# Journal of Psychoeducational Assessment

http://jpa.sagepub.com/


## Peeking Inside the "Black Box" of the Flynn Effect: Evidence From Three Wechsler Instruments


Xiaobin Zhou, Jianjun Zhu and Lawrence G. Weiss



*Journal of Psychoeducational Assessment* 2010 28: 399 originally published online 16 June 2010
DOI: 10.1177/0734282910373340

Journal of Psychoeducational Assessment
28(5) 399–411

DOI: 10.1177/0734282910373340
http://jpa.sagepub.com

# Peeking Inside the "Black Box" of the Flynn Effect: Evidence From Three Wechsler Instruments

**Xiaobin Zhou[1], Jianjun Zhu[1], and Lawrence G. Weiss[1]**

**Abstract**

This study investigated the Wechsler Performance IQ (PIQ) or Perceptual Reasoning Index (PRI)/ Perceptual Organization Index (POI) change over time and its relation to ability levels. PIQ or PRI/ POI was analyzed because of the known sensitivity of nonverbal scales to the Flynn effect. Scores were analyzed using two methods. First, analysis of covariance was applied to the combination of four representative samples of individuals who were administered the following pairs of Wechsler batteries in counterbalanced order: *Wechsler Preschool and Primary Scale of Intelligence–Revised* (WPPSI-R) and WPPSI-III ($N = 174$), *Wechsler Intelligence Scale for Children – Third Edition* (WISC-III) and WISC-IV ($N = 239$), *Wechsler Adult Intelligence Scale–Revised* (WAIS-R) and WAIS-III ($N = 191$), and WAIS-III and WAIS-IV ($N = 240$). Second, equal percentile equating was applied to each of the samples independently. Although the two methods produced different patterns of results, both methods showed some evidence of variation in the magnitude of the Flynn effect across ability levels. These results call into question the practice of adjusting IQs based on an average expected Flynn effect in routine clinical evaluations.

**Keywords**

Flynn effect, IQ, Wechsler intelligence scales

Research in the past two decades has shown cumulative support for the existence of the Flynn effect (FE)—the massive IQ gain at about 0.3 points per year (Daley, Whaley, Sigman, Espinosa, & Neumann, 2003; Flynn, 1984, 1987; Nettelbeck & Wilson, 2003). However, increasing evidence also suggests that there are large variations in IQ change over time. For example, the magnitude of IQ change was found to vary in different nations (e.g., Must, Must, & Raudik, 2002) and in different times of history (e.g., Colom, Lluis-Font, & Andres-Pueyo, 2005). Some researchers have demonstrated that the FE may have ceased or reversed in recent years (e.g., Sundet, Barlaug, & Torjussen, 2004; Teasdale & Owen, 2005). Variations in the magnitude and the direction of FE are also found among different instruments (e.g., Wicherts et al., 2004; Yang, Zhu, Pinon, & Wilkins, 2006). In one article, Flynn (2006) stated that the Wechsler Adult

[1]Pearson, San Antonio, TX, USA

**Corresponding Author:**
Xiaobin Zhou, Pearson, 19500 Bulverde Road, San Antonio, TX 78259, USA
Email: xiaobin.zhou@pearson.com

Intelligence Scale–Third Edition (WAIS-III) standardization sample is substandard and a 2.34-point adjustment to individual IQ scores is required. The only evidence Flynn provided for this statement is that WAIS-III scores did not fit expectations based on the average FE. As Weiss (2008) pointed out, however, this amounts to adjusting data to fit theory and is therefore contrary to the orderly progress of science.

Along with exploration of the magnitude and direction of the FE, researcher have also focused on the "black box" behind it (e.g., is it a real IQ gain due to social, population, and or genetic factors or it is simply a psychometric artifact). Factors such as environment (e.g., Dickens & Flynn, 2001), nutrition (e.g., Colom et al., 2005), education (e.g., Teasdale & Owen, 2005), genetic (e.g., Rodgers & Wanstrom, 2007), changes in assessment construct or content (Zhu & Tulsky, 1999), and methodological issues (e.g., Rodgers, 1998; Yang et al., 2006) have been identified as potential contributors to the IQ changes. In the meantime, policy makers, test developers, and clinical psychologists are also making reference to the FE in their practice (Beaujean & Gulling, 2006; Kanaya, Scullin, & Ceci, 2003). However, much is still left to be learned about the nature of this phenomenon and how usable it is in clinical evaluation.

This article concerns the variability of the FE across ability ranges. In his initial research, Flynn (1984) cautioned researchers that applying the allowance calculated based on the FE when estimating IQs may be "reliable only for scores in the normal range of 90 to 110" (p. 39). This word of caution was supported by later studies. For example, based on Dutch draftees' IQs obtained from a group intelligence test composed of four subtests—letter matrices, verbal analogies, number sequences, and geometric figures—Teasdale and Owen (1989) identified that the magnitude of IQ change varies across the distribution of intelligence level. The authors found that IQ increase is more evident at lower intelligence levels than at high intelligence levels. In another study, Spitz (1989) collected investigations that compared WAIS and WAIS-R full-scale IQ change across the IQ range of 50 to 130. The researcher showed that larger IQ difference and reversed IQ change are evident at IQ levels below average and above average, respectively. The most recent evidence came from the WAIS-IV standardization studies. A mild intellectual disability (MID) group and a borderline intellectual functioning (BIF) group were tested on both the WAIS-IV and the WAIS-III. The full-scale IQ (FSIQ) increase was 4.1 for the former and 2.2 for the latter; whereas the FSIQ increase for the average test–retest sample was 2.9 (Wechsler, 2008). Thus, the results from the MID group support the higher IQ change rate in lower ability groups; the results from the BIF group suggest possible variations even within the lower intellectual functioning groups.

With converging evidence that the FE might be a more typical phenomenon in the average ability range, it would be uncertain how generalizable this aggregated IQ change is to any particular ability group of concern. Despite this legitimate concern, the FE has already had impact on the definition of intellectual and developmental disability (formerly mental retardation) and is consequently affecting practice in special education and judiciary evaluation (e.g., Kanaya, Scullin, & Ceci, 2003).

The current research investigated the variability of the FE using the data from a series of Wechsler intelligence scales that assess individuals in a wide age range. We focused on (a) the relationship between the FE and ability level, and (b) the within ability group, or individual, variability of the FE. Changes in Performance IQ (PIQ) between the recent editions of three Wechsler intelligence scales—Wechsler Preschool and Primary Scale of Intelligence (i.e., WPPSI-R and WPPSI-III; Wechsler, 1989, 2002), Wechsler Intelligence Scale for Children (i.e., WISC-III and WISC-IV, Wechsler, 1991, 2003), and Wechsler Adult Intelligence Scale (i.e., WAIS-R, WAIS-III, and WAIS-IV, Wechsler, 1981, 1997, 2008)—were investigated. Because PIQ is not available in WISC-IV and WAIS-IV, the comparable index scores, Perceptual Organization Index (POI) and Perceptual Reasoning Index (PRI), were used in comparisons involving these two instruments (see Table 2 footnote for subtest compositions of the indexes compared). To simplify the discussions in the remaining sections of this article, the term *PIQ* will be used when referring to comparisons involving PRI/POI. The PIQs are analyzed because, compared with verbal scales, this composite is a better measure of fluid intelligence, the type of cognitive ability that has been shown in repeated studies to be the most sensitive to the IQ gain (e.g., Flynn, 1998; Kaufman & Lichtenberger, 2006; Truscott & Frank, 2001).

# Method

## Samples

The samples used for examining each of the three instruments were collected during the standardizations of the later editions of the tests. All samples were collected to represent the percentages of key national demographics (i.e., age groups, sex, ethnicity, and self or parent education level). The test administration was counterbalanced, such that approximately half of the sample was tested on the earlier edition first and the other half was tested on the newer edition first. The testing interval between the two administrations ranged from 5 days to 12 weeks with mean testing interval of 28 to 35 days. The total sample sizes were 174, 239, 191, and 240 for WPPSI-R and WPPSI-III, WISC-III and WISC-IV, WAIS-R and WAIS-III, and WAIS-III and WAIS-IV, respectively. Within each sample, the demographic compositions of the two testing orders are balanced on most categories. Detailed demographics information for each sample is presented in Table 1.

## Analysis

Analyses were conducted to investigate differences across all tests by ability groups. To reduce the influence of the regression-to-the-mean effect in comparing PIQ change across ability levels, the verbal composite score obtained on the newer edition of the tests was used to categorize the ability level of the examinee (i.e., Verbal Intelligence quotient [VIQ] was used in WPPSI-III and Verbal Comprehension Index [VCI] was used in the other three comparisons). The following five verbal ability levels were used: ≥120, 110 to 119, 90 to 109, 80 to 89, and ≤79.

The year of publication for each instrument was used to estimate the time gap between the two versions of the test. The publication gaps are 13 years, 12 years, 16 years, and 11 years for WPPSI-R and WPPSI-III, WISC-III and WISC-IV, WAIS-R and WAIS-III, and WAIS-III and WAIS-IV, respectively. The PIQ change per year for each examinee was then calculated.

Preliminary analysis of variance (ANOVA) confirms that, on PIQ change per year, there is no significant main effect of test battery, $F(3, 838) = 0.90$, $p = .44$, partial $\eta^2 = .003$. There is also no significant interaction between ability level and test battery, $F(12, 838) = 1.35$, $p = .19$, partial $\eta^2 = .02$. These results suggest a consistent relation between ability level and PIQ changes across data sets. Therefore, the four data sets were merged to increase statistical power.

The first analysis used descriptive statistics to examine the mean PIQ change across instruments and ability levels. The second analysis further modeled the trend in PIQ change using an analysis of covariance (ANCOVA) model with ability level used as the main effect predicting PIQ change per year. Test battery and demographic information were used as covariates. The actual and the demographic-adjusted PIQ change rates for each ability group were plotted and compared.

To further investigate the variation of the PIQ change across ability levels, the third set of analyses used equal percentile equating on each of the four samples. This methodology, unlike

**Table 1.** Percentages of Demographic Categories in the Four Samples

| | WPPSI-III and WPPSI-R | | | WISC-IV and WISC-III | | | WAIS-III and WAIS-R | | | WAIS-IV and WAIS-III | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | WPPSI-III First | WPPSI-R First | Total | WISC-IV First | WISC-III First | Total | WAIS-III First | WAIS-R First | Total | WAIS-IV First | WAIS-III First | Total |
| Age (years) | | | | | | | | | | | | |
| Mean | 4.9 | 5.0 | 4.9 | 11.3 | 11.4 | 11.4 | 47.4 | 47.2 | 47.3 | 54.2 | 51.3 | 52.7 |
| Standard deviation | 1.1 | 1.2 | 1.1 | 3.0 | 2.9 | 2.9 | 21.0 | 20.4 | 20.6 | 24.8 | 24.2 | 24.5 |
| Gender | | | | | | | | | | | | |
| Female | 45.5 | 50.0 | 47.7 | 53.1 | 52.3 | 52.7 | 52.6 | 51.0 | 51.8 | 61.7 | 62.5 | 62.1 |
| Male | 54.5 | 50.0 | 52.3 | 46.9 | 47.7 | 47.3 | 47.4 | 49.0 | 48.2 | 38.3 | 37.5 | 37.9 |
| Race/ethnicity | | | | | | | | | | | | |
| White | 56.8 | 52.3 | 54.6 | 76.6 | 81.7 | 78.7 | 82.1 | 76.0 | 79.1 | 65.8 | 67.5 | 66.7 |
| African American | 15.9 | 26.7 | 21.3 | 3.1 | 1.8 | 2.5 | 8.4 | 14.6 | 11.5 | 17.5 | 18.3 | 17.9 |
| Hispanic | 23.9 | 18.6 | 21.3 | 16.4 | 10.8 | 13.8 | 6.3 | 7.3 | 6.8 | 10.0 | 7.5 | 8.8 |
| Other | 3.4 | 2.3 | 2.9 | 3.9 | 6.3 | 5.0 | 3.2 | 2.1 | 2.6 | 6.7 | 6.7 | 6.7 |
| Education (years) | | | | | | | | | | | | |
| ≤8 | 5.7 | 3.5 | 4.6 | 1.6 | 4.5 | 2.9 | 8.4 | 3.1 | 5.8 | 7.5 | 4.2 | 5.8 |
| 9-11 | 13.6 | 8.1 | 10.9 | 10.9 | 6.3 | 8.8 | 9.5 | 9.4 | 9.4 | 13.3 | 5.8 | 9.6 |
| 12 | 31.8 | 32.6 | 32.2 | 28.9 | 28.8 | 28.9 | 43.2 | 24.0 | 33.5 | 27.5 | 36.7 | 32.1 |
| 13-15 | 37.5 | 31.4 | 34.5 | 39.1 | 36.0 | 37.7 | 20 | 30.2 | 25.1 | 26.7 | 29.2 | 27.9 |
| ≥16 | 11.4 | 24.4 | 17.8 | 19.5 | 24.3 | 21.8 | 18.9 | 33.3 | 26.2 | 25.0 | 24.2 | 24.6 |
| *N* | 88 | 86 | 174 | 128 | 111 | 239 | 95 | 96 | 191 | 120 | 120 | 240 |

Note: WPPSI = Wechsler Preschool and Primary Scale of Intelligence; WISC = Wechsler Intelligence Scale for Children; WAIS = Wechsler Adult Intelligence Scale.

**Table 2.** Overall Average PIQ Scores and Average PIQ Change Over Time

| | Overall PIQ | | | | |
|---|---|---|---|---|---|
| Scales (Years Apart[a]) | Mean 1 | *SD* 1 | Mean 2 | *SD* 2 | PIQ Change/Year |
| WPPSI-R and WPPSI-III (13)[b] | 102.59 | 13.34 | 99.47 | 14.82 | 0.24 |
| WISC-III and WISC-IV (12)[c] | 107.36 | 14.71 | 103.90 | 13.95 | 0.29 |
| WAIS-R and WAIS-III (16)[d] | 108.17 | 14.41 | 103.54 | 15.38 | 0.29 |
| WAIS-III and WAIS-IV (11)[e] | 103.70 | 15.49 | 100.27 | 15.43 | 0.31 |

Note: PIQ = Performance Intelligence Quotient; *SD* = standard deviation; WPPSI = Wechsler Preschool and Primary Scale of Intelligence; WISC = Wechsler Intelligence Scale for Children; WAIS = Wechsler Adult Intelligence Scale.

a. Years between publication.
b. The WPPSI-R PIQ comprises Block Design, Object Assembly, Picture Completion, Geometric Design, and Mazes; The WPPSI-III PIQ for ages 2 to 3 years comprises Block Design and Object Assembly; WPPSI-III PIQ for ages 4 to 7 years comprises Block Design, Matrix Reasoning, and Picture Concepts.
c. The WISC-III Perceptual Organization Index (POI) is compared with the WISC-IV Perceptual Reasoning Index (PRI). The WISC-III POI comprises Picture Completion, Picture Arrangement, Block Design, and Object Assembly; The WISC-IV PRI comprises Picture Concepts, Block Design, and Matrix Reasoning.
d. The WAIS-R PIQ comprises Picture Completion, Picture Arrangement, Block Design, Object Assembly, and Digit Symbol; The WAIS-III PIQ comprises Picture Completion, Digit Symbol-Coding, Block Design, Matrix Reasoning, and Picture Arrangement.
e. The WAIS-III POI is compared with the WAIS-IV PRI. WAIS-III POI comprises Block Design, Matrix Reasoning, and Picture Completion; The WAIS-IV PRI is comprises Block Design, Matrix Reasoning, and Visual Puzzles.

the ANCOVA approach, eliminates any possible influence of the regression effect. The equivalent PIQ on the two versions of the test across the score range of the scale was examined.

## Results

### *Average PIQ Change on Wechsler Scales*

Table 2 shows the average PIQs on the four instruments. The average PIQ change is 0.24, 0.29, 0.29, and 0.31 points increase per year for WPPSI-R, WISC-III, WAIS-R, and WAIS-III, respectively. Applying the terminology used by Kaufman (2010), the FE for the total samples ranged from 2.4 points per decade on the WPPSI-R to 3.1 points per decade on the WAIS-III, with a weighted mean of 2.85 points per decade for $N = 844$.

The descriptive statistics of PIQ change at each verbal ability level are presented in Table 3. PIQ change is calculated as the "increase" from the older version to the newer version. Thus, a positive discrepancy indicates IQ gain whereas the negative value indicating IQ decrease. Given the publication intervals between the two versions of the test, the expected discrepancy based on the FE prediction would be within the range of 3 to 5 points across all four comparisons. Table 3 shows a clear decrease in PIQ gain at the above-average ability levels. The average PIQ change per year in the middle and lower ability groups is in the range of 0.31 to 0.37; whereas this change rate drops to 0.06 and 0.15 in the two higher ability groups. Figure 1 shows the percentages of PIQ change per year by ability level. In addition to the rather small proportion of examinees having PIQ change around 0.30 (i.e., between 0.20 and 0.40) per year from the FE prediction, two other patterns are notable in Figure 1: (a) higher proportion of examinees in the three lower ability groups than in the two higher groups have PIQ change greater than 0.40 points per year and (b) higher proportion of examinees in the two higher ability groups than in the lower groups have reversed FE, which contributes to the shrinking average PIQ change rate in these groups.

**Table 3.** Descriptive Statistics of PIQ Change by Verbal Ability[a]

| PIQ Change (Total) | VCI/VIQ[b] Levels | | | | |
|---|---|---|---|---|---|
| | ≤79 (*N* = 54) | 80-89 (*N* = 129) | 90-109 (*N* = 421) | 110-119 (*N* = 156) | ≥120 (*N* = 79) |
| Mean | 4.2 | 4.6 | 4.6 | 0.9 | 1.9 |
| *SD* | 8.9 | 9.7 | 10.5 | 9.8 | 11.0 |
| Median | 4.5 | 5.0 | 5.0 | –1.0 | 2.0 |
| PIQ change (per year) | | | | | |
| Mean | 0.31 | 0.36 | 0.37 | 0.06 | 0.15 |
| *SD* | 0.69 | 0.78 | 0.84 | 0.77 | 0.88 |
| Median | 0.37 | 0.38 | 0.38 | –0.08 | 0.13 |

Note: PIQ = Performance IQ; VCI = Verbal Comprehension Index; VIQ = Verbal IQ.

a. PIQ discrepancy is calculated using the score obtained on the older version minus score obtained on the newer version.
b. VIQ is used on the Wechsler Preschool and Primary Scale of Intelligence (WPPSI).

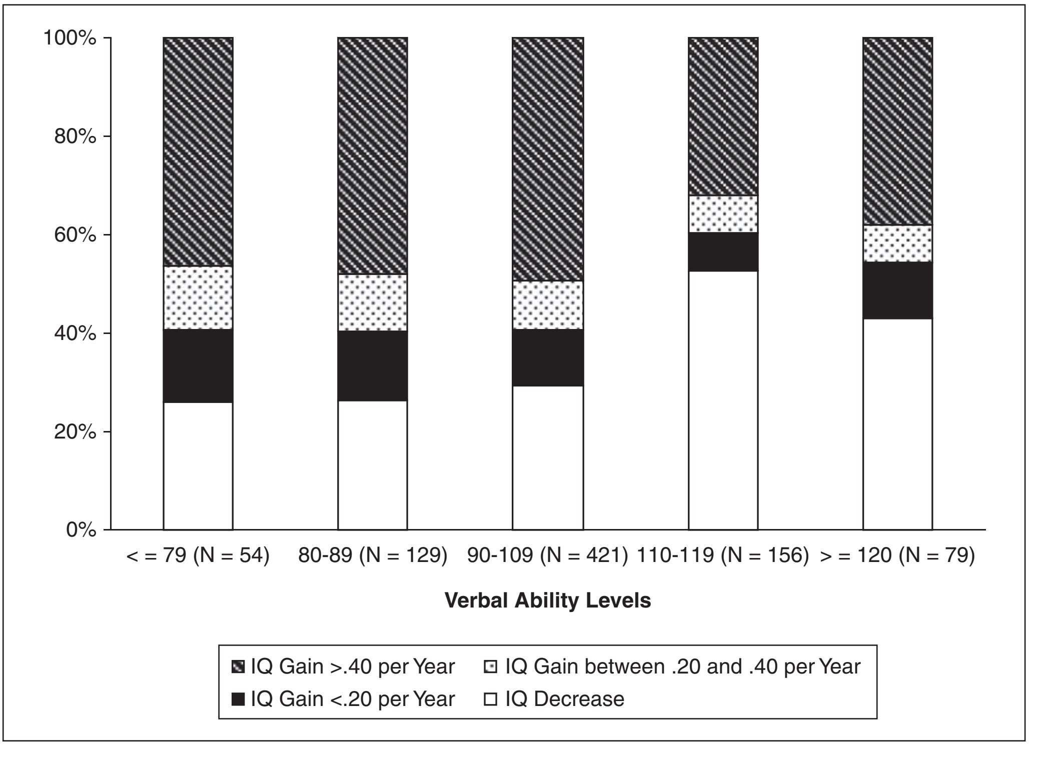


**Figure 1.** Percentages of examinees obtaining various PIQ change by ability level


## *PIQ Change Per Year: Predictive Approach by Ability Level*

The ANCOVA analysis was conducted on the consolidated sample. Ability level was investigated as the main effect predicting PIQ change per year. Testing group categories (i.e., test battery and

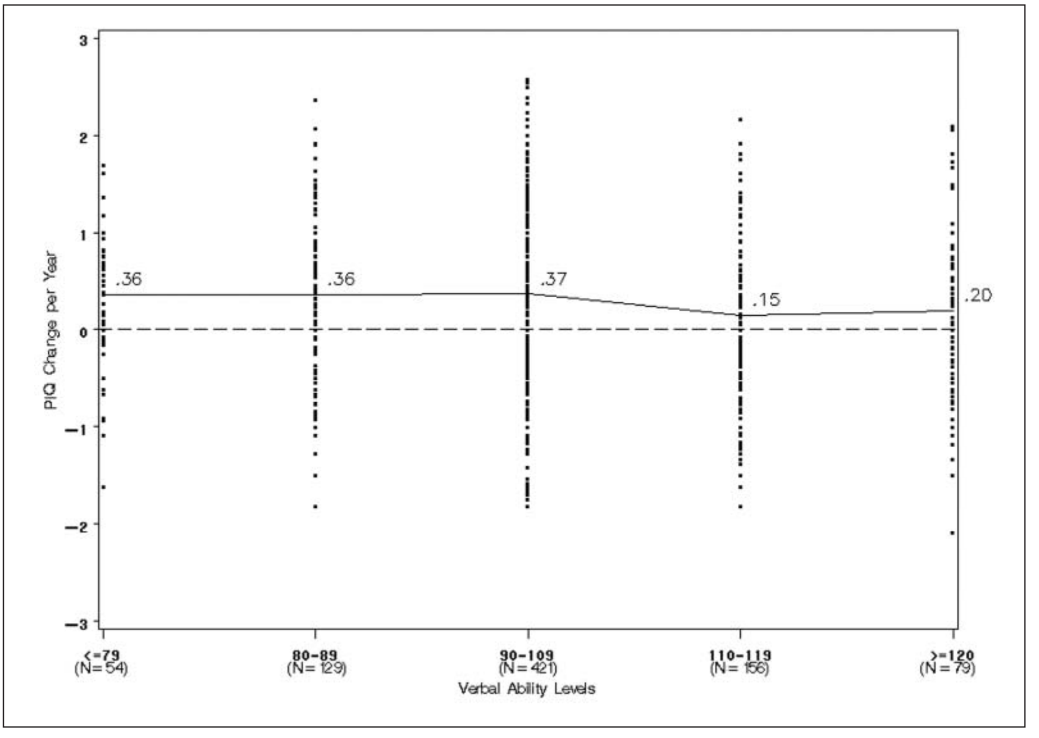


**Figure 2.** Actual and demographically adjusted average PIQ change per year by ability level.
Average PIQ changes at each ability level adjusted by demographics are plotted by the solid line. Source: Data and figure copyright Pearson, Inc., 2009. All rights reserved.

testing order), and main demographics (i.e., age, sex, ethnicity, education level) were used as covariates. Significant effects of ability level, $F$ (4, 817) = 2.48, $p$ = .04, and testing order, $F$ (1, 817) = 64.75, $p$ < .01, were found. For both factors, the effect sizes are small (partial $\eta^2$ = .01 for ability level; partial $\eta^2$ = .07 for testing order).

To control the influences of testing group and demographics composition in the samples, least-squares adjusted average PIQ change for each ability level was calculated. Figure 2 shows the observed and the adjusted PIQ change per year at each ability level. The least-squares adjusted average PIQ changes at each ability level are plotted by the solid line. This figure shows two important patterns of PIQ change with respect to ability: (a) the adjusted PIQ change is higher at low ability levels than at high ability levels—the PIQ change at the middle and lower ability levels is 0.36 or 0.37, which about doubles the change rate for ability level 110 –119 (0.15) and ≥120 (0.20) and (b) within each ability level, the range of observed change is rather large.

The variability of the PIQ change across verbal ability levels is evident in the above results. However, given the moderate-to-strong correlation between verbal and perceptual scales, it is possible that the inversed relation between the magnitude of PIQ change and ability level observed in Figure 2 is partially inflated by the effect of regression to the mean. Therefore, non-linear equating on the performance scale was used to validate this finding.

## *PIQ Change: Equal Percentile Equating*

Equal percentile equating was conducted independently on the four samples. The results are shown in Figures 3 to 6. If the PIQ increase were at a fixed rate regardless of individual's ability

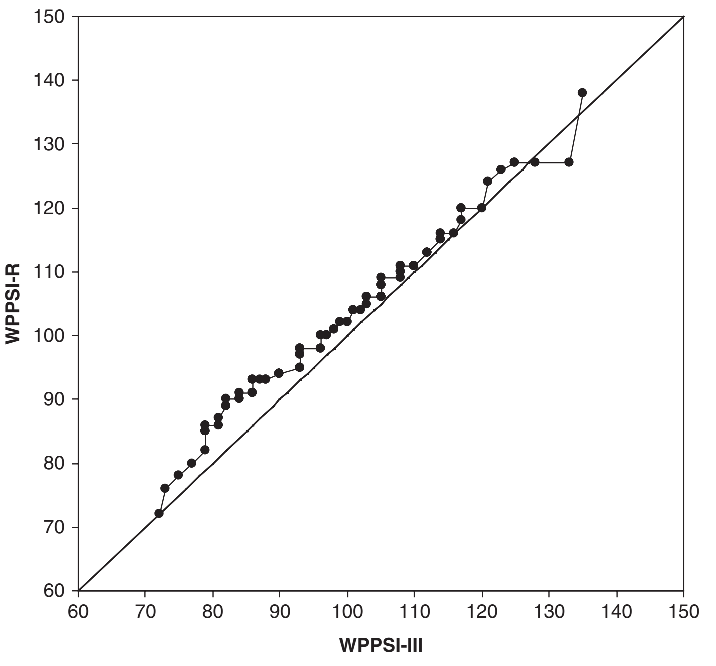


**Figure 3.** Equal percentile equating of the WPPSI-R and the WPPSI-III
Source: Data and figure copyright Pearson, Inc., 2009. All rights reserved.

level, we would expect to see a parallel line on top of the identity (diagonal) line. That is, at every percentile of the sample, the PIQ would be consistently higher in the older version than in the newer version by the same amount. This pattern, however, is not observed in any of the four analyses. In contrast, in the WPPSI (Figure 3) and WAIS-R (Figure 5), the equated scores move closer to the identity line at the higher percentile level; in the WISC (Figure 4) and the WAIS-III (Figure 6), the equated scores move away from the identity line at the higher end of the percentile. The nonparallel relation between the equating line and the identity line, again, confirms the variability of the change in PIQ across the ability distribution in the sample.

If inferred directly from the ANCOVA results, a certain pattern of the equated score is expected. The ANCOVA findings suggest higher IQ gain at the lower end of the ability level and lower IQ gain at the higher end. Therefore, the equated scores could be expected to rise higher from the identity line at the bottom than at the top. This pattern is only observed in the WPPSI and WAIS-R results. For example, at about one standard deviation below the mean, the PIQ of 86 on WPPSI-III is equivalent to the PIQ of 91 on WPPSI-R—a 5-point IQ gain. In contrast, at about one standard deviation above the mean, the PIQ of 114 on the WPPSI-III is equivalent to the PIQ of 115 on WPPSI-R—only 1-point IQ gain. The equating result for the WISC and the WAIS-IIII seems to contradict the direction of the magnitude change by ability levels suggested in the ANCOVA analysis.

## Discussion

The present exploration of the relation between PIQ change and ability level provides critical and valuable insight into the nature of the FE across ability levels. The magnitude of the FE

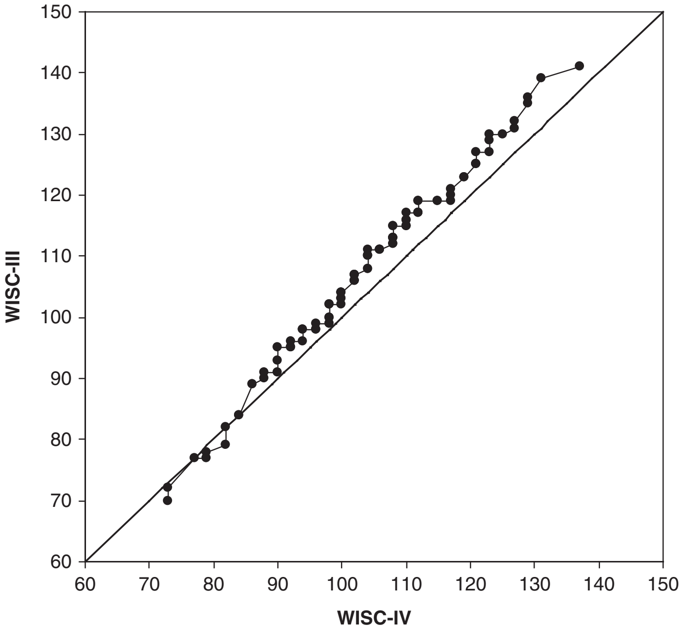


**Figure 4.** Equal percentile equating of the WISC-III and the WISC-IV


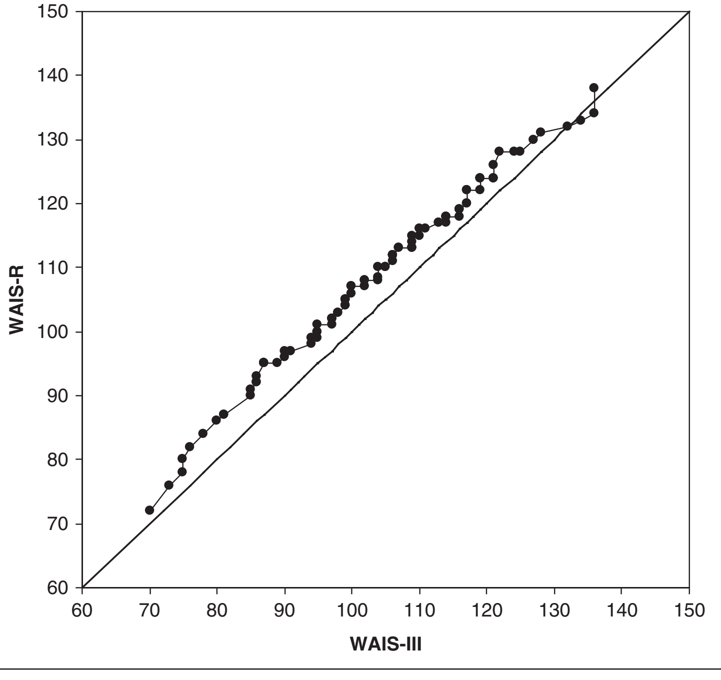


**Figure 5.** Equal percentile equating of the WAIS-R and the WAIS-III

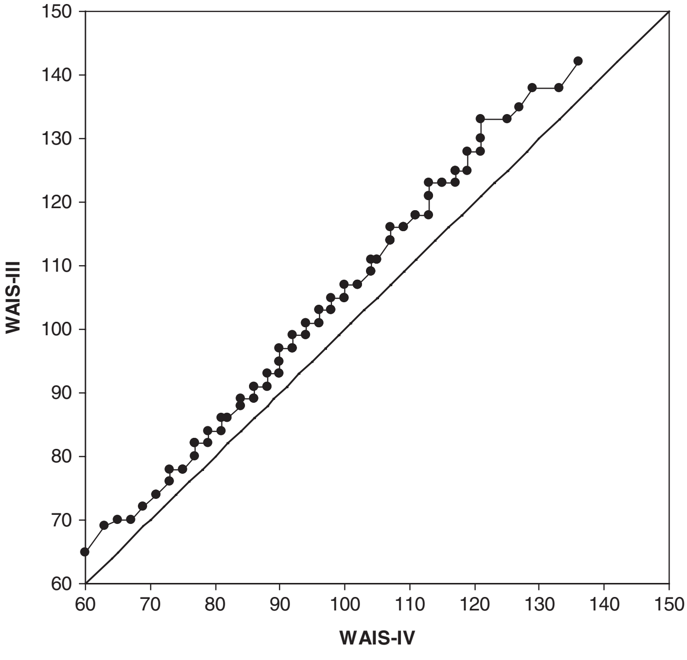


**Figure 6.** Equal percentile equating of the WAIS-III and the WAIS-IV
Source: Data and figure copyright Pearson, Inc., 2009. All rights reserved.

(as measured by change in PIQ scores) varies across ability groups, although the direction of the variation across ability groups is inconclusive. This finding is consistent for all Wechsler intelligence tests studied spanning the range from preschool to older adult. Overall, our findings suggest that the average IQ gain Flynn initially described may only be valid as an aggregated phenomenon. The variation by ability group we demonstrated implies that adjusting an individual observed IQ by a fixed rate obtained from the overall sample may yield systematic over or under estimates of IQ depending on the individual's ability level.

The ANCOVA analysis allowed the four samples to be combined and results suggested that the rate of change in PIQ scores is larger at the middle and lower portion of the distribution and smaller in the upper portion of the distribution of scores. The equating analyses were necessarily conducted on each of the four samples separately. Results of the equating analyses agreed with findings of the ANCOVA analysis for the WPPSI-R and WPPSI-III and the WAIS-R and WAIS-III samples in that both methods suggested larger rates of change in the middle and lower portions of the distribution. Results of the equating analyses did not agree with findings of the ANCOVA analysis for the WISC-III and WISC-IV or the WAIS-III and WAIS-IV samples in that the equating analyses suggested larger rates of change in the upper portion of the distribution for these samples.

Multiple analysis approaches were used in the current study with the intent that the strength of each method could complement the limitations of the other method. A strength of the ANCOVA method is that it allows the four samples to be combined to increase statistical power. A possible limitation of the ANCOVA method when examinees are categorized by ability level is its vulnerability to exaggerated findings in the tails of the distribution because of the regression-to-the-mean effect. Although this risk was ameliorated by categorizing ability based on VIQ, the high correlation between VIQ and PIQ might still produce some regression effects on the PIQ change scores. A strength of the equating method is that it allows for nonlinear solutions. However, the equating method is vulnerable to sample fluctuations at the upper and lower tails of the distribution because each data point has the same weight regardless of the frequency of the value. Equating was necessarily conducted on the four samples separately possibly producing less stable outcomes because of the smaller sample sizes. Finally, the equating procedure introduces additional errors and the error at the extreme ends of the distribution is often several times larger than the equating error in the middle of the distribution. Clearly, further analyses are needed to better understand why findings from these two methods diverge for some samples but not others.

Both the ANCOVA and the equating methods demonstrate converging evidence of the variability of the PIQ change across ability groups, although the inconsistent directions in the variation merit further exploration. Taken together, these findings suggest that the FE may vary by ability level but the pattern across ability levels may be different for different tests or age groups tested. Further research is needed to untangle these findings. Although the different methods suggest different patterns across ability levels, both methods show that the magnitude of the IQ change is not the same across IQ levels. This convergent aspect of the two methods presents a challenge for the emerging view that 0.3 points per year is an appropriate adjustment in routine clinical practice.

The debate over adjusting IQ scores based on the FE is further complicated by applied issues in clinical practice, theoretical issues in test development, and research methods selected. When the FSIQ cannot be obtained for a particular examinee, the clinician sometimes must use VIQ, PIQ, or an abbreviated IQ as the best estimates of intellectual ability. In such instances, applying an adjustment derived based on FSIQ could yield an erroneous interpretation of the person's ability because different domains of intelligence may have different patterns of change over time. At the same time, test developers are changing the mix of constructs included in the FSIQ based on new research in human intelligence, recently, by replacing older subtests measuring crystallized knowledge and visual spatial skills with new working memory and processing speed subtests that may not show the same rate of change over time. For example, in the WAIS-III FSIQ, 3 out of 11 subtests, or 27% are working memory and processing speed subtests; in the WAIS-IV FSIQ, 4 out of 10 subtests, or 40% measure these two domains. Such change in test structure could also lead to variation from the expected rate of IQ change across time and/or among instruments. Thus, without adequate definition of restrictions, forcing an IQ adjustment using a fixed rate, such as moving the cutoff point for intellectual and developmental disability classification, could cause misleading results and potentially misclassify a proportion of examinees. At the same time, clinical practice must continue even as research continues that may affect practice. Although the evidence for differential adjustments based on ability level is still nascent, early indications appear to favor slightly larger adjustments in the lower range of scores where high-stakes legal evaluations are most likely to occur.

The FE is a much more complicated phenomenon than a simple overall increase of IQs. Further research in the following directions is necessary to evaluate and extend the findings and hypotheses of the current study:

1. To go deeper into the history—to include more data from earlier versions of Wechsler, such as WPPSI, WISC-R, and WAIS. This will not only allow us to test the robustness of the current results but will also give us an understanding of the development of IQ during a longer historical period.
2. To go wider into the human mind—to extend the research into other domains such as language, achievement, and adaptive behavior assessment. The findings from diverse domains might provide a better understanding of the nature of the FE.

3. To go wider into the diverse societies—to apply similar analysis strategies to international data.
4. To go wider into examining and comparing the various statistical methodologies used for deciphering the IQ change.

If converging results could be found in other populations, across different historical time points, or on various aspects of human intellectual properties we would then be in a better position to open the “black box” behind the FE.

## Declaration of Conflicting Interests

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

## Funding

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

## References

Beaujean, A. A., & Gulling, S. F. (2006). The Lynn-Flynn effect and school psychology: A call for research. *The School Psychologist*, *Winter*, 17-20.

Colom, R., Lluis-Font, J. M., Andres-Pueyo, A. (2005). The generational intelligence gains are caused by decreasing variance in the lower half of the distribution: Supporting evidence for the nutrition hypothesis. *Intelligence, 33*, 83-91.

Daley, T. C., Whaley, S. E., Sigman, M. D., Espinosa, M. P. & Neumann, C. (2003). IQ on rise. The Flynn effect in rural Kenyan children. *Psychological Science, 14*, 215-219.

Dickens W. T., & Flynn, J. R. (2001). Heritability estimates versus large environmental effects: The IQ paradox resolved. *Psychological Bulletin, 108*, 346-369.

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

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

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* (pp. 25-66). Washington, DC: American Psychological Association.

Flynn, J. R. (2006). Tethering the elephant capital cases, IQ and the Flynn effect. *Psychology, Public Policy, and Law, 12*, 170-189.

Kanaya, T., Scullin, M. H., & Ceci, S. J. (2003) The Flynn effect and U.S. policies: The impact of rising IQs on American society via mental retardation diagnoses. *American Psychologist, 58*, 778-790.

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

Kaufman, A. S., & Lichtenberger, E. O. (2006). *Assessing adolescent and adult intelligence* (3rd ed.). New York, NY: Wiley.

Must, O., Must, A., & Raudik, V. (2003). The secular rise in IQs: In Estonia, the Flynn effect is not a Jensen effect. *Intelligence, 31*, 461-471.

Nettelbeck, T., & Wilson, C. (2003). The Flynn effect: Smarter not faster. *Intelligence, 32*, 85-93.

Rodgers, J. L. (1998). A critique of the Flynn effect: Massive IQ gains, methodological artifacts, or both? *Intelligence, 26*, 337-356.

Rodgers, J. L., & Wanstrom, L. (2007). Identification of a Flynn effect in the NLSY: Moving from the center to the boundaries. *Intelligence, 35*, 187-196.

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

Sundet, J. M., Barlaug, D. F., & Torjussen, T. M. (2004). The end of the Flynn effect? A study of secular trends in mean intelligence test scores of Norwegian conscripts during half a century. *Intelligence, 32*, 349-362.

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

Teasdale, T. W., & Owen, D. R. (2005). A long-term rise and recent decline in intelligence test performance: The Flynn effect in reverse. *Personality and Individual Differences, 39*, 837-843.

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.

Wechsler, D. (1981). *Wechsler Adult Intelligence Scale–Revised*. San Antonio, TX: Psychological Corporation.

Wechsler, D. (1989). *Wechsler Preschool and Primary Scale of Intelligence–Revised*. San Antonio, TX: Psychological Corporation.

Wechsler, D. (1991). *Wechsler Intelligence Scale for Children–Third Edition*. San Antonio, TX: Psychological Corporation.

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale–Third Edition*. San Antonio, TX: Psychological Corporation.

Wechsler, D. (2002). *Wechsler Preschool and Primary Scale of Intelligence–Third Edition*. San Antonio, TX: Psychological Corporation.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children–Fourth Edition*. San Antonio, TX: Psychological Corporation.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale–Fourth Edition*. San Antonio, TX: Pearson.

Weiss, L. G. (2008). *WAIS-III technical report: Response to Flynn*. Retrieved from http://www.pearsonassessments.com/NR/rdonlyres/98BBF5D2-F0E8-4DF6-87E2-51D0CD6EE98C/0/WAISIII_TR_lr.pdf

Wicherts, J. M., Dolan, C. V., Hessen, D. J., Oosterveld, P., Caroline, G., van Baal, M., . . . Span, M. M. (2004). Are intelligence tests measurement invariant over time? Investigating the nature of the Flynn effect. *Intelligence, 32*, 509-537.

Yang, Z., Zhu, J., Pinon, M., & Wilkins, C. (2006, August). *Comparison of the Bayley–III and the Bayley–II*. Paper presented at the annual meeting of the American Psychological Association, New Orleans, LA.

Zhu, J., & Tulsky, D. (1999). Can IQ gain be accurately quantified by a simple difference formula? *Perceptual and Motor Skill, 88*, 1255-1260.