8.32). The average WISC-R VIQ and PIQ was 92.50 (*SD* = 9.76) and 98.40 (*SD* = 10.02), respectively. Although there was no control for differences in classification procedures, the sample was relatively homogeneous on the important LD criteria of a significant difference between IQ and achievement. Of the 155 records with IQ and achievement data that could be used to calculate an IQ-achievement discrepancy, nearly 75% of the students exhibited a 15-point or greater standard score difference between IQ and achievement on their most recent evaluation (mean discrepancy = 28.09, *SD* = 12.25); almost all of which were literacy related.

Measures

Sample 1: Woodcock Johnson-Cognitive. The Standard Cognitive Batteries from the latest two version of the WJ-Cog—WJ-R Cog (Woodcock & Johnson, 1989, 1990) and WJ-III Cog (Woodcock & Johnson, 2001)—were used with Sample 1. The test versions are both carefully constructed and normed measures of cognitive ability, with excellent psychometric properties. Both are constructed from similar contemporary theoretical models and are considered good measures of general IQ.

To measure the FE for this study, the researchers had to assume that both measures of IQ in Sample 1 (WJ-R Cog and WJ-III Cog) measured the same construct. However, despite the continuation of the name, the newest revision is a substantially different test than its predecessor. Because this difference could potentially impact the significance of the results of Sample 1, a brief comparison of the WJ-R and WJ-III is justified. Since each subtest in both the WJ-R Cog and the WJ-III Cog was created using item response theory (IRT), the tests will be compared according to subtest similarity. When a test is formed using IRT, its construction process is based upon having a collection of items from which to choose. Such collections of items are known as item pools. All of the items in a particular item pool are constructed and tested to measure the same trait (Baker, 2001).

The WJ-III Cog Standard Battery contains 10 subtests, whereas the WJ-R consists of only 7 subtests. However, the Visual-Auditory Learning Delayed subtest on the WJ-III Cog was not administered for these analyses. Therefore, when comparing the WJ-III Cog with the WJ-R Cog, the tests contain 9 and 7 subtests, respectively. Based upon the researchers' calculations, out of a total of 9 administered subtests on the WJ-III Cog, 2 of the subtests (Visual Matching and Incomplete Words) are directly comparable to subtests found on the WJ-R Cog. Because these 2 subtests were created using item response theory, the items can be considered 100% alike. A third subtest on the WJ-III Cog, Verbal Comprehension, is split into four different tasks. Only one of these tasks is included in the WJ-R Cog (Picture Vocabulary). In addition, on the WJ-R Cog, Picture Completion is a separate subtest and is not grouped with other tasks. Therefore, when computing the similarity of the tests, the percentage of Picture Completion items in Verbal Comprehension (WJ-III Cog) was computed (32.3%).

According to these calculations, the WJ-III Cog and the WJ-R Cog contain 2.32 subtests that are 100% alike. Consequently, based upon analyses of subtest content, the overall WJ-III Cog (minus the Visual-Auditory Learning Delayed subtest) contains 25.7% of the same items found in the WJ-R Cog.

Sample 2: Wechsler Intelligence Scale for Children. The study employing Sample 2 used the previous two revisions of the WISC—WISC-R (Wechsler, 1974) and WISC-III (Wechsler, 1991). According to Wechsler (1991), correlation coefficients for the two tests are high, averaging .89, .90, and .81 for the Full Scale (FS), Verbal (V), and Performance (P) Scales, respectively. In addition, the WISC-R and WISC-III share many of the same items, attesting to their overall high degree of similarity. Only about one third of the items on the WISC-III are new or modified (Kaufman, 1993), with approximately 73% of the items being the same on both tests (Edelman, 1996). Additional items were added to the WISC-III to improve discrimination for exceptional students. This resulted in the inclusion of floor and/or ceiling items for five subtests, including Similarities, Arithmetic, Picture Arrangement, Block Design, and Mazes. As supported by several researchers, the tests are mostly the same and are highly correlated (e.g., Bolen et al., 1995; Carlton & Sapp, 1997; Gaskill & Brantley, 1996; Graf & Hinton, 1994).

Procedures

Sample 1: Woodcock Johnson-Cognitive. Sample 1 participants were selected by district school psychologists using the previously described criteria (LD with a 15-point standard score difference between IQ and achievement). Graduate school psychology students were trained in the administration of the WJ-R and WJ-III Cog tests, and they then tested the participants in their home school buildings. Data were recorded on a form that was coded so the researchers could not identify participants. The local school kept original test protocols. All students were administered both the WJ-R Cog Standard Battery and the WJ-III Cog Standard Battery. A period of about 8 weeks (M = 8.38, SD = 3.15) separated the administrations to minimize the impact of practice effects. Tests were administered in counterbalanced order to decrease test order effects. Twenty-one subjects were first administered the WJ-R Cog followed by the WJ-III Cog. The remaining 19 students were administered the WJ-III Cog and then the WJ-R Cog. Students were given the entire Standard Battery of both tests with the exception of the Visual-Auditory Learning Delayed subtest of the WJ-III Cog, which does not contribute to Broad Cognitive Ability.

After completion of testing, the 47 original subjects were grouped according to Broad Cognitive Ability (BCA) as determined on the WJ-R Cog. The groups were formed in accordance with Graf and Hinton's (1994) study that separated subjects based on IQ. Graf and Hinton (1994) did not specify their rationale for splitting the groups in the manner in which they did. However, the mean IQ of Sample 1 in our study (using WJ-R Cog scores) was about 90. Thus, the Graf and Hinton groups represent one *SD* above and below the mean of our Sample 1. In addition, the relevant categories have some intuitive appeal and we used the same grouping strategy to allow comparison of our results with their research. Because Sample 1 did not have as wide a range of IQ as the sample studied by Graf and Hinton (which included general education and students labeled as MR), two groups were formed. The first group consisted of 22 subjects whose BCA (WJ-R Cog) fell within the range from 76 to 90. The second group consisted of 18 subjects whose BCA (WJ-R Cog) fell within the range from 91 to 105. Seven subjects from Sample 1 were not used in this analysis because their BCA fell outside these two IQ groups.

The Predicted True Score (PTS; Atkinson, 1991, see below) was calculated to statistically adjust for regression to the mean and the reliability of the initial test score. The adjusted WJ-R BCA was then compared with the WJ-III General

Intellectual Ability (WJ-III GIA) to determine if the FE influences IQ groups differently.

Sample 2: Wechsler Intelligence Scale for Children. Sample 2 employed a longitudinal archival design using extant school records. School employees were asked to select students who were (a) classified as LD and (b) had existing data from both the WISC-R and the WISC-III. Employees of the respective schools, usually the school psychologist, created a code for each student and recorded the data on forms provided by the researchers. The sample consisted of students classified as LD who had been administered first the WISC-R, then the WISC-III approximately 3 years later (M = 3.36; SD = .991). Unlike Sample 1, the Wechsler tests could not be counterbalanced due to the longitudinal archival design. As in Sample 1, the PTS was calculated for the first administration (WISC-R) to statistically adjust for regression to the mean and the reliability of the initial test score. Due to the 3-year time span between test administrations, some researchers have also calculated the average IQ change attributable to this time difference (estimated mean change; EMC) (Truscott & Frank, 2001). However, 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). Therefore, in order to be consistent with the study using Sample 1, the EMC was not calculated.

Due to the larger sample size and range of WISC-R scores, three groups were formed based upon WISC-R Full Scale (FS) scores. Groups 1 and 2 corresponded with Sample 1. A third group was formed that contained students who obtained a FS score from 106 to 120 on the WISC-R. For each IQ group, the FS score adjusted for the PTS was compared with the obtained WISC-III score for evidence of the FE according to IQ group.

Explanation of Predicted True Score Calculation

Obtained scores are subject to two influences that, independent of the participants' "true" abilities, can result in score variations when compared to subsequent scores. These influences are regression to the mean and the reliability of the test. To account for score changes that might be attributed to statistical regression to the mean and test reliability issues, the PTS was calculated for both Sample 1 and Sample 2. The PTS produces a modified score that is more suitable for comparison than simple obtained scores. To calculate the PTS for each student's score on either the WJ-R Cog or the WISC-R, the following procedure was used: "(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).

The PTS produces a modified score that is more suitable for comparison than simple obtained scores. Because the WJ-Cog and WISC are highly reliable and the range of scores included in these samples is relatively restricted, the PTS calculation resulted in only small adjustments to the scores.

RESULTS

Sample 1: Woodcock Johnson-Cognitive

To compare the differences between BCA (WJ-R Cog) and GIA (WJ-III Cog) for each of the two IQ groups, a two-way repeated measures analysis of variance (ANOVA) was performed. The dependent variable was the students' overall IQ scores on the WJ-R Cog and the WJ-III Cog. The two independent variables were test administration (either the WJ-R or WJ-III) and IQ group (1 or 2).

Results for the overall Flynn Effect by IQ group for Sample 1 are reported in Table 2. Mean differences in the between-group comparisons of the WJ-R Cog and the WJ-III Cog were not statistically significant. Overall, 23 students' IQ scores decreased (mean change = 7.52 pts.) and 17 students' scores increased (mean change = 6.76 pts.). The interaction of test administration and IQ group was also not significant. This finding suggests that the FE did not affect IQ nor did it affect the IQ groups differently.

Table 1
Descriptive Data for Sample 1 and Sample 2

	Sample 1			Sample 2			
	Group 1	Group 2	Total	Group 1	Group 2	Group 3	Total
	(76-90)	(91-105)		(76-90)	(91-105)	(106-120)	
N	22	18	40	68	80	21	169
Gender (girls)	6	8	14	29	22	4	55
Age (*SD*)	12.63 (2.38)	12.13 (1.55)	12.40 (2.03)	9.47 (1.66)	9.24 (1.83)	8.31 (1.32)	9.21 (1.73)
Grade (*SD*)	6.67 (2.48)	6.17 (1.47)	6.44 (2.05)	2.76 (1.54)	2.88 (1.70)	2.33 (1.49)	2.76 (1.59)
WJ-R Cog (BCA)	84.02 (4.12)	95.80 (3.15)					
WJ-III Cog (GIA)	83.13 (10.99)	92.38 (9.60)	87.75 (11.27)				
WISC-R							
Full Scale				86.85 (2.53)	96.72 (4.02)	110.51 (3.04)	95.95 (8.32)
Verbal				85.83 (6.75)	93.89 (6.75)	108.74 (5.67)	92.49 (9.76)
Performance				90.92 (7.16)	101.15 (7.30)	109.79 (5.53)	98.40 (10.02)
WISC-III							
Full Scale				80.77 (8.32)	88.91 (8.51)	101.10 (7.58)	87.39 (11.08)
Verbal				80.50 (10.23)	87.14 (8.74)	99.15 (6.91)	85.98 (10.87)
Performance				83.75 (10.94)	92.78 (11.15)	102.96 (13.09)	90.73 (13.26)

Table 2
Analysis of Variance for WJ-R and WJ-III Broad Cognitive Comparison

Source	*df*	*F*	*p*
	Between subjects		
Test Administration (A)	1	1.53	0.22
IQ Group (B)	1	36.16*	0.00
A X B	1	.50	0.48
Within-group error	76	(60.81)	

Note.—Values enclosed in parentheses represent mean square errors.
* $p < .05$.

Table 3
Difference Scores on the WJ-R and III, and WISC-R and III by IQ Groups

Test	Group 1 (76-90)	Group 2 (91-105)	Group 3 (106-120)	Total
WJ (WJ-R – WJ III)	-.930	-3.411		-2.170
WISC (WISC-R – WISC-III)				
FS	-6.388	-7.805	-9.367	-7.429
V	-5.337	-6.742	-9.505	-6.521
P	-7.173	-8.278	-6.892	-7.662

Even though statistically significant differences were not found for Sample 1, there was a noticeable decrease in observed scores from the WJ-R Cog to the WJ-III Cog. This finding suggested that a closer inspection of the data was warranted. Table 1 shows the mean IQ scores for both administrations, whereas Table 3 shows the overall IQ difference scores between administrations of the tests. As seen in Tables 1 and 3, test IQ scores dropped a little over 2 points from the WJ-R Cog administration to the WJ-III administration. Table 3 also shows the discrepancy between Group 1 (76-90) and Group 2 (91-105) with regard to mean change in IQ. Group 1 showed approximately a 1-point difference between test administrations, whereas Group 2 showed about a 3½-point difference between administrations. These observed differences are discussed later in this paper.

Sample 2: Wechsler Intelligence Scale for Children

To be consistent in statistical analyses with respect to Sample 1, a two-way repeated measures ANOVA was performed on the data from Sample 2. The goal of this analysis was to compare the WISC-R FSIQ to the WISC-III FSIQ to determine whether the FE was present and if it differentially affected the IQ according to Group 1 (76-90), Group 2 (91-105), and Group 3 (106-120). Table 4 reports the results for the overall FE by IQ group for Sample 2. Results of the between-group differences show that test administration (WISC-R to WISC-III) showed significant differences with an approximately –7.5 point difference in IQ. Overall, 134 students' IQ scores decreased (mean change = 9.32 pts.) and 31 students' scores increased (mean change = 4.44 pts.). However, the interaction of test administration by IQ group is not significant and therefore the mean change in IQ does not differ according to IQ group.

Table 4
Analysis of Variance Table for WISC-R to WISC-III FSIQ Comparison

Source	*df*	*F*	*p*
	Between subjects		
Test Administration (A)	1	91.96*	0.00
IQ Group (B)	2	211.39*	0.00
A X B	2	1.02	.363
Within-group error	332	(40.343)	

Note.—Values enclosed in parentheses represent mean square errors.
$* p < .05$.

Albeit not significant, an inspection of the mean difference scores by IQ group parallels the analysis performed in Sample 1. The mean Full Scale IQ scores of both administrations are reported in Table 1, and the mean difference scores are reported in Table 3. Observed results reported in Table 3 show that all IQ groups decreased in mean scores by at least 6 points.

A second analysis was performed in which a multivariate analysis of variance was performed on the WISC VIQ and PIQ. Results are reported in Table 5. Specifically, researchers wanted to know whether the mean change in VIQ and PIQ was different depending upon IQ range. Differences between test administration were significant for both VIQ and PIQ. However, interaction effects between test administration and IQ group were not significant. Therefore, the FE is noted to a significant degree in both VIQ and PIQ, but mean IQ difference scores do not vary according to IQ group. Table 1 reports the means and standard deviations of both administrations regarding Verbal and Performance scores. Table 3 reports the mean difference scores by IQ for VIQ and PIQ. All mean difference scores were negative, indicating that students consistently performed better on the WISC-R as compared with the WISC-III.

Table 5
Multivariate Analysis of Variance for WISC-R to WISC-III VIQ and PIQ Comparisons

	Multivariate			Univariate	
Source	*df*	*F*	*p*	Verbal	Performance
Test Administration (A)	2	59.247*	.000	76.659*	70.068*
IQ Group (G)	4	58.319*	.000	82.222*	52.892*
A X G	4	1.082	.366	1.762	0.310
Mean square within-group error				87.624	130.320

Note.—Multivariate *F* ratios were generated from Wilks's Lambda statistic.
**p* < .05.

DISCUSSION

The study investigated whether the FE was evident in two LD samples using two different measures of cognitive ability, and, whether the FE differentially affected scores of students at different IQ levels. Even though results for Sample 1 were not statistically significant, overall, this research presents clear evidence that the FE affects at least some IQ scores from students in the LD population. This is an important finding because IQ scores are a central diagnostic tool used to identify students with LD.

For Sample 1, which employed the Woodcock Johnson-R and -III versions of the cognitive ability test, the more than 2-point IQ difference between the test versions (WJ-R > WJ-III) was not statistically significant for this relatively small sample. However, this result must be interpreted with caution. The observed difference in scores is in both the direction and to nearly the degree expected from the FE. Because Flynn (1984) reported that the expected change in IQ in the U. S. is approximately 3 IQ points per decade and the versions of the WJ-Cog were revised and renormed about 10 years apart, the 2+ difference in scores evident in this research would be consistent with the FE if it was maintained over a larger sample size. Other researchers have found evidence of the FE with other cognitive tests, but not necessarily at the commonly accepted estimate of 3 points per decade (Cattell, 1951; Graf & Hinton, 1994; Truscott & Frank, 2001). Previous research and the practical importance of accounting for an extraneous influence that might systematically alter the IQ scores of LD students suggest that, despite the lack of statistical significance for Sample 1, it is safest not to overinterpret the nonsignificant results of this portion of the current research.

As demonstrated in previous research (Truscott & Frank, 2001), the FE was statistically present in the Wechsler IQ scores of the LD students in the larger Sample 2. Overall, scores from this group exhibited about a 7½-point difference between versions (WISC-R > WISC-III). Because the WISC versions were normed approximately 17 years apart (WISC-R, approximately 1972; WISC-III, approximately 1989), this corresponds to approximately a 4.2 point per decade increase in the normative performance on the Wechsler scale for this sample.

The second purpose of this study was to determine whether the FE differentially affects LD students' scores at different IQ levels. Results from both samples presented here were not statistically significant. This means that the FE did not differentially affect the scores for students in the IQ groups investigated here. Neither the WJ-Cog nor the WISC results were significantly different across the identified IQ groups—i.e., IQ groups were 76-90, 91-105, and 106-120 (WISC only). However, this result should not be interpreted as the definitive answer to the concern about potentially different FE affects by IQ level. A visual inspection of the results reveals that difference scores (Table 3) between the two administrations steadily increased from Group 1 (low IQ) through Group 3 (high IQ) for both samples. For example, the WJ-Cog data show about a 1-point difference for Group 1 (below the 3 points expected from the FE) and a 3½-point difference for Group 2 (about at the expected level). Thus, the visual inspection of data is consistent with previous researchers who reported that the FE is diminished at lower IQ ranges but follows the FE expectation at IQs above 90 (Graf & Hinton, 1994; Spitz, 1989), while the statistical analysis suggests that such differences are not significant. The visual inspection pattern for Sample 2 is similar. Given the potentially high stakes represented by reaching an incorrect conclusion about the effect of the FE on the scores of students whose diagnosis depends largely on scores, it seems prudent to simply conclude that there is not sufficient evidence at this point to determine this question one way or another. It is, however, clear that there is a serious need to answer this question using a large, representative sample of students with educational disabilities who have IQ scores that represent the range of cognitive functioning.

This research has several important implications for researchers and practitioners. Regardless of whether IQ gains as reported by Flynn (e.g., 1999) and the current study reflect actual increases in the population's cognitive abilities, increases in the normative performance on IQ tests over time mean that IQ tests must be standardized frequently. Otherwise, individuals whose test performances are scored against outdated, obsolete norms will obtain inflated IQ scores. Use of such obsolete norms will interfere with the accurate diagnosis of children thought to have a learning disability by making it more likely that the child will qualify as LD and less likely that the child will score in the mentally retarded range. In contrast, when a previously identified LD student is retested with a new version of an IQ test with up-to-date norms, the student's IQ will likely decrease, making it less likely that the student will continue to be eligible for LD services. For example, Gaskill and Brantley's (1996) analysis indicated that fewer children exhibited a 15-point discrepancy between IQ and achievement after being testing with the WISC-III than when they were tested with the WISC-R. Such differences in the likelihood and specifics of special education classification can cause distress and confusion for parents, multidisciplinary teams, teachers, and students.

The results reported here also suggest two important concerns for LD classification procedures. First, psychologists who use a discrepancy between IQ and achievement as a major component in LD eligibility decisions should probably use IQ and achievement tests that are normed concurrently, or at least at approximately the same time. Using differently normed tests for LD evaluations may inaccurately identify (or miss) students for special services based on the FE rather than intrinsic, child-related factors. Second, based on this and other research (Graf & Hinton, 1994; Truscott & Frank, 2001), there is increasing reason to suspect that LD diagnostic procedures that rely on discrepancy criteria are substantially affected by factors external to the child. This suggests that such criteria have been overemphasized and should be used with caution.

Although we did not find statistically significant differences in the FE by IQ level, the potential for this confound should not be ruled out. Other research (Graf & Hinton, 1994; Spitz, 1989) suggests that a child's IQ level may influence the degree to which the FE is apparent. This could mean that not all children are affected equally by the FE. When testing or retesting students for special education eligibility, the cognitive ability of the child may be a factor to consider when administering tests with aging norms or when considering reclassification.

Perhaps most important, this research suggests that there is a clear need for large-scale, well-designed representative investigations of test scores commonly used to diagnose educational disabilities. Millions of children are classified as eligible for special education based substantially on tests, yet there is relatively little information about the characteristics of these tests with the populations who are potentially most affected by their use. There is a critical need, for example, for research using special education populations that examines the stability characteristics of these tests, the predictive validity of diagnosis based on these tests, the sensitivity of these tests to external influences (such as the FE), and the comparability of tests and test batteries used in diagnosis.

There are several limitations to this study. Subjects in this study were classified as LD by their respective schools. However, the definition of LD in New York State lacks objective and clear criteria for classification. The regulations stipulate that students must demonstrate a 50% discrepancy between actual ability and expected achievement. However, the regulations fail to specify exactly how this difference is to be calculated. The limited geographic range represented by these samples also makes the generalizability of the results unknown. Another limitation involves using existing triennial evaluation data for Sample 2. Utilizing these data means that there was a 3-year time gap between administrations of the Wechsler tests. Therefore, for Sample 2, intervention history was not controlled. In addition, Sample 2 was not counterbalanced with respect to initial test administered, thus allowing for the possibility of test order effects.

The small size of Sample 1 is a major limitation that potentially affected the results. A larger sample might have yielded statistically significant results with respect to finding the FE in a LD population when using the WJ-Cog. Previous results and the trends in the data suggest that this may indeed be the case. The use of different versions of the WJ-Cog is also a limitation. Comparisons between the two tests found that they varied significantly with respect to item sameness. This dissimilarity could have negatively affected our ability to detect the FE. Using tests with two versions that are more alike (such as was evident in Sample 2 with the WISC-R and WISC-III) might have resulted in finding greater differences between the performances. Finally, it is not clear that practitioners typically use the PTS modification for score comparisons. However, in this case, the high reliability of the cognitive measures and relative lack of scores that were substantially different from the test mean resulted in PTS modifications that were quite small. This suggests that the results reported for this study are probably applicable for practitioners whether they use the PTS or not.

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.

Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). USA: ERIC Clearinghouse on Assessment and Evaluation.

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.

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

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

Cattell, R. B. (1951). The fate of national intelligence: Test of a thirteen year prediction. *Eugenics Review, 42,* 136-148.

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 from 1932 to 1978. *Psychological Bulletin, 95,* 29-51.

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

Flynn, J. R. (1998). 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. (1999). Searching for justice: The discovery of IQ gains over time. *American Psychologist, 54*(1), 5-20.

Flynn, J. R. (2001, November). *The history of the American mind in the 20th century: A scenario to explain IQ gains over time and a case for the irrelevance of g.* Paper presented at the Quadrennial Spearman 2001 Conference, Sydney, Australia.

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, 14,* 128-133.

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

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.

Slate, J. R., & Saarino, 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.

Teasdale, T. W., & Owen, D. R. (1989). Continuing secular increases in intelligence and a stable prevalence of high intelligence levels. *Intelligence, 13,* 255-262.

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

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.

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.

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.

Woodcock, R. W., & Johnson, M. B. (1989, 1990). *Woodcock-Johnson Psycho Educational Battery–Revised.* Itasca, IL: Riverside.

Woodcock, R. W., & Johnson, M. B. (2001). *Woodcock-Johnson Psycho Educational Battery–III.* Itasca, IL: Riverside.