Original Article

Assessing the Flynn Effect in the Wechsler Scales

Alexander Beaujean and Yanyan Sheng

Department of Educational Psychology, Baylor University, Waco, TX, USA

Abstract. The current study examined the Flynn Effect (i.e., the increase in IQ scores over time) across all editions of the Wechsler Adult Intelligence Scale (WAIS), Wechsler Intelligence Scale for Children (WISC), and Wechsler Preschool and Primary Scale of Intelligence (WPPSI). By reverse engineering the correlation and scale score transformations from each Wechsler edition's technical manual, we made a mean and covariance matrix using the subtests and age groups that were in common for all editions of a given instrument. The results indicated that when aggregated, there was a FE of 0.44 IQ points/year. This Wechsler instrument used, however, moderates the FE, with the WISC showing the largest FE (0.73 IQ points/year) and the WAIS showing a smallest FE (0.30 IQ points/year). Moreover, this study found that the amount of invariant indicators across instruments and age groups varied substantially, ranging from 51.53% in the WISC for the 7-year-old group to 10.00% in the WPPSI for the 5- and 5.5-year-old age groups. Last, we discuss future direction for FE research based on these results.

Keywords: Flynn effect, Wechsler Intelligent Scale, multi-group confirmatory factor analysis, measurement invarience

The Flynn Effect (FE) is the average rise in psychometric IQ scores, an effect that has been evident since the early 20th century (Flynn, 2007). Although scholars started to notice IQ gains in the 1940s (e.g., Tuddenham, 1948), Herrnstein and Murray (1996) first coined the term, naming it after James R. Flynn who, along with Richard Lynn, popularized the area of research. Flynn's (1982, 1983, 1984) and Lynn's (1982, 1987) initial work in this area focused on IQ scores from Japan and the United States, but since this initial foray they, along a host of other investigators, have reported similar effects on every inhabited continent (Flynn, 2007; Neisser, 1998). Moreover, the FE has been studied in a multiplicity of populations, ranging from infants (Lynn, 2009b) to adults (Schaie, Willis, & Pennak, 2005) and from inhabitants of both developed and impoverished countries (Flynn, 1987; Khaleefa, Sulman, & Lynn, 2009; te Nijenhuis, 2013; te Nijenhuis, Murphy, & van Eeden, 2011).

Societal Influence

While research on IQ gains was initially of interest to researchers interested in cognitive ability, it has grown to have a substantial influence on modern society in the United States, whether directly or indirectly. Two of the places where the FE has had the most impact are in education and law. First, as IQ scores are used in many Special Education decisions, the FE has influenced the Special Education determination process (Flynn, 2000), especially with the determination of an Intellectual Disability (ID; Flynn & Widaman, 2008; Kanaya, Ceci, & Scullin, 2003) or Specific Learning Disability (Sanborn, Truscott, Phelps, & McDougal, 2003; Truscott & Frank, 2001; Truscott & Volker, 2005) diagnosis.

Second, the FE has had a forensic influence, most notably since the US Supreme Court ruled in *Atkins v. Virginia* (2002) that it was cruel and unusual to execute someone with an ID. As IQ scores are a substantial piece of evidence when determining the presence of an ID (American Psychiatric Association, 2000; Reschly, Myers, & Hartel, 2002), the FE has become an issue in many capital punishment cases, with some courts even ruling that the FE must be considered when determining a defendant's IQ (Flynn, 2006). Since the Atkins ruling, there has been much debate by forensic psychologists on how to account for the FE in capital punishment cases (Hagan, Drogin, & Guilmette, 2010; Weiss, Haskins, & Hauser, 2004), an issue that has still not found any consensus (Kaufman & Weiss, 2010).

Concerns About the Flynn Effect

While descriptions of the FE often put its magnitude to be approximately 0.30 IQ points/year (Neisser, 1998; VandenBos, 2007), there is reason to believe that this estimate is too oversimplified. First, there is some indication that the magnitude might be moderated by intellectual ability, being more concentrated in people with lower cognitive abilities (Colom, Lluis-Font, & Andrés-Pueyo, 2005;

DOI: 10.1027/1614-0001/a000128

Teasdale & Owen, 1989, but see Wai & Putallaz 2012). Second, the FE appears to be moderated by test content (Flynn, 2009). While initial FE studies used aggregated scores (e.g., Full Scale IQs [FSIQ], Verbal IQ [VIQ], Performance IQ [PIQ]), there appears to be a larger FE on measures of Fluid reasoning (e.g., induction, deduction) than on tests of Verbal-Comprehension (e.g., vocabulary knowledge) (Flynn, 1987; Lynn, 1990, 2009a, but see Flynn, 2009), and, likewise, a larger FE on tests of visual memory than verbal memory (Baxendale, 2010). There appears to be a minimal-to-no effect, however, on general intelligence (Kane & Oakland, 2000; Rudnick, 2001), reaction time (Nettelbeck & Wilson, 2004), or Piagetian tasks (Shayer, Ginsburg, & Coe, 2007). To add to the confusion, some more recent studies have shown IQ score decreases across time (Beaujean & Sheng, 2010; Teasdale & Owen, 2005, 2008).

Perhaps a critique more potent than that brought up by the moderation evidence is that summoned by examining the methods used in most FE research, positing that the types of FE data researchers analyze, as well as methods they use for the analysis, may be a contributing factor to the FE estimates. Most FE studies simply examine mean changes in aggregated IQ scores (e.g., FSIQ, PIQ, VIQ). Rodgers (1998) was one of the first critiques of such methods, stating that difference in observed scores could be due to a change in average scores, change in variances, or both. Since then, other authors (e.g., Beaujean & Osterlind, 2008; Wicherts et al., 2004) have extended Rodgers critique, stating that a major assumption in comparing manifest scores is that they are measuring the same construct, the same way, across groups. In making such an assumption, authors exclude the hypothesis that test items and/or scores could change properties over time, without their necessarily being a change in cognitive ability – a situation not uncommon with longitudinal types of studies (Bontempo & Hofer, 2007; Millsap & Hartog, 1988; Ployhart & Vandenberg, 2010). Thus, without examining the measurement instruments themselves, comparing mean differences only gives *prima facie* evidence about the magnitude and nature of the FE.

Flynn Effect and the Wechsler Intelligence Instruments

The Wechsler intelligence scales have a long history of use and refinement (Boake, 2002; Thorndike & Lohman, 1990). While initially there was only a scale for adults (Wechsler-Bellevue, which evolved into the Wechsler Adult Intelligence Scale [WAIS]), eventually there came editions for school-aged (Wechsler Intelligence Scale for Children [WISC]) and preschool children (Wechsler Preschool and Primary Scale of Intelligence [WPPSI]) (Coalson & Weiss, 2002). Since their development, they have consistently been some of the most widely used instruments used by psychological professionals (Archer & Newsom, 2000; Camara, Nathan, & Puente, 2000; Lubin & Lubin, 1972; Stinnett, Havey, & Oehler-Stinnett, 1994) and taught in psychological assessment classes (Alfonso, Oakland, LaRocca, & Spanakos, 2000). It is not surprising, then, that much of the scholarship concerning the FE, at least in the United States, has used one or more of the Wechsler scales in their analysis (Shiu & Beaujean, 2010).

Often, when studies use a Wechsler test, they give two editions of the same instrument (e.g., WAIS and WAIS-Revised) to a single sample and then compare the average FSIQ, PIQ, and/or VIQ scores. While this might seem a natural way to compare scores, this process is predicated on the belief that different editions of the same instrument measure the same construct(s), the same way. However, there can be – and often are – many differences among editions of the same Wechsler instrument, both in test content (Lewis, Sanborn, McGreevy, Tarquin, & Truscott, 2004) and test administration (Kaufman, 2010). While the constructs measured by the various editions might be similar, especially at the aggregated level (Floyd, Clark, & Shadish, 2008), there is little evidence to support that the scaling of these constructs between editions are necessarily the same, thus making mean comparisons tenuous (Nugent, 2006). These between-edition mean comparisons are akin to comparing average temperatures at two different geographic locations with thermometers that use different scales. While mean differences could be due to different temperatures, they could also be the result of the scales having different origins (e.g., Fahrenheit vs. Rankine), different units (e.g., Kelvin vs. Rankine), or both (e.g., Fahrenheit vs. Kelvin) (cf. Golembiewski, Billingsley, & Yeager, 1976 α and β type changes).

With the temperature scales, there are methods to convert one scale's units to another scale's units. Likewise, there are methods available to link different intelligence scales (Kolen & Brennan, 2004), too, although such methods are sparsely used in FE research (for some exceptions, see Beaujean & Osterlind, 2008; Beaujean & Sheng, 2010; Must, te Nijenhuis, Must, & van Vianen, 2009; Wicherts et al., 2004). One such way to link scores is by the use of Multi-Group Confirmatory Factor Analysis (MG-CFA; Chan, 1998; Millsap, 2011). MG-CFA is often used to examine bias/non-invariance in a single test among two or more groups, but the same methods can be used to link the scores of tests that measure the same construct across different samples, akin to calibrating/vertical linking procedures done with many educational assessments (Linn, 1993; Mislevy, 1992). In such cases, the major concern for the MG-CFA is twofold. First, it separates invariant from non-invariant indicators; second, it equates the scale of the latent variable among the groups using only the invariant indicators. This method allows all of a test's indicators to contribute to the estimation within that test, while allowing simultaneous estimation of invariant indicators to contribute to the linking of latent variable scores across tests (Byrne, Shavelson, & Muthen, 1989).

Purpose of Current Study

The purpose of the current study was to examine the FE across the multiple editions of the Wechsler scales:

three editions of the WPPSI (Wechsler, 1967, 1989, 2002), four editions of the WISC (Wechsler, 1949, 1974, 1991, 2003), and four editions of the WAIS (Wechsler, 1955, 1981b, 1997, 2008). This study measured the FE by comparing mean differences in between-edition latent variable scores that have been linked using MG-CFA.

Method

Sample

The sample for this study came from the norming groups of the three most commonly used Wechsler intelligence instruments: WPPSI (Wechsler, 1967, 1989, 2002), WISC (WISC, 1949; Wechsler, 1974, 1991, 2003), and WAIS (Wechsler, 1955, 1981b, 1997, 2008). The exact demographic information can be found in each of the instruments' technical/administration manuals.

Variables

FE research is longitudinal in nature. One of the strongest ways to analyze such data, especially when there are multiple measures at each data wave, is to use latent variable models (Ployhart & Vandenberg, 2010). Consequently, we formed latent variable models for each edition of the Wechsler instruments using the data presented in each edition's technical manual. While all of the Wechsler intelligence manuals present subtest correlation tables for selected age groups, not all editions of a given instrument, contain the same subtests, nor do all the manuals report the subtest correlations for the same age groups. Therefore for a given Wechsler instrument, we only used subtests and ages that were available across all of its editions. A list of common subtests and ages to all editions of a Wechsler instrument is given in Table 1. For each instrument, the common subtests measure a variety of abilities, including Verbal-Comprehension, Fluid Reasoning, Visual-Spatial abilities, Short-Term/Working Memory, and Processing Speed (Carroll, 1993).

Solely comparing the correlations across an instrument's editions has multiple problems (Cudeck, 1989), including the inability to assess for differences in the latent variable's mean across time. Consequently, we transformed each correlation matrix to a mean and covariance matrix. We did this transformation by reverse engineering the raw score conversion tables for each edition of each instrument in order to obtain the mean and standard deviation for each subtest.

Reverse Engineering the Raw Score Conversion Tables

Within an edition of each Wechsler instrument, each subtest is placed on a standard score scale (mean: 10; standard deviation: 3) for a given age group. As these scaled scores are the same for each edition, comparing them across editions of an instrument would be futile because the average and variability are the same. These standard scores can be placed back onto the raw score metric, however, which then allows for a comparison of score changes across time, within an age group.

Table 1. Common subtests and ages across Wechsler Intelligence Instrument Editions

WPPSI	WISC	WAIS
	Subtests	
Block design	Block design	Block design
Comprehension	Comprehension	Comprehension
Picture completion	Picture completion	Picture completion
Similarities	Similarities	Similarities
Vocabulary	Vocabulary	Vocabulary
	Arithmetic	Arithmetic
	Coding	Coding
	Digit span	Digit span
		Information
	Ages	
4	7	18–19
4.5	10	25–34[a]
5	13	45–54
5.5		
6[b]		

Notes. WPPSI = Wechsler Preschool Primary Intelligence Test; WISC = Wechsler Intelligence Scale for Children; WAIS = Wechsler Adult Intelligence Test. [a]For the third and fourth editions, the 25–29 and 30–34-year-old groups were combined. [b]For the first and second editions, the 6:0–6:5 and 6:6–6:11-year-old groups were combined.

For a given age group, we converted the means and variances for a subtest into raw score units in each edition of a Wechsler instrument using the following steps:

1. To obtain the mean score, we used the raw score equivalent to a scaled score of 10.
2. To obtain the standard deviation we used three steps.
   2.1. We found the raw scores equivalent to one standard deviation above and below the mean (i.e., scaled scores of 7 and 13, respectively).
   2.2. We calculated how many raw score points each score was from the mean and squared it to get two estimates of the variance.
   2.3. We averaged the two variances and then took the square root of the average to get an average standard deviation.
3. In instances where there was a range of raw scores for a single scaled score, we used the mean of the highest and lowest values. For example, if raw scores of 57–60 all gave the same scaled score, we used 58.5 as the raw score value.
4. In instances where the manual did not give a raw score for a scaled score, we took the average of the two raw scores immediately higher and lower.

5. In cases where there were multiple scaled score tables that applied to a correlation table, we averaged the statistics across all the tables.

There were two exceptions to this procedure. For the first edition of the WAIS, the raw score conversions were the same for all age groups, so there was only one mean and covariance matrix for this edition. Second, the first and second editions of the WAIS had a 25–34 age group, but the third and fourth editions had a 25–29-year-old group and a 30–34-year-old group. Consequently, we averaged the coefficients across the 25–29 and 30–34 groups so that each edition had a single 25–34-year-old group. Likewise, the first and second edition of the WPPSI has separate correlation matrices for ages 6:0–6:5 and 6:6–6:11, but the third edition has only one matrix for all the 6-year-old respondents. Consequently, we averaged the correlations, means, and variances across the two age groups for the WPPSI and WPPSI-R so that there was only one 6-year-old age group. An example of these matrices is given in the Appendix. All the data matrices are available on the first author's web page (see Electronic Supplementary Metarial).

Factor Analysis

For each age group in a given instrument's edition (37 groups in total due to the same raw score conversion tables for first edition of the WAIS), we did a confirmatory factor analysis fitting a single factor model. We chose to fit a single factor model because the subtests included for each instrument were relatively diverse, and likely only measured one common construct (i.e., *g*; Jensen, 1998). For each model, the single factor model fits the data relatively well, although the fit of some models improved when we allowed some residual variances to covary.

Subsequently, within an age group, we conducted a MG-CFA (Millsap, 2011; Wicherts & Dolan, 2010) for each Wechsler instrument, using the edition as the grouping variable. For all MG-CFAs, the latent variable's variance was constrained to unity for identification purposes. In addition, the first edition of an instrument was used as the reference group, constraining its latent mean to 0. Thus, the mean latent variable score for the subsequent editions indicates the distance from the first edition's mean in standard deviation units, as is typically done when estimating between-group effect sizes (Kelley & Preacher, 2012).

To be able to compare the latent variable means across editions, the factor model needs to have at least strong/scalar invariance, which requires that the factor loadings and intercepts for a given indicator (i.e., subtest) are constrained to be equal across editions (Vandenberg & Lance, 2000). The presence of scalar invariance indicates that individuals with the same level of the latent trait would have the same observed score on a given indictor variable irrespective of an instrument's edition. Moreover, it renders the latent variables' means to be comparable across editions of an instrument (Little & Slegers, 2005).

As the content of some of the subtests, as well as the demographics of the normative samples, likely changed from one edition to another (Coalson & Weiss, 2002; Kaufman, 2010; Wechsler, 1981a), we did not expect every indictor to exhibit scalar invariance.

Byrne et al. (1989) used the term *partial measurement invariance* to describe the situation where some, but not all, of an instrument's indicators are invariant. There is no consensus as to the minimum number of invariant indicators needed in a factor model to make comparisons, as scholars have argued for as small as one (Steenkamp & Baumgartner, 1998) to as large as a majority of the indicators (Vandenberg & Lance, 2000). Consequently, we recorded the number of invariant loadings and intercepts for each between-edition comparison and then examined if this was related to the magnitude of the differences in the latent means.

Evaluating Model Fit

When evaluating the adequacy of a particular MG-CFA model, we examined multiple indices (McDonald & Ho, 2002) that represent a variety of fit criteria (Marsh, Hau, & Grayson, 2005). Specifically, we examined (a) the χ^2, (b) the comparative fit index (CFI), (c) Tucker-Lewis index (TLI), (d) root mean square error of approximation (RMSEA), and (e) standardized root mean square residual (SRMR). For all models, we looked for patterns in the fit statistics, and judged acceptance/rejection of the specific model based on the majority of the indices. For this study's criteria of overall model-data fit, we used the following: (a) CFI $\geq$ 0.95; (b) TLI $\geq$ 0.95; (c) RMSEA $\leq$ 0.06; and (d) SRMR $\leq$ 0.08, (Browne & Cudeck, 1993; Hu & Bentler, 1999; Sivo, Fan, Witta, & Willse, 2006; Yu, 2002).

Model modification is usually an inevitable process in fitting latent variable models, especially when examining multiple groups (Chou & Huh, 2012). For this study, if a particular model did not fit the data well based on the aforementioned criteria, we examined the modification indexes (MIs) for parameter constraints to release within a group or between groups. We used two criteria before deciding to use a MI. First, freeing a parameter had to make sense theoretically, so we only considered releasing constraints on intercepts and factor loadings between groups, and releasing residual covariances within a group. Second, the MI had to be higher than a threshold value of 3.84 (Brown, 2006; Hoyle, 2011; Lei & Wu, 2007). The parameters were freed in a stepwise manner such that the constraint that made sense to release that was associated with the largest MI was freed first and then we re-examined the model. This procedure continued until the subsequent model fit adequately.

Data Analysis

All analyses were done using the Mplus (Muthén & Muthén, 2010) and R (R Development Core Team, 2013)

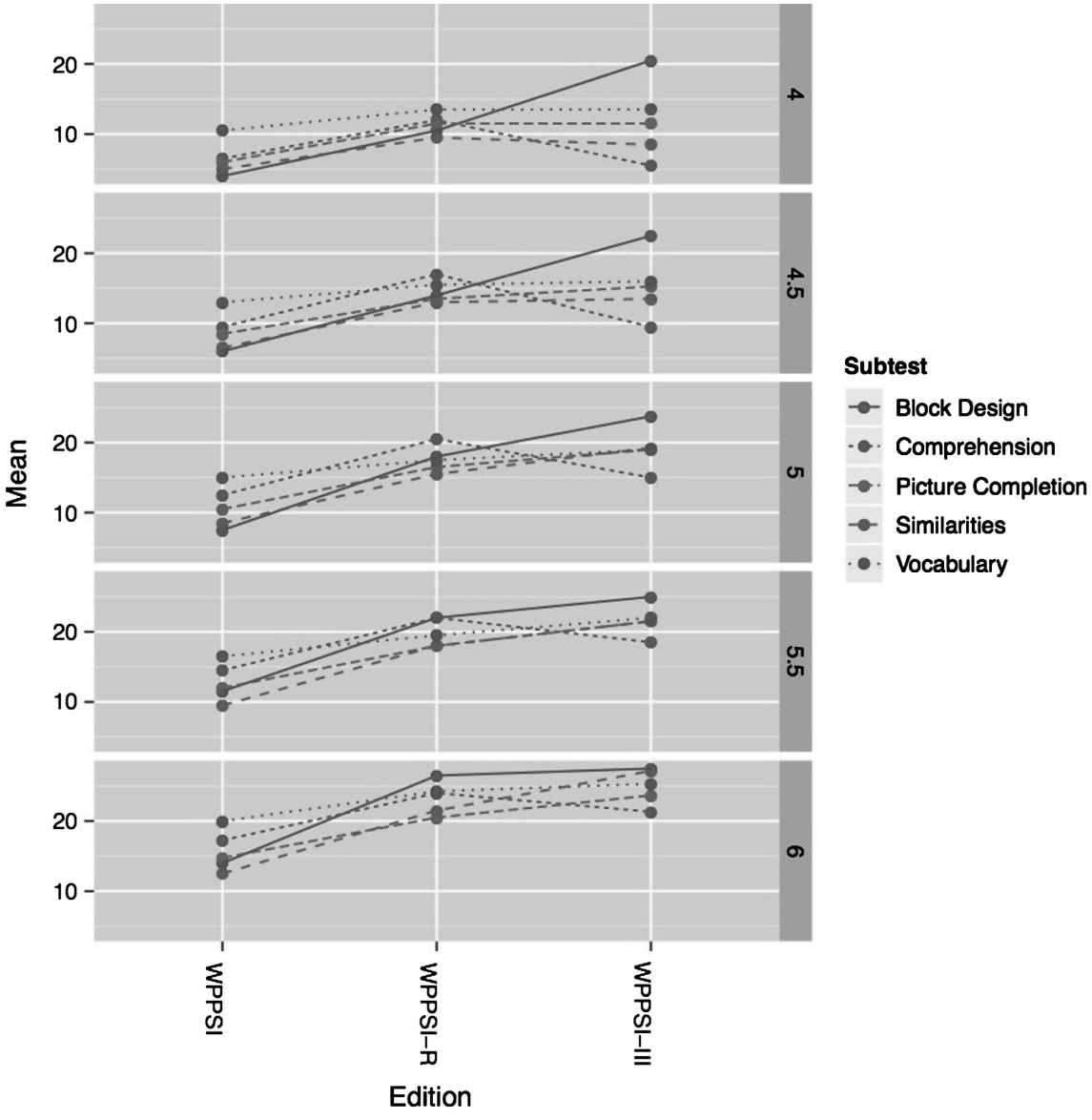

Figure 1. Mean raw score differences in the Wechsler Preschool and Primary Scale of Intelligence as a function of edition, subtest, and age group. The mean score is the raw score equivalent of a standard score of 10.

statistical programming languages. Within R, we used the *lavaan* (Rosseel, 2012) package.

Results

The average raw score for each subtest (i.e., the raw score equivalent for a standard score of 10) for each age group is given in Figures 1, 2, and 3 for the WPPSI, WISC, and WAIS, respectively. For some subtests (e.g., Block Design in the WPPSI) there was an increase in the raw score average age over time for all the age groups, whereas for other subtests (e.g., Vocabulary in the WISC) there was little-to-no change in mean score.

Wechsler Preschool and Primary Scale of Intelligence

The results from the latent variable analysis are given in Table 2. The number of invariant loadings and intercepts for a given age group ranged from 30% to 10%, although for the 5- and 5.5-year-old groups there were no invariant indicators between the second and third edition.

The effects differed substantially among the age groups. The 4-, 5-, and 5.5-year-old groups exhibited a moderate increase from the first to second edition (range: 0.63–0.75 *SD* units) and then a small decrease from the second to third edition (0.02–0.23 *SD* units). The 4.5-year-old group showed practically no change across all three editions, while the 6-year-old group showed large changes between

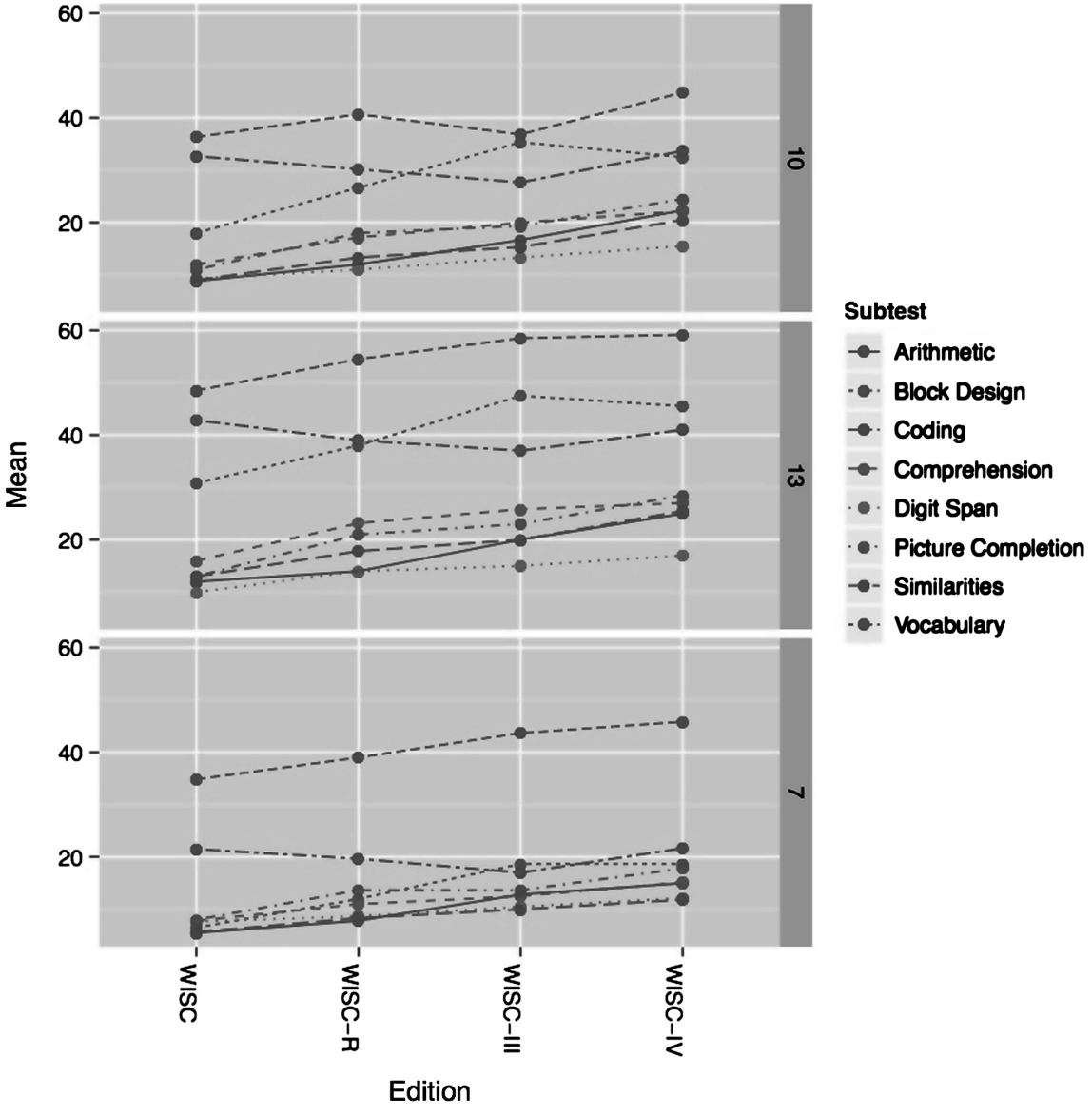

Figure 2. Mean raw score differences in the Wechsler Intelligence Scale for Children as a function of edition, subtest, and age group. The mean score is the raw score equivalent of a standard score of 10.

the first and second (2.44 *SD* units) and the second to the third (1.3 *SD* units). Despite the variability in mean differences, the precision of the mean scores was consistent for all age groups and all editions as the standard errors ranged from 0.11 to 0.17.

Following Flynn and Weiss (2007), we computed the sum of the subtests' mean observed scores to compare with the latent mean scores. To make the summed scores more interpretable, we subtracted the first editions' score from the subsequent editions and divided the difference by the standard deviation of the first edition. As with the latent mean scores, there was a general increase across time, although the magnitude of the increase was substantially larger for all ages except the 6-year-old group, where the increase was somewhat smaller.

Wechsler Intelligence Scale for Children

The results from the latent variable analysis are given in Table 3. The number of invariant loadings and intercepts for a given age group ranged from 23.44 to 51.56%, although there were no invariant indicators between the second and fourth edition for the 13-year-old group.

Unlike with the WPPSI, the effects are relatively consistent across the three age groups. There was a relatively large average score increase from the first to second edition (between 0.97 and 1.35 *SD* units), from the second to third edition (between 0.70 and 1.19 *SD* units), and from the third and fourth edition (between 0.45 and 1.37 *SD* units). The precision of the mean scores was relatively consistent for the age groups and editions, ranging from 0.12 to 0.24,

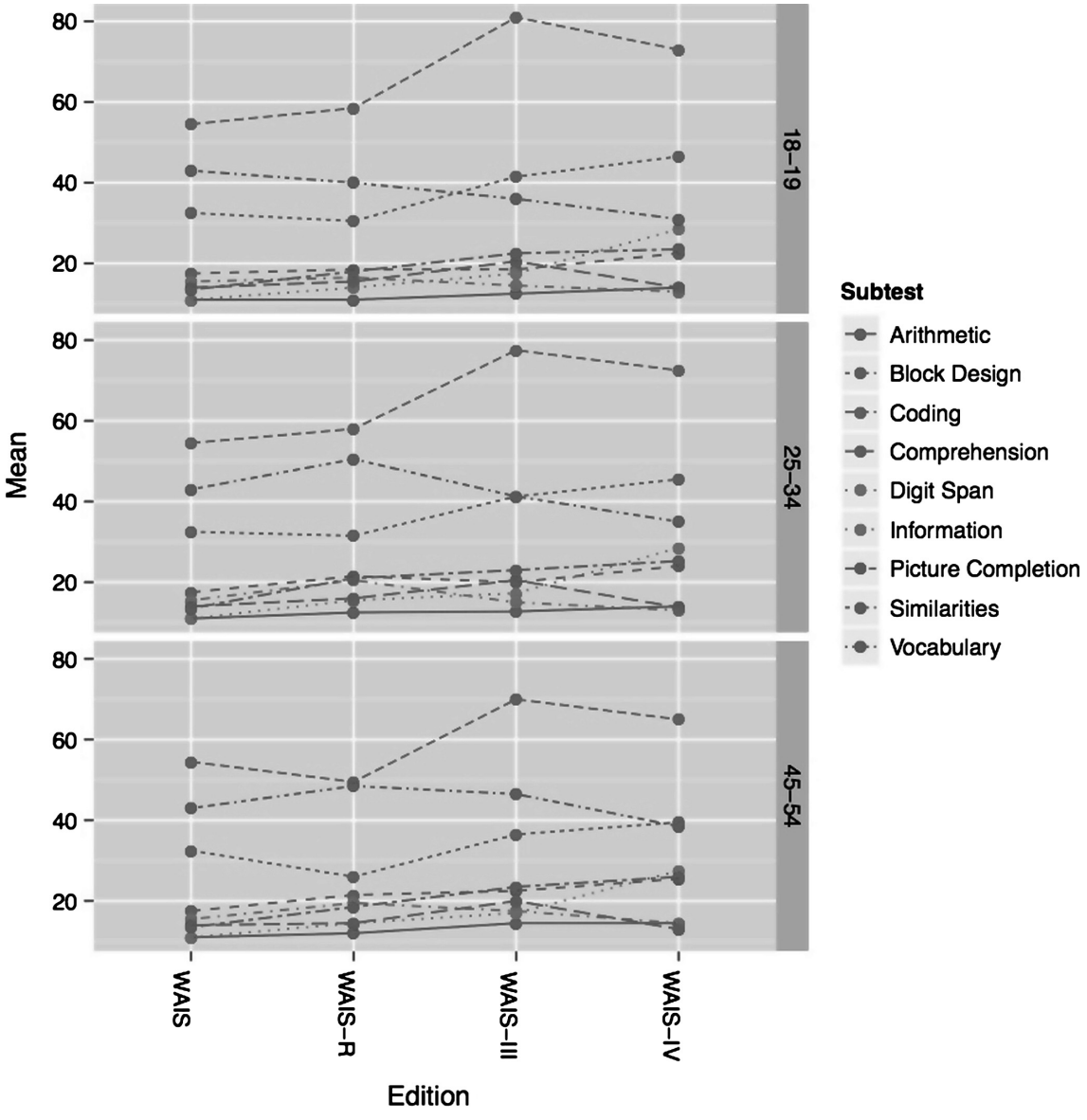

Figure 3. Mean raw score differences in the Wechsler Adult Intelligence Scale as a function of edition, subtest, and age group. The mean score is the raw score equivalent of a standard score of 10.

although the fourth edition's scores tended to be less precise than the second or third edition's score for all age groups.

The subtests' mean observed scores vacillated between being larger and smaller than the latent mean scores for the 7-year-old group, while for the 10-year-old group the observed mean scores were consistently lower than the latent mean scores. For the 13-year-old group, which had zero invariant indicators, the latent and observed mean scores showed very little difference.

Wechsler Adult Intelligence Scale

The results from the latent variable analysis are given in Table 4. The number of invariant indicators ranged from 23.61% to 44.44%. Across all editions, there was an increase in the latent variables' means.

Across all age groups, there was an increase from the first to second edition, ranging between 0.20 and 0.52 *SD* units. There was an increase from the second to the third edition as well, although there was much variability in amount of increase, ranging from 0.11 to 0.94 *SD* units. Likewise, there was an increase from the third to the fourth edition, but much variability in the increase, ranging from 0.07 to 0.62 *SD* units. Despite the variability in mean increases, the precision of the mean scores was very consistent for all age groups and all editions as the standard errors ranged from 0.08 to 0.13.

The subtests' mean observed scores vacillated between being larger and smaller than the latent mean scores for

Table 2. Mean scores on the Wechsler Preschool and Primary Scales of Intelligence

		Edition		
Age	Statistic	First	Second	Third
4	Latent mean	0.00	0.74	0.65
	Standard error	(–)	(0.13)	(0.14)
	Observed mean	0.00	1.61	1.77
4.5	Latent mean	0.00	−0.10	−0.00
	Standard error	(–)	(0.14)	(0.12)
	Observed mean	0.00	1.67	1.88
5[a]	Latent mean	0.00	0.63	0.40
	Standard error	(–)	(0.15)	(0.11)
	Observed mean	0.00	2.00	2.47
5.5[a]	Latent mean	0.00	0.65	0.63
	Standard error	(–)	(0.13)	(0.12)
	Observed mean	0.00	1.84	2.31
6	Latent mean	0.00	2.44	3.80
	Standard error	(–)	(0.12)	(0.17)
	Observed mean	0.00	2.07	2.52

Notes. For both the latent and observed mean, the metric used is distance (in standard deviation units) from the first edition's mean score. The standard deviation was calculated using the variances and covariances of the subtest scores (Nunnally & Bernstein, 1994, Equation 5-3a, p. 162). The proportion of invariant loadings and intercepts for each age group is as follows: 4 (30%), 4.5 (30%), 5 (10%), 5.5 (10%), and 6 (43.33%). [a]There were no invariant indicators between the second and third edition.

Table 3. Mean scores on the Wechsler Intelligence Scale for children

		Edition			
Age	Statistic	First	Second	Third	Fourth
7	Latent mean	0.00	0.97	2.16	3.53
	Standard error	(–)	(0.17)	(0.17)	(0.23)
	Observed mean	0.00	1.08	1.97	2.89
10	Latent mean	0.00	1.35	2.05	3.02
	Standard error	(–)	(0.12)	(0.14)	(0.16)
	Observed mean	0.00	1.07	1.59	2.65
13[a]	Latent mean	0.00	1.03	2.12	2.57
	Standard error	(–)	(0.16)	(0.16)	(0.24)
	Observed mean	0.00	1.06	1.81	2.47

Notes. For both the latent and observed mean, the metric used is distance (in standard deviation units) from the first edition's mean score. The standard deviation was calculated using the variances and covariances of the subtest scores (Nunnally & Bernstein, 1994, Equation 5-3a, p. 162). The proportion of invariant loadings and intercepts for each age group is as follows: 7 (51.56%), 10 (34.38%), and 13 (23.44%). [a]There were no invariant indicators between the second and fourth edition.

Table 4. Mean scores on the Wechsler Adult Intelligence Scale

		Edition			
Age	Statistic	First	Second	Third	Fourth
18–19	Latent mean	0.00	0.20	0.53	1.12
	Standard error	(–)	(0.11)	(0.13)	(0.11)
	Observed mean	0.00	0.17	0.9	0.93
25–34	Latent mean	0.00	0.52	0.63	1.25
	Standard error	(–)	(0.09)	(0.08)	(0.08)
	Observed mean	0.00	0.61	1.00	1.05
45–54	Latent mean	0.00	0.29	1.23	1.30
	Standard error	(–)	(0.11)	(0.11)	(0.12)
	Observed mean	0.00	0.21	0.97	0.9

Notes. For both the latent and observed mean, the metric used is distance (in standard deviation units) from the first edition's mean score. The standard deviation was calculated using the variances and covariances of the subtest scores (Nunnally & Bernstein, 1994, Equation 5-3a, p. 162). The proportion of invariant loadings and intercepts for each age group is as follows: 18–19 (44.44%), 25–34 (23.61%), and 45–54 (38.89%).

Table 5. Publication dates of Wechsler Instrument Editions

Edition	WPPSI	WISC	WAIS
First	1967	1949	1955
Second	1989	1974	1981
Third	2002	1991	1997
Fourth	–	2003	2008

Notes. WPPSI = Wechsler Preschool Primary Intelligence Test; WISC = Wechsler Intelligence Scale for Children; WAIS = Wechsler Adult Intelligence Test.

both the 18–19-year-old and 25–34-year-old groups, while for the 45–54-year-old group the observed mean scores were consistently lower than the latent mean scores.

Scores by Time

As there were different amounts of time between the publications of one edition to another (see Table 5), we examined the relationship between the time between-edition publication and changes in the latent variable scores within an instrument using multiple regression. In addition, we examined the influence of the amount of between-edition invariance using the same regression models.

The regression results are given in Table 6. For all models, we first removed the three comparisons with no invariant indicators. We entered the variables in a hierarchical manner to examine the effects of individual predictors and their interactions (Aiken & West, 1991; Beaujean, 2008). Initially, we included only the number of years between editions as the predictor (Model 1), then added

Table 6. Hierarchical regression analysis summary for predicting differences in average latent ability (n = 48)

Model	Variable	B	SE	β	p	r_s	R^2
1	Years	0.03	0.01	0.46	< .00	1.00	0.21
2	Years	0.03	0.01	0.46	< .00	0.69	0.45
	WPPSI	0.20	0.27	0.21	.46	−0.32	
	WISC	1.02	0.25	1.08	< .00	0.74	
3	Years	0.04	0.01	0.54	< .00	0.63	0.54
	WPPSI	0.19	0.25	0.2	.44	−0.29	
	WISC	0.98	0.23	1.04	< .00	0.68	
	Invariance	2.43	0.84	0.31	< .01	0.26	
4	Years	0.02	0.01	0.31	.07	0.60	0.58
	WPPSI	0.18	0.25	0.19	.48	−0.28	
	WISC	0.94	0.22	0.99	< .00	0.65	
	Invariance	2.55	0.83	0.33	< .00	0.25	
	WPPSI × Years	0.02	0.02	0.30	.36	0.34	
	WISC × Years	0.03	0.01	0.43	< .06	0.56	

Notes. B, β, and r_s are the unstandardized regression coefficients, standardized regression coefficients, and structure coefficients, respectively. The editions were dummy coded using the WAIS as the reference group. The Years and Invariance variables were mean centered (means were 27.19 and 0.25, respectively). The constants for models 1, 2, 3, and 4 are 1.07, 0.66, 0.67, and 0.70 respectively. All p-values are two-tailed.

the Wechsler edition as a predictor (Model 2), then added the amount of invariance as a predictor (Model 3), and finally added interaction terms between the edition and the number of years between editions (Model 4).

Adding the Wechsler edition to the model explained an additional 23.66% of the variance in latent variable score differences than years alone, which explained 21.09% of the variance. Most of this additional explained variance, however, is due to the difference between the WISC and WAIS instruments.

Next, we added the amount of invariance between instruments to the model. While this explained an additional 8.99% of the variance in latent variable score differences, the standardized and structure coefficients are relatively small, indicating it does not have a strong relationship to score differences, at least not as strong as the years and edition variables.[1] When adding the interaction terms, there was a small increase in the amount of explained variance (3.97%), but the standardized and structure coefficients are larger than those associated with the WPPSI, amount of invariance, and years main effects. Moreover, the coefficients associated with the WISC × Years interaction are not much smaller than those associated with the WISC main effect. Thus, it appears that keeping the interaction terms is warranted.

A graph of the interaction is given in Figure 4. An interpretation of the interaction is as follows. When ignoring the Wechsler instrument used, there is a FE of 0.46 IQ points/year. When accounting for the type of Wechsler instrument and the amount of invariance, the WPPSI and WISC instruments show an average FE of 0.61 and 0.73 IQ points/year, but the WAIS instruments, on average, only show a FE of 0.30 IQ points/year.

Discussion

This study investigated changes over time in the three most commonly used Wechsler intelligence scales: WPPSI (Wechsler, 1967, 1989, 2002), WISC (WISC, 1949; Wechsler, 1974, 1991, 2003), and WAIS (Wechsler, 1955, 1981b, 1997, 2008). We formed covariance matrices for each edition of each instrument by converting the subtest correlation matrices into covariance matrices by reverse engineering each instrument's standard score conversion tables to obtain the mean and standard deviations in raw score units. From the covariance matrices, we then formed single factor latent variable models and examined if there was invariance in the factor loading and intercepts (i.e., scalar invariance) via a MG-CFA. The results indicated that there does appear to be an average FE of 0.44 IQ points/year across all the Wechsler scales. This result, however, must be tempered with both the invariance findings, and the interaction between years and Wechsler instrument. While the WISC and WPPSI had many invariant indicators across all their age groups, only the 4-year-old and 6-year-old age groups on the WPPSI, and the 7-year-old group on the WISC had half or more of the indicators be invariant. Conversely, the 5- and 5.5-year-old groups on the WPPSI

[1] The difference in signs for the WPPSI variable for the structure coefficient (r_s) than the regression coefficients (B and β) is due to the change in meaning of the dummy coding for the coefficients. r_s are zero-order correlations (Pedhazur, 1997), so the WPPSI variable is comparing scores from the WPPSI to combined WISC and WAIS scores. B and β, however, account for all the other variables in the model, so the WPPSI variable is comparing the WPPSI scores with only the reference group (i.e., the WAIS scores).

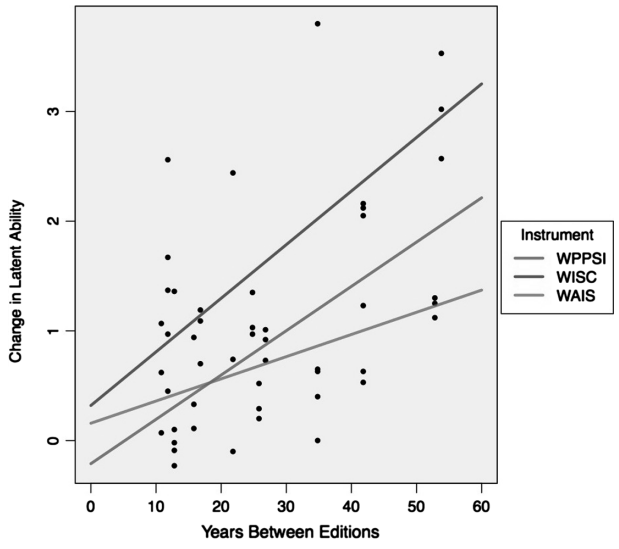

Figure 4. Differences in average latent ability as a function of years, amount of invariance, and Wechsler edition. Computed slopes of regression line for Wechsler edition and years interaction, with amount of invariance held at the mean (i.e., 0), are presented.

had no invariant indicators between the second and third editions, nor did the second and fourth editions of the WISC for the 13-year-old age group. Moreover, the interaction between the instrument and years indicate that the WPPSI and WISC showed, on average, similar FEs of 0.61 and 0.73 IQ points/year, respectively. The WAIS instruments, however, showed an average FE of 0.30 IQ points/year.

Integration With Other Research Examining the Comparability of Scores

To date, few studies have examined the comparability of IQ-type scores in the context of the FE. Kane (2000) examined the correlations among the subtests for all age groups of the WAIS, WAIS-R, and WAIS-III and found a decrease in average correlation between WAIS and WAIS-R, but little difference between WAIS-R and WAIS-III. They concluded that the Flynn Effect exerted its greatest influence between 1955 and 1981 (the standardization dates of the first and second WAIS editions), but decelerated during the 16-year period between the second and third editions of the WAIS, resulting in a leveling off in the secular decline of Spearman's *g*. The current study's results mimic Kane's (2000) results for the 45–54-year-old age group, but not the 18–19- or 25–34-year-old groups, indicating that the FE, at least in WAIS scales, is moderated by age.

In one of the more comprehensive comparability studies, Floyd et al. (2008) found minimal differences between test batteries that influenced the FSIQ score (or its equivalent) school-aged children. For adults, however, they found that more than 20% of the total variance in FSIQ-type scores is attributable to the differences in test batteries. Moreover, it was the WAIS (specifically the third edition) that produced IQ estimates that were notably higher (on average between 5 and 9 IQ points) than other instruments. The results for the current study show an increase from the first and second edition of the WAIS to the third edition, but the results, at least for the 18–19 and 25–34-year-old groups, show continued increases to the fourth edition, indicating that there might generally be something with the WAIS that is conducive to showing a FE instead of something with a specific edition. This argument is strengthened when combined with the results from the WISC analysis of the current study. As with the WAIS, the WISC shows a continual increase in latent ability across all four editions. This interpretation needs to be qualified, though, as the results from the WPSSI show a different pattern from the WAIS and WISC.

Limitations

The largest limitation of the current study is the scores used for the data analysis. In the best situation, there would be raw scores available from all editions of a given Wechsler instrument that were either collected using the same sample at one time point or on comparable samples at various time points. Such data, to the best of our knowledge, do not exist. Thus, the current study had to reverse engineer the Wechsler technical manuals to obtain mean and covariance matrices.

As the data for the current study had to be obtained from the Wechsler technical manuals, we were limited to only using age groups and subtests that were common across all editions of an instrument. While the number of common subtests was substantial for the WAIS (10 subtests) and WISC (9 subtests) it was much less for the WPPSI (6 subtests). Conversely, the WPPSI had five age groups in common, while the WISC and WAIS only had three age groups. Thus, the comparisons we made, and differences we found, could be the result of only being able to use part of the total number of subtests and only part of the total number of age groups.

Last, we conceptualized cognitive ability as general latent variable, *g*, and measured the FE as changes in this latent variable. While there is much support for this conceptualization (Gottfredson, 2002; Jensen, 1998; Kuncel, Hezlett, & Ones, 2004), it is not the only one. Flynn (2003, 2007), for example, argues that a better approach to understanding changes in cognitive ability is to examine changes in functional skills and the potency of active social multipliers.

Implications from the Current Study

First, although many studies have used scores from Wechsler instruments to examine the FE (e.g., Flynn, 2007; Kanaya & Ceci, 2011; Neisser, 1998), they all have used traditional mean comparisons of the FSIQ (or another index score) and assuming that the manifest scores are directly comparable (i.e., all subtests have at least scalar invariance). The results from the current study call the

assumptions for such comparisons into question. In the best case scenario there are nine overlapping subtests (although not all contribute to calculating the FSIQ, e.g., Arithmetic) and 55% of the subtests' loadings and intercepts show invariance across all editions. This best case scenario only occurred for one instrument (WISC) and one age group (7-year-olds). At its worse (e.g., the 5- and 5.5-year-old groups for the WPPSI), there are only five common subtests and 10% of the subtests' loadings and intercepts show invariance across all editions, with there being no invariant subtests betweens the second and third editions. While the amount of invariance did not have an appreciable influence on the score differences in the current study, this is likely because of the simultaneous estimation of parameters for a given age group (Kolen & Brennan, 2004).

When we compared the sum of the observed score means to the latent variable means, there was no systematic pattern. Sometimes the differences in latent means were higher than the observed means, and sometimes they were lower; sometimes the differences were large (e.g., 2.07 *SD* in WPPSI and WPPSI-III comparison for the 5-year-old age group), and sometimes the differences were small (e.g., 0.03 *SD*s in the WISC and WISC-R comparison for the 13-year-old age group). Without being able to better understand the reason for such differences, the comparison of Wechsler scores across editions, at least those comparisons that do not account for nonequivalence, likely adds considerable bias to the mean-difference estimates.

Second, although the FE is often estimated to be 0.30 IQ points/year, this estimate should be given with an asterisk that it is likely moderated by a host of factors, such as age of the respondents and the editions being compared, at least when using Wechsler scales.

For example, based on the current study's results, while the 0.30 estimate would be applicable for the average WAIS scores, it would be an underestimate for WISC scores. Moreover, the WPSSI scores showed much variability, both in pattern and magnitude of the score changes. This variability was so large that we are probably better off without giving an average FE for the WPPSI and, instead, looking at the results by age group.

From a more practical perspective, these results do indicate that an application of a FE correction for death penalty (Flynn, 2006; Reynolds, Niland, Wright, & Rosenn, 2010) might continue to be warranted, as the WISC and WAIS instruments both showed continued score increases across all editions. Nonetheless, the magnitude of such a correction needs more consideration, as the amount of change showed considerable variability between different editions and for different age groups (cf. Kanaya & Ceci, 2007a). "Although a uniform adjustment will be a better and fairer [adjustment] than no adjustment at all, it should only be regarded as a temporary solution" (Kanaya & Ceci, 2007b, p. 63).

Future Directions

Future studies should specifically examine possible reasons *why* the FE occurs poignantly, yet rather unsystematically, in the Wechsler scales. Flynn (2006) has previously suggested that atypical or substandard norms might be a reason. The demographic information for each standardization sample is given in each of the Wechsler administration or technical manuals, while such an analysis would be costly in terms of time in acquiring the information and coding the data (different demographic information is presented in different manuals), such a hypothesis is now testable using the results in Tables 2, 3, and 4.

Another possible causal area to examine in future analyses of the Wechsler scales is the difference in manifest content. The current study examined invariance statistically, but did not examine the content of the subtests where there was a lack of invariance. Such examinations should examine not only number of items and item characteristics (e.g., difficulty), but also subtest placement and whether the subtest is used in the core battery (i.e., contributes to the FSIQ).

A third area for future investigation would be to do a similar type of analysis using other intelligence tests. The Stanford-Binet (Roid, 2003) instruments, for example, have a long history with multiple revisions (Becker, 2003) and provide subtest correlations and standard score conversion tables. Using the reverse engineering steps outlined in the Method section, such FE examinations would be doable after acquiring the appropriate technical manuals.

Electronic Supplementary Material

All other matrices can be found at http://blogs.baylor.edu/psychometric_lab/files/2013/05/FEWechslerMatrices-1qhip1y.pdf

References

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.

Alfonso, V. C., Oakland, T. D., LaRocca, R., & Spanakos, A. (2000). The course on individual cognitive assessment. *School Psychology Review, 29*, 52–64.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (Revised 4th ed.). Washington, DC: Author.

Archer, R. P., & Newsom, C. R. (2000). Psychological test usage with adolescent clients: Survey update. *Assessment, 7*, 227–235. doi: 10.1177/107319110000700303

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

Baxendale, S. (2010). The Flynn effect and memory function. *Journal of Clinical and Experimental Neuropsychology, 32*, 699–703. doi: 10.1080/13803390903493515

Beaujean, A. A. (2008). Mediation, moderation, and the study of individual differences. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 422–442). Thousand Oaks, CA: Sage.

Beaujean, A. A., & Osterlind, S. J. (2008). Using item response theory to assess the Flynn effect in the National Longitudinal Study of Youth 79 Children and Young Adults data. *Intelligence, 36*, 455–463. doi: 10.1016/j.intell.2007.10.004

Beaujean, A. A., & Sheng, Y. (2010). Examining the Flynn effect in the general social survey vocabulary test using item

response theory. *Personality and Individual Differences, 48*, 294–298. doi: 10.1016/j.paid.2009.10.019

Becker, K. A. (2003). *History of the Stanford-Binet Intelligence Scales: Content and psychometrics*. Itasca, IL: Riverside.

Boake, C. (2002). From the Binet-Simon to the Wechsler-Bellevue: Tracing the history of intelligence testing. *Journal of Clinical and Experimental Neuropsychology, 24*, 383–405. doi: 10.1076/jcen.24.3.383.981

Bontempo, D. E., & Hofer, S. M. (2007). Assessing factorial invariance in cross-sectional and longitudinal studies. In A. D. Ong & M. van Dulmen (Eds.), *Handbook of methods in positive psychology* (pp. 153–175). New York, NY: Oxford University Press.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Byrne, B. M., Shavelson, R. J., & Muthen, B. O. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*, 456–466. doi: 10.1037/0033-2909.105.3.456

Camara, W. J., Nathan, J. S., & Puente, A. E. (2000). Psychological test usage: Implications in professional psychology. *Professional Psychology: Research and Practice, 31*, 141–154. doi: 10.1037/0735-7028.31.2.141

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York, NY: Cambridge University Press.

Chan, D. (1998). The conceptualization and analysis of change over time: An integrative approach incorporating longitudinal mean and covariance structures analysis (LMACS) and multiple indicator latent growth modeling (MLGM). *Organizational Research Methods, 1*, 421–483. doi: 10.1177/109442819814004

Chou, C.-P., & Huh, J. (2012). Model modification in structural equation modeling. In R. A. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 232–246). New York, NY: Guilford.

Coalson, D., & Weiss, L. G. (2002). The evolution of Wechsler intelligence scales in historical perspective. *Assessment Focus Newsletter, 11*, 1–3.

Colom, R., Lluis-Font, J. M., & Andrés-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.

Cudeck, R. (1989). Analysis of correlation matrices using covariance structure models. *Psychological Bulletin, 105*, 317–327. doi: 10.1037/0033-2909.105.2.317

Floyd, R. G., Clark, M. H., & Shadish, W. R. (2008). The exchangeability of IQs: Implications for professional psychology. *Professional Psychology: Research and Practice, 39*, 414–423. doi: 10.1037/0735-7028.39.4.414

Flynn, J. R. (1982). Lynn, the Japanese, and environmentalism. *Bulletin of the British Psychological Society, 35*, 409–413.

Flynn, J. R. (1983). Now the great augmentation of the American IQ. *Nature, 301*, 65. doi: 10.1038/301655a0

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin, 95*, 29–51. doi: 10.1037/0033-2909.95.1.29

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

Flynn, J. R. (2000). The hidden history of IQ and special education: Can the problems be solved? *Psychology, Public Policy, and Law, 6*, 191–198. doi: 10.1037/1076-8971.6.1.191

Flynn, J. R. (2003). Movies about intelligence: The limitations of *g*. *Current Directions in Psychological Science, 12*, 95–99.

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

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

Flynn, J. R. (2009). The WAIS-III and WAIS-IV: Daubert motions favor the certainly false over the approximately true. *Applied Neuropsychology, 16*, 98–104. doi: 10.1080/09084280902864360

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

Flynn, J. R., & Widaman, K. F. (2008). The Flynn effect and the shadow of the past: Mental retardation and the indefensible and indispensible role of IQ. In L. M. Glidden (Ed.), *International review of mental retardation* (Vol. 35, pp. 121–149). Boston, MD: Elsevier.

Golembiewski, R. T., Billingsley, K., & Yeager, S. (1976). Measuring change and persistence in human affairs: Types of change generated by OD designs. *The Journal of Applied Behavioral Science, 12*, 133–157. doi: 10.1177/002188637601200201

Gottfredson, L. S. (2002). g: Highly general and highly practical. In R. J. Sternberg & E. L. Grigorenko (Eds.), *The general factor of intelligence: How general is it?* (pp. 331–380). Mahwah, NJ: Erlbaum.

Hagan, L. D., Drogin, E. Y., & Guilmette, T. J. (2010). IQ scores should not be adjusted for the Flynn effect in capital punishment cases. *Journal of Psychoeducational Assessment, 28*, 474–476. doi: 10.1177/0734282910373343

Herrnstein, R. J., & Murray, C. (1996). *The Bell curve: Intelligence and class structure in American life*. New York, NY: Free Press.

Hoyle, R. H. (2011). *Structural equation modeling for social and personality psychology*. London, UK: Sage.

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55. doi: 10.1080/10705519909540118

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger/Greenwood.

Kanaya, T., & Ceci, S. J. (2007a). Are all IQ scores created equal? The differential costs of IQ cutoff scores for at-risk children. *Child Development Perspectives, 1*, 52–56.

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

Kanaya, T., & Ceci, S. J. (2011). The Flynn effect in the WISC subtests among school children tested for special education services. *Journal of Psychoeducational Assessment, 29*, 125–136. doi: 10.1177/0734282910370139

Kanaya, T., Ceci, S. J., & Scullin, M. H. (2003). The rise and fall of IQ in special ed.: Historical trends and their implications. *Journal of School Psychology, 41*, 453–465.

Kane, H. D. (2000). A secular decline in Spearman's *g*: Evidence from the WAIS, WAIS-R and WAIS-III. *Personality and Individual Differences, 29*, 561–566. doi: 10.1016/s0191-8869(99)00217-2

Kane, H. D., & Oakland, T. D. (2000). Secular declines in Spearman's *g*: Some evidence from the United States. *The Journal of Genetic Psychology: Research and Theory on Human Development, 161*, 337–345. doi: 10.1080/00221320009596716

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. doi: 10.1177/0734282910373346

Kaufman, A. S., & Weiss, L. G. (2010). Guest editors' introduction to the special issue of JPA on the Flynn effect. *Journal of Psychoeducational Assessment, 28*, 379–381. doi: 10.1177/0734282910373344

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods, 17*, 137–152. doi: 10.1037/a0028086

Khaleefa, O., Sulman, A., & Lynn, R. (2009). An increase of intelligence in Sudan, 1987–2007. *Journal of Biosocial Science, 41*, 279–283. doi: 10.1017/s0021932008003180

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2004). Academic performance, career potential, creativity, and job performance: Can one construct predict them all? *Journal of Personality and Social Psychology, 86*, 148–161. doi: 10.1037/0022-3514.86.1.148

Lei, P.-W., & Wu, Q. (2007). Introduction to structural equation modeling: Issues and practical considerations. *Educational Measurement: Issues and Practice, 26*, 33–43. doi: 10.1111/j.1745-3992.2007.00099.x

Lewis, M., Sanborn, K., McGreevy, S., Tarquin, K., & Truscott, S. D. (2004). What every school psychologist should know about the Flynn effect. *NASP Communiqué, 32*, 29–32.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6*, 83–102. doi: 10.1207/s15324818ame0601_5

Little, T. D., & Slegers, D. W. (2005). Factor analysis: Multiple Groups. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science* (Vol. 2, pp. 617–623). Chichester, UK: Wiley.

Lubin, B., & Lubin, A. W. (1972). Patterns of Psychological Services in the U.S.: 1959–1969. *Professional Psychology, 3*, 63–65. doi: 10.1037/h0021498

Lynn, R. (1982). IQ in Japan and the United States shows a growing disparity. *Nature, 306*, 291–292.

Lynn, R. (1987). Japan: Land of the rising IQ: A reply to Flynn. *Bulletin of the British Psychological Society, 40*, 464–468.

Lynn, R. (1990). Differential rates of secular increase of five major primary abilities. *Biodemography and Social Biology, 37*, 137–141. doi: 10.1080/19485565.1990.9988753

Lynn, R. (2009). Fluid intelligence but not vocabulary has increased in Britain, 1979–2008. *Intelligence, 37*, 249–255. doi: 10.1016/j.intell.2008.09.007

Lynn, R. (2009). What has caused the Flynn effect? Secular increases in the development quotients of infants. *Intelligence, 37*, 16–24. doi: 10.1016/j.intell.2008.07.008

Marsh, H. W., Hau, K.-T., & Grayson, D. (2005). Goodness of fit in structural equation models. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics: A festschrift for Roderick P. McDonald* (pp. 275–340). Mahwah, NJ: Erlbaum.

McDonald, R. P., & Ho, M.-H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods, 7*, 64–82. doi: 10.1037/1082-989X.7.1.64

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.

Millsap, R. E., & Hartog, S. B. (1988). Alpha, beta, and gamma change in evaluation research: A structural equation approach. *Journal of Applied Psychology, 73*, 574–584. doi: 10.1037/0021-9010.73.3.574

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects*. Princeton, NJ: Educational Testing Service.

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

Muthén, L. K., & Muthén, B. O. (2010). Mplus (version 6) [Computer software]. Los Angeles, CA: Muthén and Muthén.

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

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

Nugent, W. R. (2006). The comparability of the standardized mean difference effect size across different measures of the same construct. *Educational and Psychological Measurement, 66*, 612–623. doi: 10.1177/0013164405284032

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory* (3rd ed.). New York, NY: McGraw-Hill.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research* (3rd ed.). Fort Worth, TX: Harcourt Brace.

Ployhart, R. E., & Vandenberg, R. J. (2010). Longitudinal research: The theory, design, and analysis of change. *Journal of Management, 36*, 94–120. doi: 10.1177/0149206309352110

R Development Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available from http://www.R-project.org

Reschly, D. J., Myers, T. G., & Hartel, C. R. (Eds.). (2002). *Mental retardation: Determining eligibility for social security benefits*. Washington, DC: National Academies Press.

Reynolds, C. R., Niland, J., Wright, J. E., & Rosenn, M. (2010). Failure to apply the Flynn correction in death penalty litigation: Standard practice of today maybe, but certainly malpractice of tomorrow. *Journal of Psychoeducational Assessment, 28*, 477–481. doi: 10.1177/0734282910373348

Rodgers, J. L. (1998). A critique of the Flynn effect: Massive IQ gains, methodological artifacts, or both? *Intelligence, 26*, 337–356. doi: 10.1016/s0160-2896(99)00004-5

Roid, G. H. (2003). *Stanford-Binet intelligence scales* (5th ed.). Itasca, IL: Riverside.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*, 1–36.

Rudnick, A. (2001). Ethics of ECT for children. *Journal of the American Academy of Child & Adolescent Psychiatry, 40*, 387–388.

Sanborn, K. J., Truscott, S. D., Phelps, L., & McDougal, J. L. (2003). Does the Flynn effect differ by IQ level in samples of students classified as learning disabled? *Journal of Psycheducational Assessment, 21*, 145–159.

Schaie, K. W., Willis, S. L., & Pennak, S. (2005). An historical framework for cohort differences in intelligence. *Research in Human Development, 2*, 43–67.

Shayer, M., Ginsburg, D., & Coe, R. (2007). Thirty years on – a large anti-Flynn effect? The Piagetian test Volume Heaviness norms 1975–2003. *British Journal of Educational Psychology, 77*, 25–41. doi: 10.1348/000709906x96987

Shiu, W., & Beaujean, A. A. (2010, December). *The Flynn effect in adults: A meta-analysis.* Poster presented at the annual meeting of the International Society for Intelligence Research, Arlington, VA.

Sivo, S. A., Fan, X., Witta, E. L., & Willse, J. T. (2006). The search for "optimal" cutoff properties: Fit index criteria in structural equation modeling. *Journal of Experimental Education, 74*, 267–288. doi: 10.3200/JEXE.74.3.267-288

Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer

research. *Journal of Consumer Research, 25*, 78–107. doi: 10.1086/209528

Stinnett, T. A., Havey, J. M., & Oehler-Stinnett, J. (1994). Current test usage by practicing school psychologists: A national survey. *Journal of Psychoeducational Assessment, 12*, 331–350. doi: 10.1177/073428299401200403

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. doi: 10.1016/j.paid.2005.01.029

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

te Nijenhuis, J. (2013). The Flynn effect, group differences, and *g* loadings. *Personality and Individual Differences, 55*, 224–228. doi: 10.1016/j.paid.2011.12.023

te Nijenhuis, J., Murphy, R., & van Eeden, R. (2011). The Flynn effect in South Africa. *Intelligence, 39*, 456–467. doi: 10.1016/j.intell.2011.08.003

Thorndike, R. L., & Lohman, D. (1990). *A century of ability testing*. Chicago, IL: Riverside.

Truscott, S. D., & Frank, A. J. (2001). Does the Flynn effect affect IQ scores of students classified as LD? *Journal of School Psychology, 89*, 319–334.

Truscott, S. D., & Volker, M. A. (2005). The Flynn effect and LD classification: Empirical evidence of IQ score changes that could affect diagnosis. In A. Columbus (Ed.), *Advances in psychology research* (Vol. 35, pp. 173–204). New York, NY: Nova Science.

Tuddenham, R. D. (1948). Soldier intelligence in World Wars I and II. *American Psychologist, 3*, 54–56.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4–70. doi: 10.1177/109442810031002

VandenBos G. R. (Ed.). (2007). *APA dictionary of psychology*. Washington, DC: American Psychological Association.

Wai, J., & Putallaz, M. (2012). The Flynn effect puzzle: A 30-year examination from the right tail of the ability distribution provides some missing pieces. *Intelligence, 39*, 443–455. doi: 10.1016/j.intell.2011.07.006

Wechsler, D. (1949). *Wechsler Intelligence Scale for children*. San Antonio, TX: Psychological Corporation.

Wechsler, D. (1955). *Wechsler adult intelligence scale*. New York, NY: The Psychological Corporation.

Wechsler, D. (1967). *Wechsler preschool and primary scale of intelligence*. New York, NY: The Psychological Corporation.

Wechsler, D. (1974). *Wechsler intelligence scale for children* (Revised ed.). New York, NY: The Psychological Corporation.

Wechsler, D. (1981). The psychometric tradition: Developing the wechsler adult intelligence scale. *Contemporary Educational Psychology, 6*, 82–85. doi: 10.1016/0361-476X(81)90035-7

Wechsler, D. (1981). *Wechsler adult intelligence scale* (Revised ed.). San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1989). *Wechsler preschool and primary scale of intelligence* (Revised ed.). San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1991). *Wechsler intelligence scale for children* (3rd ed.). San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1997). *Wechsler adult intelligence scale* (3rd ed.). San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2002). *Wechsler preschool and primary scale of intelligence* (3rd ed.). San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2003). *Wechsler intelligence scale for children* (4th ed.). San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2008). *Wechsler adult intelligence scale* (4th ed.). San Antonio, TX: The Psychological Corporation.

Weiss, K. J., Haskins, B., & Hauser, M. J. (2004). Commentary: Atkins and clinical practice. *Journal of American Academic Psychiatry Law, 32*, 309–313.

Wicherts, J. M., & Dolan, C. V. (2010). Measurement invariance in confirmatory factor analysis: An illustration using IQ test performance of minorities. *Educational Measurement: Issues and Practice, 29*, 39–47. doi: 10.1111/j.1745-3992.2010.00182.x

Wicherts, J. M., Dolan, C. V., Hessen, D. J., Oosterveld, P., van Baal, G. C. M., Boomsma, D. I., & Span, M. M. (2004). Are intelligence tests measurement invariant over time? Investigating the nature of the Flynn effect. *Intelligence, 32*, 509–537. doi: 10.1016/j.intell.2004.07.002

Yu, C.-Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes*, Unpublished doctoral dissertation.

Date of acceptance: January 13, 2014
Published online: June 2, 2014

Alexander Beaujean

Department of Educational Psychology
One Bear Place #97301
Baylor University, Waco, TX
USA
E-mail Alex_Beaujean@baylor.edu

Appendix

Sample Correlation Matrices, Means, and Standard Deviations of the Wechsler Subtests Used in This Study

Wechsler Intelligence Scale for Children

Table A1. WISC, 7-year-olds (n = 200)

	Comp	Arith	Sim	Voc	DigSp	PicComp	BlkDsgn	Cod
Comprehension	1.00	0.31	0.36	0.51	0.29	0.39	0.32	0.22
Arithmetic	0.31	1.00	0.40	0.46	0.40	0.29	0.27	0.32
Similarities	0.36	0.40	1.00	0.45	0.33	0.27	0.29	0.15
Vocabulary	0.51	0.46	0.45	1.00	0.43	0.36	0.33	0.22
Digit span	0.29	0.40	0.33	0.43	1.00	0.33	0.24	0.27
Picture completion	0.39	0.29	0.27	0.36	0.33	1.00	0.28	0.12
Block design	0.32	0.27	0.29	0.33	0.24	0.28	1.00	0.26
Coding	0.22	0.32	0.15	0.22	0.27	0.12	0.26	1.00
SD	2.69	1.50	2.36	6.06	1.85	2.18	5.97	9.94
Mean	7.83	5.50	5.67	21.50	7.67	8.00	6.50	34.83

Table A2. WISC-R, 7-year-olds (n = 200)

	Comp	Arith	Sim	Voc	DigSp	PicComp	BlkDsgn	Cod
Comprehension	1.00	0.43	0.45	0.61	0.34	0.33	0.43	0.18
Arithmetic	0.43	1.00	0.43	0.50	0.47	0.25	0.42	0.22
Similarities	0.45	0.43	1.00	0.59	0.34	0.36	0.50	0.18
Vocabulary	0.61	0.50	0.59	1.00	0.35	0.37	0.43	0.27
Digit span	0.34	0.47	0.34	0.35	1.00	0.18	0.32	0.18
Picture completion	0.33	0.25	0.36	0.37	0.18	1.00	0.42	0.13
Block design	0.43	0.42	0.50	0.43	0.32	0.42	1.00	0.26
Coding	0.18	0.22	0.18	0.27	0.18	0.13	0.26	1.00
SD	3.63	1.77	3.26	5.07	2.86	4.25	8.82	8.05
Mean	11.00	7.83	8.33	19.67	8.67	13.67	12.00	39.00

Table A3. WISC-III, 7-year-olds (n = 200)

	Comp	Arith	Sim	Voc	DigSp	PicComp	BlkDsgn	Cod
Comprehension	1.00	0.33	0.46	0.61	0.22	0.23	0.22	0.13
Arithmetic	0.33	1.00	0.46	0.41	0.39	0.30	0.36	0.22
Similarities	0.46	0.46	1.00	0.64	0.36	0.33	0.37	0.18
Vocabulary	0.61	0.41	0.64	1.00	0.30	0.31	0.35	0.19
Digit span	0.22	0.39	0.36	0.30	1.00	0.25	0.30	0.09
Picture completion	0.23	0.30	0.33	0.31	0.25	1.00	0.47	0.24
Block design	0.22	0.36	0.37	0.35	0.30	0.47	1.00	0.20
Coding	0.13	0.22	0.18	0.19	0.09	0.24	0.20	1.00
SD	3.52	2.19	3.19	4.76	2.53	3.85	10.25	10.51
Mean	12.50	12.83	10.00	17.00	10.50	13.67	18.67	43.67

Table A4. WISC-IV, 7-year-olds (n = 200)

	Comp	Arith	Sim	Voc	DigSp	PicComp	BlkDsgn	Cod
Comprehension	1.00	0.46	0.58	0.63	0.27	0.45	0.33	0.15
Arithmetic	0.46	1.00	0.55	0.43	0.51	0.38	0.52	0.27
Similarities	0.58	0.55	1.00	0.73	0.37	0.37	0.49	0.16
Vocabulary	0.63	0.43	0.73	1.00	0.33	0.43	0.41	0.09
Digit span	0.27	0.51	0.37	0.33	1.00	0.13	0.29	0.12
Picture completion	0.45	0.38	0.37	0.43	0.13	1.00	0.43	0.25
Block design	0.33	0.52	0.49	0.41	0.29	0.43	1.00	0.23
Coding	0.15	0.27	0.16	0.09	0.12	0.25	0.23	1.00
SD	4.93	4.10	5.20	6.54	2.72	5.35	9.36	10.44
Mean	15.17	15.00	11.83	21.67	12.17	17.83	18.67	45.83

