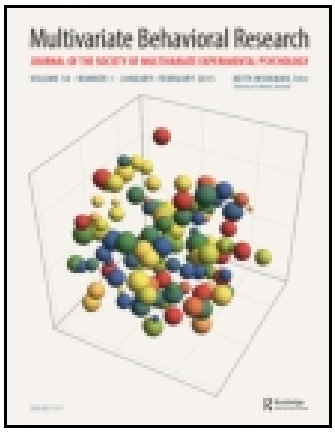

Multivariate Behavioral Research
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hmbr20

Using Score Equating and Measurement Invariance to Examine the Flynn Effect in the Wechsler Adult Intelligence Scale

Nicholas Benson[a], A. Alexander Beaujean[b] & Gordon E. Taub[c]
[a] Division of Counseling and Psychology in Education, The University of South Dakota
[b] Department of Educational Psychology, Baylor University
[c] Department of Child, Family, and Community Sciences, University of Central Florida
Published online: 28 Jul 2015.

Multivariate Behavioral Research, 50:398–415, 2015

ISSN: 0027-3171 print / 1532-7906 online
DOI: 10.1080/00273171.2015.1022642

Routledge
Taylor & Francis Group

Using Score Equating and Measurement Invariance to Examine the Flynn Effect in the Wechsler Adult Intelligence Scale

Nicholas Benson
Division of Counseling and Psychology in Education, The University of South Dakota
A. Alexander Beaujean
Department of Educational Psychology, Baylor University
Gordon E. Taub
Department of Child, Family, and Community Sciences, University of Central Florida

The Flynn effect (FE; i.e., increase in mean IQ scores over time) is commonly viewed as reflecting population shifts in intelligence, despite the fact that most FE studies have not investigated the assumption of score comparability. Consequently, the extent to which these mean differences in IQ scores reflect population shifts in cognitive abilities versus changes in the instruments used to measure these abilities is unclear. In this study, we used modern psychometric tools to examine the FE. First, we equated raw scores for each common subtest to be on the same scale across instruments. This enabled the combination of scores from all three instruments into one of 13 age groups before converting raw scores into Z scores. Second, using age-based standardized scores for standardization samples, we examined measurement invariance across the second (revised), third, and fourth editions of the Wechsler Adult Intelligence Scale. Results indicate that while scores were equivalent across the third and fourth editions, they were not equivalent across the second and third editions. Results suggest that there is some evidence for an increase in intelligence, but also call into question many published FE findings as presuming the instruments' scores are invariant when this assumption is not warranted.

There are well-documented secular changes in mean IQ scores in America (Flynn, 1984, 2012), as well as many other countries (Kanaya, Ceci, & Scullin, 2005; Flynn & Rossi-Casé, 2012). This phenomenon was observed as early as the 1930s, although the Flynn effect (FE) moniker was first coined by Herrnstein and Murray (1996) in recognition of James Flynn's scholarship in the area (Lynn, 2013). While the typical FE is between 3 to 5 IQ points per decade, the effect's magnitude and direction have shown considerable variation. As Williams (2013) pointed out, despite a proliferation of research pertaining to the FE the phenomenon remains enigmatic. Results from FE studies frequently conflict and few findings generalize across time and location. Moreover, it is unclear if FE gains are concentrated at the left tail of the bell curve (Colom, Lluis-Font, & Andres-Pueyo, 2005), concentrated at the right tail of the bell curve (Wai & Putallaz, 2011), or occur across the distribution (Flynn 1996, 2009a).

Its enigmatic nature notwithstanding, the FE has important implications for cognitive ability scholarship, the practice of psychology, and society in general (Kaufman & Weiss, 2010). As psychologists routinely administer intelligence tests, accurate norm-referenced comparisons are critical as these test scores are used to inform high-stakes decisions such as making or ruling out psychiatric diagnoses as well as eligibility decisions for special education, the Social Security Administration, and the death penalty (Flynn, 2006; Gresham & Reschly, 2011; Kanaya, Scullin, & Ceci, 2003). The influence of FE research on the practice of psychology is highlighted by the fact that test publishers note the FE as one

Correspondence should be sent to Nicholas Benson, Delzell 205D, 414 East Clark Street, Vermillion, SD 57069 E-mail: Nicholas.Benson@usd.edu

reason they obtain new nationally representative normative samples approximately every 10 years in an effort to control for norm obsolescence (Weiss, 2010).

Although the existence of the FE is widely accepted by professional psychologists, there is little agreement regarding causal mechanisms. Some have argued that the FE reflects an actual increase in cognitive abilities, due to either environmental changes such as nutrition (Lynn, 1998) or education (Blair, Gamson, Thorne, & Baker, 2005) or heterosis arising from changes in the ratio of heterozygous to homozygous genotypes (Mingroni, 2004). While at least some of the FE appears to reflect an actual increase in abilities (Shiu, Beaujean, Must, te Nijenhuis, & Must, 2013), many researchers have found that the FE is unrelated to general intelligence (*g*; e.g., Kane & Oakland, 2000; Must, Must, & Raudik, 2003; te Nijenhous & van der Flier, 2013; te Nijenhuis, van Vianen, & van der Flier, 2007), although some have found a rise in *g* (Shiu, Beaujean, & Wells, 2015).

At present, little research has examined relations between the FE and biological markers of brain function (e.g., diffusion coefficients, glucose metabolic rate, nerve conduction velocity; Williams, 2013), although head size reportedly has increased over time (Lynn, 2009). As head size correlates primarily with *g* rather than group factors (Jensen, 1998), and head size correlates highly with brain size, it could be argued that the FE is unrelated to brain growth given that previous research suggests its effects are not on *g*. The FE is not associated with improvements in inspection time (Nettelbeck & Wilson, 2004) and appears to be inversely related to changes in reaction times (Woodley, te Nijenhuis, & Murphy, 2013). Thus, there is no evidence to suggest that the FE can be accounted for by changes in brain efficiency.

Jensen (1998) proposed that the practical significance of the FE should be evaluated using tests of predictive bias. By this standard, the meaningfulness of gains in observed IQ scores is tenuous at best. As Jensen noted, if the FE reflected meaningful differences in intelligence then re-norming should change estimates of predictive validity. There is no evidence indicating that renorming changes estimates of predictive validity, while observed IQ scores may be increasing SAT scores are in fact declining (cf. Rodgers, 1998).

Evidence suggests that gains in observed IQ scores arise, at least in part, from issues other than genuine changes in the cognitive abilities that intelligence tests are purported to measure. Such issues include methodological and psychometric concerns (e.g., Beaujean & Osterlind, 2008) as well as substantial changes in the tests themselves (Kaufman, 2010). Any meaningful differences in intelligence that do exist are likely to be confounded by artifactual issues that inflate IQ scores (Williams, 2013). In fact, a recent meta-analysis suggests that variability between FE studies can be explained in aggregate by sampling error, unreliability, and restriction of range (te Nijenhuis & van der Flier, 2013).

The Appropriateness of Comparing Mean IQ Scores

Although the content of intelligence tests has changed over time (Boake, 2002), most FE research has assumed the standardized scores across instruments and editions are directly comparable (i.e., measurement invariance) and represent changes in cognitive ability. Then, without examining whether these are assumptions are warranted, they interpret any mean differences in IQ scores as representing mean differences in cognitive ability. This is unfortunate, as there are a variety of reasons for score differences across time.

Golembiewski, Billingsley, and Yeager (1976) delineated three different categories of score changes: alpha, beta, and gamma. *Alpha* change occurs when score differences correspond to an actual change in the construct the scores measure. For example, IQ scores increase because it reflects the increase in cognitive ability across time. *Beta* change occurs when score differences reflect a recalibration of the instrument's metric or scale. For example, IQ score differences are a result of anchoring the average score at different levels of cognitive ability across editions, not an actual change in cognitive ability itself. *Gamma* change represents a shift in the meaning/conceptualization of the measured construct. With gamma change, score differences are due to a different construct being measured. For example, the subtests that comprise a given IQ score may be so different between editions or instruments that they represent distinct, albeit related, cognitive abilities.

While there is evidence that the *g* factors measured across intelligence tests are highly correlated (Floyd, Reynolds, Farmer, and Kranzler, 2013), IQ scores are not necessarily exchangeable, especially the non-full-scale IQ (FSIQ) scores (Floyd, Bergeron, McCormack, Anderson, & Hargrove-Owens, 2005; Floyd, Clark, & Shadish, 2008). Thus, empirical support for the FE is based on comparisons of scores that assume alpha change, but the score differences could be due to gamma or beta change—meaning the equivalence of the scores is questionable and, subsequently, rendering the meaning of these findings indeterminate. Beaujean and Sheng (2014) liken the situation 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).

In order to ensure between-instrument score comparisons reflect differences in the level of the construct the instruments' scores intend to measure, it is first necessary to establish that the numerical values of the scores are comparable. One way to accomplish this is to administer the same edition of an instrument across multiple time-separated samples (e.g., Schaie, Willis, & Pennak, 2005). Another way to determine this comparability is to examine measurement invari-

ance (Millsap & Hartog, 1988). If measurement invariance is present, then it is appropriate to compare the observed scores across instruments because the probability of obtaining a given observed score is independent of the instrument used. Thus, individuals with the same level of the construct will, on average, produce the same observed score no matter what instrument is used (Meredith, 1993).

Previous FE research has examined measurement invariance using both item and test scores (Beaujean & Osterlind, 2008; Beaujean & Sheng, 2010, in press; Must, te Nijenhuis, Must, & van Vianen, 2009; Pietschnig, Tran, & Voracek, 2013; Shiu et al., 2013; Wicherts et al., 2004). They all converged in finding some level of *non*-invariance, which indicates that construct-irrelevant sources of variance were, at least partially, responsible for the FE. In other words, they found some evidence for beta change. Thus, reasons other than secular changes in intelligence appear to be partly responsible for the increase in test scores. As the construct-irrelevant sources' effects have likely differed between studies, the level of influence they exert on the FE is not exactly known. One way to better understand the influence of these construct-irrelevant sources of variance is to examine the changes in an instrument that has multiple editions published at different time points, such as the Wechsler Adult Intelligence Scale.

Changes in the Wechsler Adult Intelligence Scale Across Editions

The Wechsler Adult Intelligence Scale (WAIS) was first published in 1955 and has been revised three times (Wechsler, 1981, 1997, 2008). There has been some consistency between each edition as well as some noticeable changes. For example, the scoring structure of the first three editions included a Verbal IQ (VIQ), Performance IQ (PIQ), and FSIQ. While the fourth edition retained the FSIQ score, the VIQ-PIQ dichotomy was removed in favor of using four index scores: Verbal Comprehension, Perceptual Reasoning, Working Memory, and Processing Speed. In addition to changing the composite scores, there have been changes in some of the retained subtests (Kaufman, 2010) as well as the addition and subtraction of subtests that comprise the composite score (see Table 1). The third edition added three new subtests: Matrix Reasoning, Symbol Search, and Letter-Number Sequencing. The fourth edition removed two subtests (i.e., Object Assembly and Picture Arrangement) and added three new subtests (i.e., Cancellation, Figure Weights, and Visual Puzzles).

Another major change between WAIS editions is the demographics of the norming samples (see Table 2 as well as tables in Zhou, Zhu, & Weiss, 2010) so that the scores would be generalizable to the US population at the time of the norming. Nonetheless, a consequence of using nonequivalent norming groups is that the ability required to obtain a given standardized score on one WAIS edition is not necessarily the same level of ability required to get the same score on another WAIS edition.

What all the changes across WAIS editions indicate is that comparing index score means across editions and inferring that any changes are due to changes in cognitive ability is tenuous. First, it is difficult to separate changes in scores due to an increase in ability versus changes in scores due to using norming groups with different demographic characteristics. Second, the same composite scores across editions are comprised of different subtests and some of the subtests that remained across editions had substantial revisions. Consequently, changes in mean scores across editions could be due a variety of reasons, not just an increase in cognitive ability.

In response to the WAIS changes, some have advocated comparing subtest scores across editions to measure the FE (e.g., Flynn, 2009b). There are two major problems with this approach. First, the problems of using subtests as the unit of analysis are well known (Sinharay, Puhan, & Haberman, 2011), as typically they are more unreliable and have less information than composite scores (Sinharay, 2010). Second, such comparisons cannot differentiate alpha versus beta change. An alternative way to examine changes in aggregate-level scores that minimizes the influence of the different norming samples is to create standard scores that reflect relative rank within a grand sample consisting of participants from multiple WAIS normative samples. This requires combining the raw scores for each subtest across WAIS editions and then converting the combined raw scores into Z scores as this allows for comparisons based on relative rank within the grand sample. Moreover, these Z scores can then be combined to create either composite scores or be used as indicator variables for a latent variable model to test for invariance across the WAIS editions. There are two problems with this approach, however, but each problem has a solution.

The first problem is that the WAIS norming groups consist of individuals from a wide range of ages. Thus, combining the raw scores confounds ability differences with differences due to age. This can be solved by grouping the respondents into aged-based groups before converting the raw scores into Z scores.

The second problem is that the raw scores for each edition's subtests have unique metrics due to the items changing across WAIS editions. To demonstrate this, the parenthetical values in Table 1 show the maximum possible scores for the common WAIS subtests. While the maximum score for some subtests is relatively consistent across editions (e.g., Arithmetic, Information), others show more variation (e.g., Digit Span, Coding). Thus, in order to combine the raw subtest scores across WAIS editions, they first need to be equated (Linn, 1993).

Equating Subtest Scores

There are a variety of methods and procedures to link scores from different tests (Linn, 1993; Mislevy, 1992). The most

TABLE 1
Common Wechsler Subtests Across Editions

WAIS-R	WAIS-III	WAIS-IV
Block Design (51)	Block Design (68)	Block Design (66)
Coding (93)	Coding (133)	Coding (135)
Comprehension (32)	Comprehension (33)	Comprehension (36)
Digit Span (28)	Digit Span (30)	Digit Span (48)
Information (29)	Information (28)	Information (26)
	Letter-Number Sequencing (21)	Letter-Number Sequencing (30)
	Matrix Reasoning (26)	Matrix Reasoning (26)
Object Assembly (41)	Object Assembly (52)	
Picture Arrangement (20)	Picture Arrangement (22)	
Picture Completion (20)	Picture Completion (25)	Picture Completion (24)
Similarities (28)	Similarities (33)	Similarities (36)
	Symbol Search (60)	Symbol Search (60)
Vocabulary (70)	Vocabulary (66)	Vocabulary (57)

Note. WAIS: Wechsler Adult Intelligence Scale.
Numbers in parentheses denote the maximum possible score for that subtest. The Arithmetic subtest was included in all three editions, but we did not used in the current study because it is likely a better measure of academic achievement than intelligence (Parkin & Beaujean, 2012).

stringent form of linking is equating. Here, the different tests are thought to be interchangeable versions of the same test, so the goal is to make the scores exchangeable (i.e., using the same metric to measure the same construct). Consequently, to be able to equate two tests' scores, the tests must measure the same construct and must do so with an approximately equal degree of reliability (Kolen & Brennan, 2014). The number of items as well as the mean and variance of the tests' scores do not need to be the same, however, as successful equating adjusts for these differences. The resulting equated scores have the same meaning regardless of who took the test, when they took the test, or what version of the test they took.

When using observed scores (as opposed to items), there are three common ways to equate scores (Kolen & Brennan, 2014). *Mean equating* adjusts the mean of one test to be the same as the mean for another test, while *linear equating* adjusts both the mean and variability. A more general method for equating test scores is *equipercentile equating*. This method converts the scores from one test to those on another test by finding the observed scores that have the same percentile ranks on both tests. As test scores are technically discrete variables (as opposed to being continuous), equipercentile equating can produce scores with irregular distributions. This is especially problematic when the range of possible scores is small. In such cases, it is useful to use a smoothing function to eliminate any roughness and zero frequencies in the scores' distributions.

One way to incorporate smoothing into the equating process is to smooth the raw score distributions, sometimes referred to as *pre-smoothing*. One supported method for presmoothing is the polynomial log-linear method, which is shown in Equation (1).

$$\log [F(x)] = \delta_0 + \delta_1 x^1 + \delta_2 x^2 + \delta_3 x^3 + \cdots + \delta_M x^C \qquad (1)$$

Equation (1) is the log of the cumulative score density, $F(x)$, expressed as a polynomial of degree C (Holland & Thayer, 2000). The δ terms in Equation (1) are estimable parameters (Holland & Thayer, 1987). Using the logarithm allows Equation (1) to be additive instead of multiplicative.

Choosing C is the most important part of the polynomial log-linear method. One method of choosing C is to use a goodness-of-fit test. For a given score density, the estimation of the δ terms via maximum likelihood produces a fit statistic

TABLE 2
Descriptive Statistics for Standardization Samples

		Age	*Sex*	*Race/EthnicityPercentage*				
	n	M *(SD)*	*Percentage*					
WAIS Edition			Male	W	B	A	H	O
Second	1,800	39.5 (18.80)[a]	50	88.5[b]	NA	NA	NA	NA
Third	2,450	48.1 (23.6)	53.2	78.6	11.4	NA	7.4	2.7
Fourth	2,200	44.9 (22.9)	52.2	70.0	11.8	3.2	13.1	1.8

Note. W: White; B: Black; A: Asian, H: Hispanic; O: other; NA: information not reported.
[a] Estimated from grouped frequency tables.
[b] The only race/ethnicity categories reported were White/non-White.

that follows a χ^2 distribution with *C-1* degrees of freedom. A "statistically significant" value of the statistic suggests the model does not fit. Consequently, *C* is chosen by first fitting multiple models to the score data using increasing values for *C*, then selecting the model with the smallest *C* value that also adequately fits the distribution. Moses and Holland (2009) suggested that using Akaike's information criterion (AIC) to select the value of *C* produces more accurate estimation than using the χ^2 values. Here, multiple models are still fit, but selection is based on the model with the smallest AIC value.

A third way to select *C* is to use the value that produces the smallest standard error of equating (SEE). Whenever samples of examinees are used to estimate the equating relationship, random equating error is present. Conceptually, this error is the variance of equated scores over multiple replications of the equating procedure. The square root of the random error variance is the SEE (Lord, 1982).

Factor Models of Intelligence

In order to examine invariance of the WAIS scores, we first have to form a latent variable model of the subtest scores. There are competing views about the form of the latent variable model that should be used when examining cognitive ability data. Some advocate using a higher order factor model, with first-order factors directly influencing the test scores and the second-order factor only directly influencing the first-order factors (Reynolds & Keith, 2013; Weiss, Keith, Zhu, & Chen, 2013). In keeping with Carroll's (1993) terminology, we call the first-order factors Stratum II factors and the second-order factor *g*. In the higher order model, *g* only has an *indirect* relationship with the test scores. Others advocate using a bi-factor model, which posits there are two systematic and direct influences on the test scores (Gignac, 2008; Reise, 2012). The first influence is *g*. The second influence is the set of domain-specific Stratum II factors, each of which influence only a portion of the tests. These Stratum II factors, also known as group factors, represent variance shared by subsets of tests with similar task demands. Unlike the higher order model, the bi-factor model specifies that *g* and the Stratum II factors are uncorrelated with each other.

We use a bi-factor model for the current study for multiple reasons. First, Carroll's (1993) three-stratum theory of cognitive ability is generally considered the most empirically supported model of cognitive ability currently available (Jensen, 2004). Carroll (1997) argued that a bi-factor specification was the best way to represent his three-stratum model: "[It] would be desirable to show also that a general factor so identified constitutes a true ability, independent of lower order factors, rather than being merely a measure of associations among those lower order factors. . ." (p. 144). Moreover, previous research that has utilized the bi-factor model indicates it fits data from different versions of the WAIS relatively well, and often better than alternative models (e.g., Gignac, 2005, 2006; Gignac & Watkins, 2013).

Second, bi-factor models have an interpretive advantage over higher order models. The bi-factor model specifies first-order factors that are independent of *g* instead of being influenced by both *g* and non-*g* abilities. Thus, Murray and Johnson (2013) concluded, "If 'pure' measures of specific abilities are required then bi-factor model factor scores should be preferred to those from a higher order model"(p. 420). Moreover, bi-factor models do not require that *g* be interpreted on the basis of first-order factors, which has been likened to "interpreting shadows of the shadows of mountains rather than the mountains themselves" (McClain, 1996, p. 233).

Third, higher order models disallow *g* to have a direct relationship to the individual test scores. Instead, the *g*-test score relationship is mediated by the Stratum II factors. Moreover, these mediated relationships have proportionality constraints (Brunner, 2008; Schmiedek & Li, 2004). That is, for a given set of tests influenced by the same Stratum II factor, the ratio of the test scores' variance due to the Stratum II factor to the variance attributable to *g* are constrained to be the same (for a graphical explanation of these proportionality constraints, see Beaujean, Parkin, & Parker, 2014). While these proportionality constraints make the higher order model more parsimonious than the bi-factor model, they also limit the higher order model's ability to represent the direct relation between *g* and individual test scores.

Fourth, bi-factors models have an advantage over higher order models when examining invariance (Chen, West, & Sousa, 2006). Because the bi-factor model specifies the first-order factors as being independent of *g*, lack of invariance in a first-order factor does not influence invariance in *g* and vice versa. In addition, the bi-factor model allows for a direct comparison of latent mean differences between groups on Stratum II factors over and above *g*. Thus, any differences in a Stratum II factor across group are due to changes independent of *g*. Consequently, if there is measurement invariance across the WAIS editions, a bi-factor model allows for a more direct examination of whether the FE involves *g*, Stratum II factors, or both.

Current Study

The purpose of this study is to examine the FE in the revised (second), third, and fourth editions of the WAIS using sound psychometric analysis.[1] The WAIS is one of the most popular instruments used to measure cognitive ability in adults (Camara, Nathan, & Puente, 2000), and has been utilized in much FE scholarship, especially Flynn's (2012) own work in the United States.

If scores derived from different WAIS editions are invariant, then any index score difference can be interpreted as representing meaningful differences in intelligence. If the scores lack a sufficient level of invariance, however, it would

[1]Scores from the original WAIS were not included because equating data were not available.

be wrong to conclude that observed mean differences in the FSIQ, or any other index score, only reflect differences in intelligence. Instead, non-invariance would suggest that the secular changes in scores, at least in part, reflect a difference in the tests themselves (i.e., beta change in addition to, or in lieu of, alpha change). Based on previous invariance studies of the FE, we expect to find some level of non-invariance across all the editions, although we cannot hypothesize the magnitude and influence of this invariance.

METHOD

Participants

This study used participants from the Wechsler Adult Intelligence Scale's revised (WAIS-R; $n = 1{,}800$), third (WAIS-III; $n = 2{,}450$), and fourth (WAIS-IV; $n = 2{,}200$) editions' standardization samples. The tests' publisher provided all the data. Information regarding the participants' age, sex, and race/ethnicity is presented in Table 2.

There were a few notable differences in the inclusion criteria for the standardization samples. First, during the WAIS-R norming process only two racial groups were sampled (White, Non-White), while the WAIS-III sample consists of four racial/ethnic groups (Black, White, Hispanic, Other) and the WAIS-IV sample consists of five racial/ethnic groups (Black, White, Hispanic, Other, Asian). Second, medical and psychiatric exclusionary criteria were used when norming the WAIS-III and WAIS-IV. Third, for the WAIS-R participants up to age 75 years were sampled while subsequent editions sampled up to age 90 years.

Wechsler Subtests

Wechsler subtests used in the current study are shown in Table 1. Most of the subtests were used in all three editions of the WAIS, although there were some exceptions. The WAIS-R did not include the Matrix Reasoning and Symbol Search subtests, while the WAIS-IV did not include the Object Assembly and Picture Arrangement subtests. Although the Arithmetic subtest was included in all three editions, we did not use it in the data analysis because it is likely a better measure of academic achievement than intelligence (e.g., Parkin & Beaujean, 2012).

Data Analysis

There were two parts to this study's data analysis. The first part involved equating the WAIS subtest scores, while the second part involved examining invariance of the equated scores.

Subtest Score Equating

As the datasets contained raw scores, each WAIS-R and WAIS-IV subtest was equated to the corresponding subtest raw score on the WAIS-III. Participants in the equating studies were administered two editions of the WAIS, either the WAIS-R and WAIS-III ($n = 192$) or the WAIS-III and WAIS-IV ($n = 284$), and all participants were originally part of a standardization sample. All samples were collected to represent the percentages of national demographics (i.e., age, sex, ethnicity, and 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.

One respondent, each, was missing data on the following subtests: WAIS-III and WAIS-IV Arithmetic, WAIS-III and WAIS-IV Symbol Search, and WAIS-R Picture Completion. Four respondents were missing data on the WAIS-III Picture Arrangement subtest. Respondents missing data for a given subtest were excluded from the equating of that subtest, but were included in the equating of all other subtests.

We equated each subtest's raw scores using equipercentile methods with pre-smoothing using a polynomial log-linear model [(see Equation (1)] with degrees ranging from $C = 1$–7.[2] For each model in each subtest, we examined the χ^2, AIC, and SEE values.[3] We then selected the optimal value of C for each subtest based on having relatively low SEE values, fitting the data better than other models, and producing sensible equated scores (i.e., minimum and maximum values of equated scores being close to the possible data range).

After equating the subtests' raw scores, we combined the three samples. Because of the different age ranges in the different editions' norming samples, we created age-based scores. Specifically, we placed all participants into one of 13 age groups (16–17, 18–19, 20–24, 25–39, 30–34, 35–44, 45–54, 55–64, 65–69, 70–74, 75–79, 80–84, and 85–90 years) and converted the raw scores into Z scores within each age group. We used these Z scores for all subsequent analyses.

Invariance

The second part of the study's analysis involved examining invariance of the WAIS across editions. Before investigating invariance, however, we determined the factor structure of each edition's subtests. Subsequently, we examined invariance via multi-group latent variable models, using WAIS edition as the grouping variable. For this part of the study, we used all participants from the WAIS-R ($n = 1{,}800$), WAIS-III ($n = 2{,}450$), and WAIS-IV ($n = 2{,}200$) standardization samples.

[2]Details of the equating processes are provided as supplemental materials.

[3]The SEE were estimated using bootstrap methods with 1,000 replacement samples. For details about this process, see Kolen and Brennan (2014).

TABLE 3
Levels of Measurement Invariance

Model	*Title*	*Description*
1	Configural	The editions' factor models are the same. No parameter constraints imposed.
2	Weak	1 + constrain all factor loadings to be the same between editions
3	Strong	2 + constrain all intercepts to be the same between editions
4	Strict	3 + constrain error/residual variances to be the same between editions
5		3 or 4 + constrain the latent variances to be the same between editions
6		3, 4, or 5 + constrain the latent means to be the same between editions

To assess invariance, we examined a series of increasingly restrictive models (see Table 3). First, we examined configural invariance by determining if the different editions have the same number of factors and factor loadings pattern. Next, we examined weak invariance by constraining factor loadings to be equal across editions. If such a model holds, it implies that the latent variable's units/scale is the same across editions. In the third step, we examined strong invariance by constraining the subtests' intercepts to be equal across editions. Invariant intercepts imply that any between-edition mean differences in subtest scores are only due to between-edition differences in the latent variables. Fourth, we examined strict invariance by constraining the subtests' residual/error variances to be equal across editions. Although examining strict invariance is not absolutely necessary (Little & Slegers, 2005), if there is strict invariance as well as invariance in the latent variables' variances, then this indicates the constructs were measured with equal reliability across editions. If either the strict or strong invariance model did fit the data as well as the less restrictive models, then we considered the WAIS editions to exhibit measurement invariance.

For a model exhibiting measurement invariance, we then investigated invariance of the latent variables. As these steps are not hierarchical, failure to find one type of invariance does not preclude examining another. First, we constrained the latent variances to be equal across editions. If the latent and residual variances are both invariant across editions, then the measured constructs' reliabilities are equivalent. Second, we constrained the latent means to be equal across editions, which, if true, would indicate there was no change in the constructs' mean across editions.

Assessing model fit

Although the typical measure of model fit is the χ^2 statistic, it is very sensitive to sample size (West, Taylor, & Wu, 2012). Since our sample sizes were large, we used the following alternative fit measures and criteria to determine acceptable model fit: comparative fit index (CFI; > 0.95), McDonald's noncentrality index (Mc; > 0.90), and root mean square error of approximation (RMSEA; < 0.08). In addition, we used the AIC, which is best used to compare competing models, with lower values indicating better fit.

Traditionally, the difference in the χ^2 values (i.e., likelihood ratio test) between the increasingly restrictive invariance models has been used to determine model fit because these models are nested within each other. As with single model assessment, the difference in the χ^2 values is also sensitive to sample size (Cheung & Rensvold, 2002). As an alternative, Meade, Johnson, and Braddy (2008) suggest using differences in the CFI and Mc indexes, with differences

TABLE 4
Results from Equating Wechsler Adult Intelligence Scale Subtests

	WAIS-R (n = 192)							WAIS-IV (n = 284)				
	Non-Equated			Equated		WAIS-III		Non-Equated			Equated	
Subtest	C	M (SD)	Range	M (SD)	Range	M (SD)	Range	C	M (SD)	Range	M (SD)	Range
BD	3	24.04 (12.84)	0–59	32.6 (15.13)	1.56–68.28	36.73 (13.14)	2–68	2	40.24 (13.57)	6–66	39.31 (13.41)	6.08–65.54
CD	2	45.93 (21.66)	1–93	58.75 (27.87)	0.8–118.17	66.78 (21.82)	2–133	3	65.08 (19.94)	4–135	69.32 (23.15)	3.39–135.49
CO	3	18.86 (6.58)	1–32	18.62 (6.5)	0.53–31.33	20.24 (5.66)	2–33	4	23.37 (6.06)	5–36	21 (5.85)	4.74–33.45
DS	4	14.07 (4.47)	0–28	16.04 (4.68)	–0.45–30.42	16.67 (4.34)	1–30	3	27.38 (6.1)	9–48	17.31 (4.33)	6.52–30.49
IN	4	17.02 (6.32)	1–29	15.27 (5.94)	2.9–27.16	16.04 (5.53)	0–28	6	13.9 (5.25)	0–26	17.01 (5.39)	0.51–28.39
MR	—	—				13.73 (5.8)	0–26	4	16.25 (5.48)	1–26	15.51 (5.71)	0.59–25.66
OA	3	26.21 (8.56)	1–41	25.76 (11.38)	4.2–49.43	29.37 (10.57)	1–52	—	—	—	—	—
PA	4	9.73 (5.89)	0–20	10.85 (5.9)	0.47–21.61	12.3 (5.35)	0–22	—	—	—	—	—
PC	3	12.63 (5.02)	0–20	16.43 (6.66)	0.04–25.06	18.58 (4.23)	1–25	2	12.29 (4.41)	1–24	19.78 (4.41)	5.83–25.48
SI	4	15.68 (6.95)	0–28	18.62 (7.08)	1.94–31.81	21.61 (5.8)	0–33	4	24.38 (5.66)	1–36	23.26 (5.65)	–0.5–33.49
SS	—	—				29.09 (10.79)	0–60	2	30.36 (9.67)	0–60	30.35 (11.32)	–0.33–60.35
VC	4	41.5 (16.18)	3–70	36.43 (14.47)	0.8–64.87	41.14 (12.48)	3–66	4	35.84 (10.86)	6–57	44.45 (12.74)	3.67–65.92

Note. C*:* Degree of pre-smoothing polynomial; BD: Block Design; CD: Coding; CO: Comprehension; DS: Digit Span; IN: Information; MR: Matrix Reasoning; OA: Object Assembly; PA: Picture Arrangement; PC: Picture Completion; SI: Similarities; SS: Symbol Search; VC: Vocabulary; —: subtest not included in WAIS edition.

All values are in raw score units. All subtests equated to be on WAIS-III metric.

TABLE 5
Descriptive Statistics for Wechsler Adult Intelligence Scale Subtests by Edition

	WAIS-R (n = 1,800)		WAIS-III (n = 2,450)		WAIS-IV(n = 2,200)	
Subtest	M (SD)	Range	M (SD)	Range	M (SD)	Range
BD	92.4 (14.5)	47.3–131.6	100.1 (13.7)	47.8–142.2	104.3 (14.3)	52.8–165.1
CD	95.5 (15.7)	46.0–124.8	99.7 (14.3)	49.1–131.6	104 (14.7)	50.7–154.5
CO	95.5 (15.7)	46.0–124.8	99.7 (14.3)	49.1–131.6	102.7 (14.7)	56.4–135.4
DS	95.8 (15.0)	43.2–141.7	99.8 (14.6)	47.1–161.2	102.7 (14.6)	62.3–155.7
IN	97.1 (15.3)	58.6–129.6	99.2 (15.0)	53.1–135.3	102.4 (14.6)	55.7–140.8
MR	—	—	96.7 (14.8)	49.3–156.7	102.6 (14.5)	53.9–142.0
OA	94.4 (13.9)	55.4–125.7	102.9 (14.4)	55.0–161.6	—	—
PA	94.8 (14.5)	49.5–132.4	102.8 (14.2)	42.9–148	—	—
PC	92.2 (17.8)	10.3–120.2	99.9 (12.6)	28.1–129.8	104.9 (13.2)	46.4–131.4
SI	92.3 (15.8)	41.2–128.8	100.1 (13.9)	44.3–130.9	104.6 (13.6)	43.0–132.4
SS	—	—	98.3 (14.1)	49.7–180.0	101.3 (15.5)	47.8–166.1
VC	94.6 (15.3)	47.4–125.8	99.5 (14.3)	51.6–139.7	103.9 (14.5)	51.7–133.0

Note. *Z* scores were transformed into IQ scale scores (M: 100, *SD*: 15) to aide interpretability.
BD: Block Design; CD: Coding; CO: Comprehension; DS: Digit Span; IN: Information; MR: Matrix Reasoning; OA: Object Assembly; PA: Picture Arrangement; PC: Picture Completion; SI: Similarities; SS: Symbol Search; VC: Vocabulary; —: subtest not included in WAIS edition.

in CFI values of .002 and differences in Mc values between 0.008–.009 being useful cutoff points.

Data Analysis Software

All analyses were done using the **R** statistical program. We used the *equate* (Albano, 2011) package to perform the equating and the *lavaan* (Rosseel, 2012) package to fit the latent variable models (Beaujean, 2014).

RESULTS

Equating

The results from the equating are given in Table 4. The presmoothing polynomial degree (*C*) was 4 or lower for all subtests except Information on the WAIS-IV where the degree was 6. Further inspection of this subtest showed multiple peaks and troughs in the raw scores, indicating that the degree is likely not too large. In addition, Table 4 contains the raw score means and standard deviations for equated and non-equated scores. In general, the moments for the equated scores are closer to the WAIS-III values than the moments for the non-equated scores, although this is better for the WAIS-IV subtests than the WAIS-R subtests. Thus, it appears that the equating worked as expected. Interestingly, after equating the scores the values for the subsequent editions are higher than the scores from the previous editions across all subtests. This indicates that when the subtest scores are aggregated there will be a FE, although without examining invariance not much interpretive weight should be placed on these scores. Table 5 contains descriptive statistics for each WAIS edition's equated scores after applying the within-age group standardization.

Data Screening

Missing Data

Missing data were minimal, as 99.78% of the respondents from the standardization samples had no missing data. The others were missing responses on one to three subtests. Instead of discarding these observations, we used full-information maximum likelihood estimation (FIML; Enders & Bandalos, 2001), which incorporates the information available from all the participants.

Normality Assumptions

Data screening revealed no atypical skew or kurtosis in the subtests. Multivariate normality, however, was not supported based on multivariate kurtosis estimates and quantile–quantile plots. Consequently, we used a robust estimator (MLR; Asparouhov & Muthén, 2005) for the analyses, which has been shown to work well with FIML estimation (Enders, 2001).

TABLE 6
Fit of Baseline Models in Standardization Samples

Model	Description	CFI	Mc	RMSEA	χ^2 (*df*)
B1	WAIS-R Bi-factor	.985	.964	.059	222.250(27)
B2	WAIS-III Bi-factor 1[a]	.988	.967	.050	225.776(27)
B3	WAIS-III Bi-factor 2[b]	.987	.963	.051	247.176(29)
B4	WAIS-IV Bi-factor	.989	.970	.046	178.722(29)

Note. CFI: comparative fit index; Mc: McDonald's noncentrality index; RMSEA: root-mean square error of approximation.
[a] Used in invariance analysis with WAIS-R, so did not include the Matrix Reasoning and Symbol Search subtests.
[b] Used in invariance analysis with WAIS-IV, so did not include the Object Assembly and Picture Arrangement subtests.

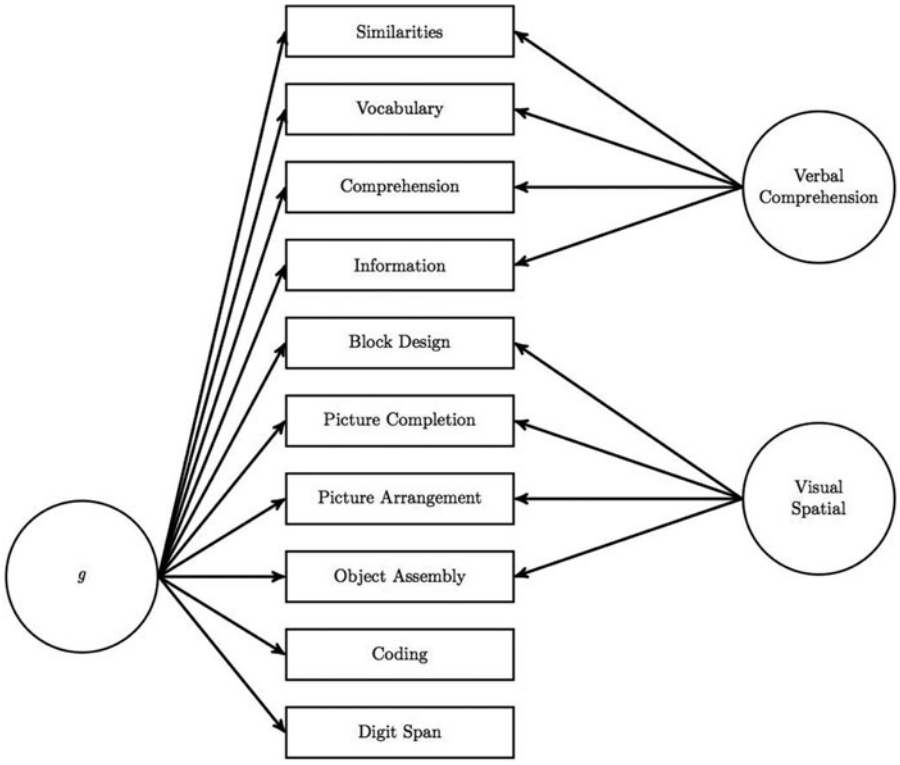

FIGURE 1 Bi-Factor Model for the Wechsler Adult Intelligence Scale subtests across the revised and third editions of the Wechsler Adult Intelligence Scale.

Testing Invariance

Revised and Third Editions

First, we determined the factor model to use for the data. Since the WAIS-R did not include the Matrix Reasoning and Symbol Search subtests, we did not include them as indicator variables for the WAIS-III either. We found the bi-factor model fit the data relatively well in both editions (see Models B1 and B2 in Table 6). For these two editions, the general factor represents general intelligence (*g*; Spearman, 1904) and the two group factors represent Verbal Comprehension and Visual Spatial Processing (see Figure 1). To identify the models, we initially constrained one loading for each factor in each edition to be one. For *g*, Verbal Comprehension, and Visual Spatial Processing, respectively, the loadings we constrained were for the Similarities, Information, and Block Design subtests. All the other parameters were freely estimated.

Next, we examined invariance between the two editions. The results are given in Table 7. The configural invariance model fit the data relatively well (Model 1), but constraining factor loadings to be equal (Model 2) caused a noticeable degradation in model fit. When we examined what factor loadings were the most discrepant in Model 1, we found the Picture Completion subtest's loading on *g* had the largest between-edition difference, so this equality constraint was re-

TABLE 7
Invariance Results for the WAIS-R and WAIS-III in Standardization Samples

Model	CFI	ΔCFI	Mc	ΔMc	χ^2 (df)	AIC	RMSEA
Configural	.987	—	.961	—	448.346 (54)	109865	.054
Weak	.980	.007[c]	.942	.019[c]	663.253 (69)	110049	.059
Partial Weak[a]	.984	.003[c]	.954	.007[c]	536.622 (68)	109924	.053
Partial Strong[b]	.977	.007[d]	.935	.019[d]	745.097 (74)	110121	.06

Note. CFI: comparative fit index; Mc: McDonald's noncentrality index; AIC: Akaike's information criterion; RMSEA: root mean square error of approximation.

[a] The *g*-loading for Picture Completion was freed between groups.

[b] The *g*-loading and intercept for Picture Completion were freed between groups.

[c] Compared with Model 1.

[d] Compared with Model 3.

TABLE 8
Standardized Parameter Estimates for WAIS-R and WAIS-III Subtests Under Partial Weak Invariance in Standardization Samples

Subtest	*g* Loading	Loading on Group Factor	WAIS-R Intercept	WAIS-R SE of Intercept	WAIS-III Intercept	WAIS-III SE of Intercept	WAIS-R Residual Variance	WAIS-III Residual Variance
Block Design	.70	.36	–.42	0.02	.06	.02	.33	.39
Coding	.62	—	–.46	.02	.11	.02	.60	.61
Comprehension	.72	.43	–.25	.02	.02	.02	.31	.31
Digit Span	.55	—	–.23	.02	.01	.02	.64	.70
Information	.67	.50	–.16	.02	–.02	.02	.25	.30
Object Assembly	.58	.62	–.3	.02	.21	.02	.21	.28
Picture Arrangement	.70	.11	–.28	.02	.20	.02	.47	.51
Picture Completion	.78[a],.69[b]	.22	–.44	.03	.04	.02	.37	.47
Similarities	.77	.36	–.45	.02	.05	.02	.32	.28
Vocabulary	.73	.55	–.32	.02	.00	.02	.13	.16

Note. Subtests are all on *Z* score scale.
g: general intelligence; *SE:* Standard Error.
[a] Estimate for the WAIS-R.
[b] Estimate for the WAIS-III.

leased. This partial weak invariance model (Model 3) showed minimal degradation in fit from Model 1. Estimates of factor loadings for the partial weak invariance model are presented in Table 8.

Last, we constrained the intercepts to be equal across editions for all subtests except Picture Completion (Model 4). These constraints caused a substantial degradation in model fit, so we examined what subtests' intercepts were the most discrepant using the results from Model 3. The results, shown in Table 8, indicate that all the subtests show substantial difference. Consequently, it appears that between-edition differences in the latent constructs do not account for all the

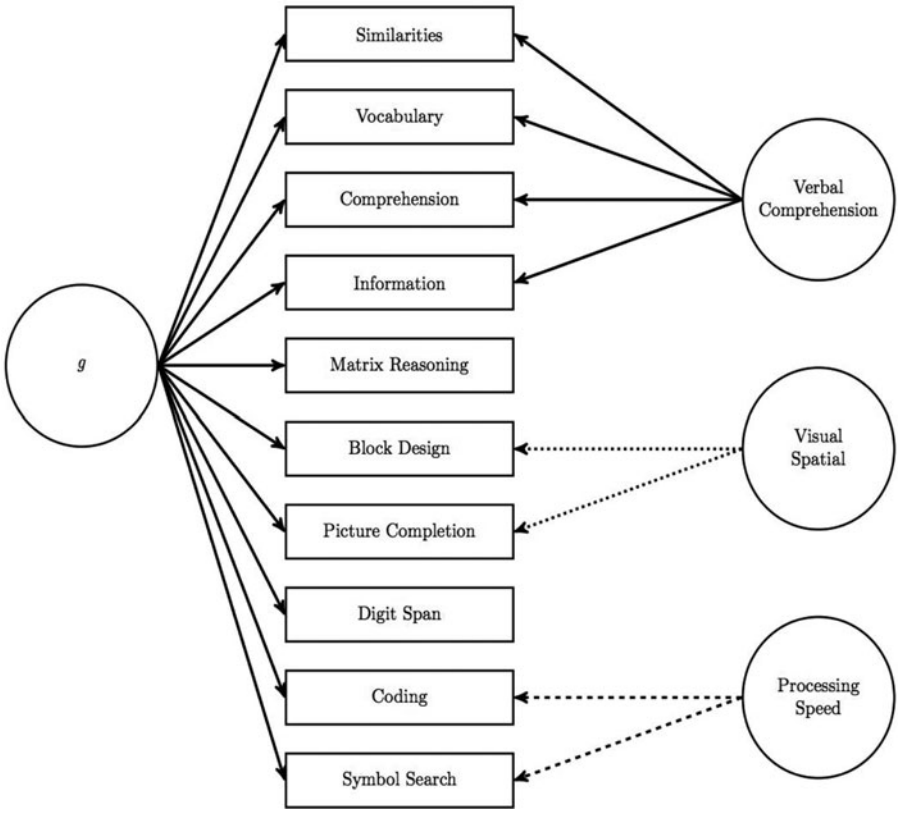

FIGURE 2 Bi-Factor Model for the third and fourth editions of the Wechsler Adult Intelligence Scale subtests. Dashed lines indicate that loadings for the Processing Speed factor are constrained to be equal within editions. Dotted lines indicate that loadings for the Visual Spatial factor are constrained to be equal within editions.

TABLE 9
Invariance Results for the WAIS-III and WAIS-IV in Standardization Samples

Model	CFI	ΔCFI	Mc	ΔMc	χ^2 (df)	AIC	RMSEA
Configural	.988	—	.967	—	425.898 (58)	118091	.048
Weak	.986	.002[a]	.961	.005[a]	498.027 (70)	118139	.048
Strong	.984	.002[a]	.955	.006[a]	568.730 (76)	118198	.049
Strict	.981	.002[a]	.949	.006[a]	650.945 (86)	118260	.049
Model 4 +latent variances constrained	.979	.002[a]	.943	.006[a]	725.828 (90)	118327	.051
Model 5 + all latent means constrained	.973	.006[a]	.928	.015[a]	898.382 (94)	118492	.056
Model 6 without constraining g's mean	.978	.001[b]	.942	.001[b]	742.705 (93)	118337	.051

Note. CFI: comparative fit index; Mc: McDonald's noncentrality index; AIC: Akaike's information criterion; RMSEA: root mean square error of approximation; *g*: general intelligence.

[a] Compared to previous model.

[b] Compared to Model 5.

For identification purposes, in Models 2 through 4 the latent variance for Processing Speed and Visual Spatial Processing were constrained to 1.0 for the WAIS-III, but freely estimated in the WAIS-IV; the variance of *g* and Verbal Comprehension were freely estimated in both groups.

differences in subtest scores. That is, WAIS exhibited substantial change between the revised and third editions, in addition to any possible changes in the two editions' standardization samples. Thus, beta change is responsible for at least part of the score differences between the two editions.

Third and Fourth Editions

Since the WAIS-IV did not include the Object Assembly and Picture Arrangement subtests, we did not include them as indicator variables for the WAIS-III, either. First, we found a bi-factor model fit the data relatively well in both editions (see Models B3 and B4 in in Table 6). For these two editions, in addition to *g*, there were three group factors: Verbal Comprehension, Visual Spatial Processing, and Processing Speed (see Figure 2). Since there were only two subtests for the Visual Spatial Processing and Processing Speed factors, we constrained their factor loadings to be equal within an edition.

The results from the invariance assessment are given in Table 9, and indicate that these two editions exhibited measurement invariance. Specifically, tests of configural, weak, strong, and strict invariance (Models 1, 2, 3 and 4, respectively) were supported by relatively small ΔCFI and ΔMc values. Parameter estimates for the final model are presented in Table 10.

Since there were only two subtests for the Visual Spatial Processing and Processing Speed factors, we identified the models differently than with the WAIS-R-WAIS-III comparison. Specifically, for Model 1 we initially constrained all the latent variances to be one and constrained the factor loadings for Visual Spatial Processing and Processing Speed to be equal within an edition. For Models 2–4, we constrained the factor loadings for Visual Spatial Processing and Processing Speed, respectively, to be equal within an edition and across editions, but constrained their latent variances to be one only for the WAIS-III edition. To identify the *g* and Verbal Comprehension factors, respectively, we constrained the loadings

TABLE 10
Parameter Estimates for the WAIS-III and WAIS-IV Final Invariance Model in Standardization Samples

Subtest	Standardized Loading on *g*	Standardized Loading on Group Factor Loading	Standardized Intercept	Standardized Residual Variance
Block Design	.73	–.25	.05	.41
Coding	.58	.55	.09	.36
Comprehension	.67	.50	–.02	.29
Digit Span	.60	—	–.01	.65
Information	.65	.51	–.05	.32
Matrix Reasoning	.74	—	–.13	.45
Picture Completion	.66	–.27	.09	.49
Symbol Search	.61	.51	–.11	.37
Similarities	.72	.44	.06	.29
Vocabulary	.71	.57	.01	.17

Note. Subtests are all on *Z* score scale. The latent mean of *g* is zero for the WAIS-III and 0.37 for the WAIS-IV. Latent means for G*c*, G*s*, and G*v* all equal zero. Latent variances all equal 1. *g*: general intelligence; *Gc*: verbal comprehension; *Gs*: processing speed; *Gv*: visual spatial processing.

for Similarities and Vocabulary to be one and estimated the latent variances.

Subsequently, we constrained the latent variances (Model 5) and latent means (Model 6) to be equal across editions. The latent variances did not appreciably differ across editions, indicating that the constructs the WAIS-III and WAIS-IV subtests measure are measured with equal precision across editions. While the latent means of the domain-specific factors showed no between-edition difference, Model 6's results indicated we needed to release *g*'s mean across editions (Model 7). The between-edition mean difference in *g* was 0.373 standard deviation units (i.e., *d* effect size) higher for the WAIS-IV's sample than the WAIS-III's sample. Thus, it appears that when comparing the WAIS-III and WAIS-IV samples on the equated scores, the score changes mostly reflect alpha change—that is, the score differences reflect changes in *g*, not instrumental changes.

DISCUSSION

The purpose of this study was to examine the Flynn effect (FE) in revised (second), third, and fourth editions of the Wechsler Adult Intelligence Scales (WAIS) using sound psychometric analysis of the editions' standardization data. We utilized data from the WAIS-R-to-WAIS-III and WAIS-III-to-WAIS-IV linking studies provided by the publisher to equate the raw scores for each subtest in the WAIS-R and WAIS-IV, separately, to be on the same scale as the WAIS-III. We then investigated invariance between the WAIS-R and WAIS-III and then between the WAIS-III and WAIS-IV via multi-group latent variable models. While only weak invariance is tenable when comparing the WAIS-R and WAIS-III, results indicate that measurement invariance is tenable when comparing the WAIS-III and WAIS-IV.

Even though score comparability across instruments depends on a minimum level of invariance, FE studies do not typically examine this assumption. Thus, any difference they report in manifest scores from these instruments (e.g., FSIQ) could just as easily be due to changes in the instrument as due to changes in the examinees (i.e., beta or gamma change vs. alpha change; Golembiewski et al., 1976). In contrast to previous studies, our use of score equating placed subtests on equivalent metrics across editions, which then allowed them to be combined to form a single reference group. After combining the scores, we converted the raw scores into *Z* scores using age-based reference groups. This approach yielded a distribution of scores based on relative rank within a grand sample comprised of participants from all three normative samples.

When comparing the WAIS-R and WAIS-III, results revealed that controlling for differences in the latent variables did not account for differences in the subtests' intercepts. Failure to establish strong measurement invariance indicates that in creating the WAIS-III, the test authors changed the WAIS-R subtest's items in such a way that differences in performance on them is partially due to one of more additional latent variables not included in our factor model (Steinmetz, 2013). As these differences extended across all the intercepts (see Table 7), one of the unmeasured variables could be related to administration/administrator differences (McDermott, Watkins, & Rhoad, 2014; for additional possible causes, see Steinmetz, pp. 3–4). Another alternative is that participants' test-taking strategies changed in the timespan between when the WAIS-R and WAIS-III were normed possibly as a response to the proliferation of standardized testing for high-stakes decisions proliferated during the 1980s and 1990s. For example, as scoring rules for many tests changed to reward speediness of responding while simultaneously reducing penalties for guessing, this could have led to respondents using different test-taking patterns (Must & Must, 2013). Indeed, as shown in Table 7 the largest intercept differences between WAIS-R and WAIS-III versions of subtests were observed for timed subtests (i.e., Coding and Object Assembly) while the smallest intercept differences were found for untimed subtests (i.e., Information and Digit Span). In any case, because scores from the WAIS-R and WAIS-III are not on the same metric, any reported between-edition mean differences (e.g., Flynn, 1998, 2009b) do not necessarily indicate changes in the constructs the scores represent. Thus, not only are these score comparisons not very informative concerning the FE research, but they should not be used in clinical practice, either.

Unlike the WAIS-R and WAIS-III comparison, results from the WAIS-III and WAIS-IV comparison indicate that measurement invariance is tenable. Thus, between-edition score comparisons, at least using scores derived from the current study's subtests, represent differences in the construct the scores represent. Moreover, as we found that *g* was the only latent variable that showed mean changes over time (0.37 *SD* increase from the WAIS-III to WAIS-IV sample), any scores differences between the two editions can be interpreted as arising largely from differences in *g*. More specifically, if the FSIQ is estimated as from the summed 10 subtests shared by the WAIS-III and WAIS-IV, then there is an increase of 4.37 IQ points when comparing the mean for the WAIS-III standardization sample ($M = 97.98$) to the mean for the WAIS-IV standardization sample ($M = 102.36$).[4] Alternatively, using the latent mean differences in *g*, the 0.37 *SD* translates to a 5.60 IQ point difference.

Comparison to Previous Flynn Effect Research

The current study is the first study we are aware of that has equated raw scores across editions to create a single reference group. We believe that our equating strategy is

[4]To create the FSIQ score, we summed the 11 common subtests and then formed *Z* scores within each of the 13 age groups. Subsequently, we multiplied each *Z* score by 15 and added 100.

directly in line with Rodgers' (1998) proposals for better FE studies. Relative to methods used in most FE research, the approach used in the present study allows for a more direct test of whether the FE arises from genuine secular changes in intelligence or simply reflects changes in the tests used to measure intelligence.

Zhou et al. (2010) previously used score equating to study the FE in Wechsler scales, but their study and use of score equating was much different than ours. First, they only examined changes in the Performance Index (PIQ) score. They found an average score increase of approximately 0.30 PIQ units per year from the WAIS-R to the WAIS-IV, but this increase was moderated by the Verbal Index (VIQ) score. Specifically, the majority of the PIQ score increase from the WAIS-R to WAIS-III was concentrated in individuals with VIQ scores in the middle and lower range, but the change in PIQ scores from the WAIS-III to WAIS-IV had a higher concentration in individuals with VIQ scores in the upper range.

Second, Zhou et al. (2010) did not examine invariance in the equated PIQ scores, so it is difficult to know if the patterns of change they found are due to an increase in the abilities the PIQ measures (i.e., Fluid Reasoning, Visual-Spatial Ability) between editions or a change in structure of the PIQ score itself. Third, Zhou et al. used percentiles from equipercentile equating as a method to examine changes in the FE. As expected, after equating they found that for a given percentile WAIS-III scores were always higher than the WAIS-IV scores. Unexpectedly, they found that the amount of difference was inconsistent across the PIQ score range as higher scores tended to show larger differences than lower scores. Likewise, WAIS-R scores were higher than the WAIS-III at differing magnitudes, except at very high percentiles where the pattern reversed and WAIS-III scores where higher than WAIS-R scores. While this somewhat maps onto our finding that WAIS-R and WAIS-III scores should not be compared, this confirmation should be interpreted with a caveat. Unlike our study, they did not report using any smoothing, which could be why their equating produced the unexpected results. Thus, it is difficult to distinguish PIQ changes due to the FE and changes due to problems with the equated scores.

As with the current study, previous invariance research of the WAIS has suggested that the mean differences in the subtests cannot be explained solely by differences in the latent variable (e.g., Beaujean and Sheng, 2014; Wicherts et al., 2004). Interestingly, Wicherts et al.'s study's found noninvariance with the WAIS intercepts and their participants came from the 1967/1968–1998/1999 Dutch standardization of the WAIS. This period encompasses the 1981 and 1997 dates that the US WAIS-R and WAIS-III were published, for which we also found trouble at the level of the intercepts.

In contrast to previous FE research (e.g., te Nijenhuis & van der Flier, 2013), we found mean differences in *g*—at least when comparing the WAIS-III and WAIS-IV samples. Most of the studies that have concluded that the FE does not represent a change in *g* have used the method of correlated vectors (MCV). In the FE context, the MCV consists of: (a) extracting a *g* factor from two batteries of tests normed at different times, (b) examining invariance of the factor loadings using a congruence coefficient, (c) calculating the mean score differences between the two batteries, and (d) measure *g*'s effect by calculating the Spearman correlation between score differences and the *g* loadings from the combined group (Jensen, 1992). The MCV has been criticized for multiple reasons (Ashton & Lee, 2005; Dolan & Hamaker, 2001). One criticism is the use of congruence coefficients to examine invariance. In the current study, we did not find invariance for the WAIS-R and WAIS-III factors, yet the congruence coefficient for *g*, extracted using the Schmid-Leiman transformation, is $>.99$. Another criticism of the MCV is that interpretation of the effect values is ill defined. For example, the Spearman correlation between *g*, extracted using the Schmid-Leiman transformation, and the differences in subtest scores between the WAIS-III and WAIS-IV equated subtests is .34. Does that mean there is, or is not, a FE on *g*?

Limitations and Future Directions

As the current study only investigated three editions of the WAIS, we have only examined a portion of the instruments used to assess the FE. Future studies should follow our procedures with other instruments, such as the Wechsler Intelligence Scale for Children and Stanford-Binet, to determine if their scores are comparable and, if so, the magnitude of the FE.

Although the method we used allowed us to create a grand sample comprised of participants from three normative samples collected over a time period of close to 30 years, the respective normative samples are, nevertheless, only cross-sectional. Studies that combine cross-sectional and longitudinal designs (e.g., Schaie et al., 2005) could likely shed more light on the FE. Even with longitudinal studies, when the same tests are administered to the same persons at different points in time, the measurement scale and meaning of scores may change (Horn & McArdle, 1992; McArdle & Cattell, 1994). Thus, studies that incorporate a longitudinal design in which the same version of the WAIS is administered to the same persons at different points in time and the scales are assessed for invariance could add to our understanding of the FE.

The present study is the first we are aware of that has examined the FE using a bi-factor model. As we discussed in the Introduction, the advantages of using a bi-factor model are manifold, but it is not the only model used to explain the covariance of the WAIS subtests. For example Weiss et al. (2013) argued that a higher order model is better for the WAIS than a bi-factor model. Likewise, using an eight-subtest version of the WAIS-R, Horn and McArdle (1992) argued for using a two-factor model based on the theory of fluid and crystallized abilities. Unlike most other two-

factor models, they allowed all subtests to load on both latent variables. Unlike the single-factor or two-factor model with loadings constrained to be zero, their full model was invariant across all their age groups. Consequently, future FE studies should look to examine if there is an influence of the factor model used in both assessing for invariance over instruments (e.g., Irwing, 2012) as well as measuring the magnitude of the FE.

Related to the issue of the factor model used for the WAIS is the model used to examine the FE. As the investigation of the FE is really an examination of change, there are a variety of methods available to assess this change (McArdle, 2009). We believe our use of a multi-group latent variable model using equated subtest scores was a robust method for handling the complexities involved with the Wechsler standardization and linking data that is in line with best practices for measuring change (McArdle & Prindle, 2013). Nonetheless, future research should compare our results with other robust ways of measuring change to see the influence of the methods. For example, Jensen (1998) noted the practical significance of any change believed to reflect the FE should be evaluated using tests of predictive bias.

Implications of the Current Study

There are four major implications from this study. First, comparing scores between instruments is a tenuous undertaking, which does not lessen just because the scores come from different editions of the same instrument. This is not necessarily because norms are obsolete (Flynn, 1998), but because the different instruments have different metrics (i.e., scales, origins). Thus, the default stance should likely be that IQ instruments' scores are on their own metrics, and not directly comparable. Only after sufficient work has been published indicating the scores are invariant and psychometric techniques have been employed to equate the instruments should the scores be compared.

Examining comparability of scores is not a novel idea, but it is one that has escaped most FE research. Although research suggests that *g* can be measured dependably and is strongly correlated across different batteries of tests (Floyd et al., 2013; Floyd, Shands, Rafael, Bergeron, & McGrew, 2009; Johnson, Bouchard, Krueger, McGue, & Gottesman, 2004; Johnson, te Nijenhuis, & Bouchard, 2008; Major, Johnson, & Bouchard, 2011), relying only on correlations to determine comparability will often lead to misleading results when attempting to quantify the FE and make score comparisons. While the present study suggests comparability across the third and fourth edition of the WAIS, it is important to keep in mind that subtest scores were equated across editions prior to multi-group comparison. Previous FE studies have not equated scores across editions before comparing values.

The second major implication is that if WAIS scores are used as the criterion for determining if American adults are getting smarter over time, then the evidence is modest. Although mean full-scale IQ (FSIQ) scores may appear to be increasing over time (Flynn, 1984, 1998, 2009b), part of this increase can be attributed to the test revision process (i.e., beta change)—at least until 1997 when the WAIS-III was published. Similarly, the stability of the FE over time is difficult to gauge because scores obtained from the WAIS-R are not equivalent to scores obtained from the WAIS-III, and the comparability of the original WAIS and WAIS-R scores is unknown, although we doubt they are comparable (Beaujean & Sheng, 2014). As the WAIS-III and WAIS-IV subtests showed invariance, we can state that over the approximately 11 years between the instruments' publication, the FSIQ increased 0.40 IQ points a year and *g* increased 0.51 IQ points a year, both of which are within the typically espoused 3- to 5-point IQ gain per decade range for the FE.

The third major implication is that the FE was observed only for *g*. Flynn's (2012) belief regarding the FE is that it arises largely from gains on specific tasks. Notably, Flynn points out that the Wechsler Similarities subtest and Raven's matrices show the largest gains. The Similarities subtest has a high *g* loading and Raven's matrices are viewed as measures of fluid reasoning, a group factor that is often statistically indistinguishable from *g* (Reynolds, Keith, Flanagan, & Alfonso, 2013). As fluid reasoning reflects abilities such as making abstractions and solving novel problems, and fluid reasoning is often statistically indistinguishable from *g*, our findings are consistent with Fox and Mitchum's (2013) hypothesis that the FE reflects improvements in the ability to "map objects at higher levels of abstraction" (p. 979). In higher order models *g* will cause mean differences in group factors, as group factors are not independent of *g*. The use of a bi-factor model makes it clear that the FE, at least as measured by the third and fourth editions of the WAIS, does indeed reflect gains in *g*. The FE was not observed for group factors. We believe that if scholars want to examine the FE in areas beyond *g*, they should employ bi-factor models instead of using higher order model or analyzing specific subtests.

The last implication of this study is that there needs to be more discussion and research on how to compare scores when measurement invariance is not found between instruments. In the short term, such solutions as using scores derived from invariant subtests or using any between-group intercept differences to correct the subtest scores might be useful. These are only stopgap solutions, though, and will become obsolete as new instruments are published. Long-term solutions will require developing and scaling IQ tests that are invariant across time. For an instrument undergoing revision, one possible solution would be to place the aggregate scores on the same metric as the previous edition. For example, construct the index scores from the fifth edition of the WAIS to be on the same metric as the WAIS-IV, making the instruments' scores directly comparable, literally, out of the box. With new

instruments, the solution will be more complex. One possible solution would be to construct the scores to be on the same metric as a referent instrument. For example, any scores from new adult intelligence test that NCS Pearson/Psychological Corporation (the company responsible for the WAIS) publishes would be constructed to be directly comparable to the WAIS-IV. Similar procedures could be used for instruments produced by other test publishers.

ARTICLE INFORMATION

Conflict of Interest Disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical Principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was not supported.

Role of the Funders/Sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication. [Not Applicable]

Acknowledgments: The authors are grateful to NCS Pearson for providing the data used in this research. Standardization data from the Wechsler Adult Intelligence Scale.s revised (WAIS-R), third (WAIS-III), and fourth (WAIS-IV) editions were used with permission. Copyright 1981, 1997, and 2008 by NCS Pearson, Inc. All rights reserved. "Wechsler Adult Intelligence Scale" and "WAIS" are trademarks, in the US and other countries, of Pearson Education, Inc. or its affiliate(s). The authors would like thank Jack McArdle, Joe Rodgers, and two anonymous reviewers for their comments on prior versions of this manuscript. The ideas and opinions expressed herein are those of the authors alone, and endorsement by the authors' institutions is not intended and should not be inferred.

SUPPLEMENTAL DATA

Supplemental data for this article can be accessed on the publisher's website.

REFERENCES

Albano, A. (2011). *equate: Statistical methods for test score equating.* [Computer software]. Retrieved from http://CRAN.R-project.org/package=equate

Ashton, M. C., & Lee, K. (2005). Problems with the method of correlated vectors. *Intelligence*, *33*, 431–444. doi:10.1016/j.intell.2004.12.004

Asparouhov, T., & Muthén, B. O. (2005, November). *Multivariate statistical modeling with survey data*. Paper presented at the annual meeting of the Federal Committee on Statistical Methodology Research Conference, Arlington, VA.

Beaujean, A. A. (2014). *Latent variable modeling using **R**: A step by step guide*. New York, NY: Routledge/Taylor and Francis.

Beaujean, 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., Parkin, J., & Parker, S. (2014). Comparing Cattell-Horn-Carroll models for predicting language achievement: Differences between bi-factor and higher-order factor models. *Psychological Assessment*, *26*, 789–805. doi:10.1037/a0036745

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*(3), 294–298. doi:10.1016/j.paid.2009.10.019

Beaujean, A. A., & Sheng, Y. (2014). Assessing the Flynn effect in the Wechsler scales. *Journal of Individual Differences*, *35*, 63–78. doi:10.1027/1614-0001/a000128

Blair, C., Gamson, D., Thorne, S., & Baker, D. (2005). Rising mean IQ: Cognitive demand of mathematics education for young children, population exposure to formal schooling, and the neurobiology of the prefrontal cortex. *Intelligence*, *33*(1), 93–106. doi:10.1016/j.intell.2004.07.008

Boake, C. (2002). From the Binet-Simon to the Wechsler-Bellevue: Tracing the history of intelligence testing. *Journal of Clinical and Experimental Neuropsychology*, *24*(3), 383–405. doi:10.1076/jcen.24.3.383.981

Brunner, M. (2008). No *g* in education? *Learning and Individual Differences*, *18*(2), 152–165. doi:10.1016/j.lindif.2007.08.005

Camara, W. J., Nathan, J. S., & Puente, A. E. (2000). Psychological test usage: Implications in professional psychology. *Professional Psychology: Research and Practice*, *31*(2), 141–154. doi:10.1037/0735-7028.31.2.141

Carroll, J. B. (1997). Theoretical and technical issues in identifying a factor of general intelligence. In B. Devlin, S. E. Fienberg, D. P. Resnick, & K. Roeder (Eds.), *Intelligence, genes, and success: Scientists respond to the bell curve* (pp. 125–156). New York, NY: Springer-Verlag.

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

Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research*, *41*(2), 189–225. doi:10.1207/s15327906mbr4102_5

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–255. doi:10.1207/S15328007SEM0902_5

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. doi:10.1016/j.intell.2004.07.010

Dolan, C. V., & Hamaker, E. L. (2001). Investigating Black-White differences in psychometric IQ: Multi-group confirmatory factor analyses of the WISC-R and K-ABC, and a critique of the method of correlated vectors. In F. Columbus (Ed.), *Advances in psychological research* (Vol. 6, pp. 31–60). Huntington, NY: Nova Science.

Enders, C. K. (2001). The impact of nonnormality on full information maximum-likelihood estimation for structural equation models with missing data. *Psychological Methods*, *6*(4), 352–370. doi:10.1037/1082-989X.6.4.352

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*(3), 430–457. doi:10.1207/S15328007SEM0803_5

Floyd, R. G., Bergeron, R., McCormack, A. C., Anderson, J. L., & Hargrove-Owens, G. L. (2005). Are Cattell-Horn-Carroll (CHC) broad ability composite scores exchangeable across batteries? *School Psychology Review*, *34*(3), 386–341.

Floyd, R. G., Clark, M. H., & Shadish, W. R. (2008). The exchangeability of IQs: Implications for professional psychology. *Professional Psychology: Research and Practice*, *39*(4), 414–423. doi:10.1037/0735-7028.39.4.414

Floyd, R. G., Reynolds, M. R., Farmer, R. L., & Kranzler, J. H. (2013). Are the general factors from different child and adolescent intelligence tests the same? Results from a five-sample, six-test analysis. *School Psychology Review*, *42*(4), 383–401.

Floyd, R. G., Shands, E. I., Rafael, F. A., Bergeron, R., & McGrew, K. S. (2009). The dependability of general-factor loadings: The effects of factor-extraction methods, test battery composition, test battery size, and their interactions. *Intelligence*, *37*(5), 453–465. doi:10.1016/j.intell.2009.05.003

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. (1996). What environmental factors affect intelligence: The relevance of IQ gains over time. In D. Detterman (Ed.), *Current topics in human intelligence (*vol. *5): The environment.* Norwood, NJ: Ablex.

Flynn, J. R. (1998). WAIS-III and WISC-III gains in the United States from 1972 to 1995: How to compensate for obsolete norms. *Perceptual and Motor Skills*, *86*(3, Pt 2), 1231–1239. doi:10.2466/pms.1998.86.3c.1231

Flynn, J. R. (2006). Tethering the elephant: Capital cases, IQ, and the Flynn effect. *Psychology, Public Policy, And Law*, *12*(2), 170–189. doi:10.1037/1076-8971.12.2.170

Flynn, J. R. (2009a). The WAIS-III and WAIS-IV: Daubert motions favor the certainly false over the approximately true. *Applied Neuropsychology*, *16*(2), 98–104. doi:10.1080/09084280902864360

Flynn, J. R. (2009b). *What is intelligence? Beyond the Flynn effect*. Cambridge UK: Cambridge University [Expanded paperback edition].

Flynn, J. R. (2012). *Are we getting smarter? Rising IQ in the twenty-first century*. New York, NY: Cambridge University.

Flynn, J. R., & Rossi-Casé, L. (2012). IQ gains in Argentina between 1964 and 1998. *Intelligence*, *40*(2), 145–150. doi:10.1016/j.intell.2012.01.006

Fox, M. C., & Mitchum, A. L. (2013). A knowledge-based theory of rising scores on "culture-free" tests. *Journal of Experimental Psychology: General*, *142*(3), 979–1000. doi:10.1037/a0030155

Gignac, G. E. (2005). Revisiting the factor structure of the WAIS-R: Insights through nested factor modeling. *Assessment*, *12*, 320–329.

Gignac, G. E. (2006). The WAIS-III as a nested factors model: A useful alternative to the more conventional oblique and higher-order models. *Journal of Individual Differences*, *27*, 73–86.

Gignac, G. E. (2008). Higher-order models versus direct hierarchical models: *g* as superordinate or breadth factor? *Psychology Science Quarterly*, *50*(1), 21–43.

Gignac, G. E., & Watkins, M. W. (2013). Bifactor modeling and the estimation of model-based reliability in the WAIS-IV. *Multivariate Behavioral Research*, *48*(5), 639–662. doi:10.1080/00273171.2013.804398

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*(2), 133–157. doi:10.1177/002188637601200201

Greshman, F., & Reschly, D. J. (2011). Standard of practice and Flynn effect testimony in death penalty cases. *Intellectual and Developmental Disabilities*, *49*(3), 131–140. doi:10.1352/1934-9556-49.3.131

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

Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics*, *25*(2), 133–183. doi:10.3102/10769986025002133

Holland, P. W., & Thayer, D. T. (1987). Notes on the use of log-linear models for fitting discrete probability distributions (Technical Report 87–79). Princeton, NJ: Educational Testing Service.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*(3–4), 117–144.

Irwing, P. (2012). Sex differences in *g*: An analysis of the US standardization sample of the WAIS-III. *Personality and Individual Differences*, *53*, 126–131. doi:10.1016/j.paid.2011.05.001

Jensen, A. R. (1992). Spearman's hypothesis: Methodology and evidence. *Multivariate Behavioral Research*, *27*, 225–233. doi:10.1207/s15327906mbr2702_5

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

Jensen, A. R. (2004). Obituary: John Bissell Carroll. *Intelligence*, *32*, 1–5. doi:10.1016/j.intell.2003.10.001

Johnson, W., Bouchard, T. J., Krueger, R. F., McGue, M., & Gottesman, I. I. (2004). Just one *g*: Consistent results from three test batteries. *Intelligence*, *32*(1), 95–107. doi:10.1016/S0160-2896(03)00062-X

Johnson, W., te Nijenhuis, J., & Bouchard, T. J. (2008). Still just 1 *g*: Consistent results from five test batteries. *Intelligence*, *36*(1), 81–95. doi:10.1016/j.intell.2007.06.001

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. doi:10.1037/0003-066X.58.10.778

Kanaya, T., Ceci, S. J., & Scullin, M. H. (2005). Age differences within secular IQ trends: An individual growth modeling approach. *Intelligence*, *33*(6), 613–621. doi:10.1016/j.intell.2005.08.001

Kane, H. D., & Oakland, T. D. (2000). Secular declines in Spearman's g: Some evidence from the United States. *The Journal of Genetic Psychology*, *161*, 337–345.

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). Flynn effect [Special issue]. *Journal of Psychoeducational Assessment*, *28*(5).

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices (3rd ed.)*. New York, NY: Springer-Verlag.

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.

Lord, F. M. (1982). The standard error of equipercentile equating. *Journal of Educational and Behavioral Statistics*, *7*(3), 165–174. doi:10.3102/10769986007003165

Lynn, R. (1998). In support of the nutrition theory. In U. Nesser (Ed.), *The rising curve: Long-term gains in IQ and related measures* (pp. 207–215). Washington, DC: American Psychological Association.

Lynn, R. (2009). What has caused the Flynn effect? Secular increases in the Development Quotients of infants. *Intelligence*, *37*(1), 16–24. doi:10.1016/j.intell.2008.07.008

Lynn, R. (2013). Who discovered the Flynn effect? A review of early studies of the secular increase of intelligence. *Intelligence*, *41*(6), 765–769. doi:10.1016/j.intell.2013.03.008

Major, J. T., Johnson, W., & Bouchard, T. J. (2011). The dependability of the general factor of intelligence: Why small, single-factor models do not adequately represent *g*. *Intelligence*, *39*(5), 418–433. doi:10.1016/j.intell.2011.07.002

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology*, *60*(1), 577–605. doi:10.1146/annurev.psych.60.110707.163612

McArdle, J. J., & Cattell, R. B. (1994). Structural equation models of factorial invariance in parallel proportional profiles and oblique confactor problems. *Multivariate Behavioral Research*, *29*(1), 63–113. doi:10.1207/s15327906mbr2901_3

McArdle, J. J., & Prindle, J. J. (2013). Basic issues in the measurement of change. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J. I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology, Vol. 1: Test theory and testing and assessment in industrial and organizational psychology* (pp. 223–243). Washington, DC: American Psychological Association.

McClain, A. L. (1996). Hierarchical analytic methods that yield different perspectives on dynamics: Aids to interpretation. In B.Thompson (Ed.), *Advances in social science methodology* (Vol. 4, pp. 229–240). Bingley, England: Emerald Group.

McDermott, P. A., Watkins, M. W., & Rhoad, A. M. (2014). Whose IQ is it? Assessor bias variance in high-stakes psychological assessment. *Psychological Assessment*, *26*(1), 207–214. doi:10.1037/a0034832

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal Of Applied Psychology*, *93*(3), 568–592. doi:10.1037/0021-9010.93.3.568

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*, 525–543. doi:10.1007/BF02294825

Millsap, R. E., & Hartog, S. B. (1988). Alpha, beta, and gamma change in evaluation research: A structural equation approach. *Journal of Applied Psychology*, *73*(3), 574–584. doi:10.1037/0021-9010.73.3.574

Mingroni, M. A. (2004). The secular rise in IQ: Giving heterosis a closer look. *Intelligence*, *32*(1), 65–83. doi:10.1016/S0160-2896(03)00058-8

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

Moses, T., & Holland, P. W. (2009). Selection strategies for univariate loglinear smoothing models and their effect on equating function accuracy. *Journal of Educational Measurement*, *46*(2), 159–176. doi:10.1111/j.1745-3984.2009.00075.x

Murray, A. L., & Johnson, W. (2013). The limitations of model fit in comparing the bi-factor versus higher-order models of human cognitive ability structure. *Intelligence*, *41*(5), 407–422. doi:10.1016/j.intell.2013.06.004

Must, O., & Must, A. (2013). Changes in test-taking patterns over time. *Intelligence*, *41*(6), 780–790. doi:10.1016/j.intell.2013.04.005

Must, O., Must, A., & Raudik, V. (2003). The secular rise in IQs: In Estonia, the Flynn effect is not a Jensen effect. *Intelligence*, *31*(5), 461–471. doi:10.1016/S0160-2896(03)00013-8

Must, O., te Nijenhuis, J., Must, A., & van Vianen, A. E. M. (2009). Comparability of IQ scores over time. *Intelligence*, *37*(1), 25–33. doi:10.1016/j.intell.2008.05.002

Nettelbeck, T., & Wilson, C. (2004). The Flynn effect: Smarter not faster. *Intelligence*, *32*(1), 85–93. doi:10.1016/S0160-2896(03)00060-6

Parkin, J., & Beaujean, A. A. (2012). The effects of Wechsler Intelligence Scale for Children Fourth Edition cognitive abilities on math achievement. *Journal of School Psychology*, *50*, 113–128. doi:10.1016/j.jsp.2011.08.003

Pietschnig, J., Tran, U. S., & Voracek, M. (2013). Item-response theory modeling of IQ gains (the Flynn effect) on crystallized intelligence: Rodgers' hypothesis yes, Brand's hypothesis perhaps. *Intelligence*, *41*(6), 791–801. doi:10.1016/j.intell.2013.06.005

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*(5), 667–696. doi:10.1080/00273171.2012.715555

Reynolds, M. R., & Keith, T. Z. (2013). Measurement and statistical issues in child assessment research. In D. H. Saklofske, V. L. Schwean, & C. R. Reynolds (Eds.), *The Oxford Handbook of Child Psychological Assessment* (pp. 48–83). New York, NY: Oxford University.

Reynolds, M. R., Keith, T. Z., Flanagan, D. P., & Alfonso, V. C. (2013). A cross-battery, reference variable, confirmatory factor analytic investigation of the CHC taxonomy. *Journal of School Psychology*, *51*(4), 535–555. doi:10.1016/j.jsp.2013.02.003

Rodgers, J. L. (1998). A critique of the Flynn Effect: Massive IQ gains, methodological artifacts, or both? *Intelligence*, *26*(4), 337–356. doi:10.1016/S0160-2896(99)00004-5

Rosseel, Y. (2012). lavaan: An **R** package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. Retrieved from http://www.jstatsoft.org/v48/i02/

Schaie, K. W., Willis, S. L., & Pennak, S. (2005). An historical framework for cohort differences in intelligence. *Research in Human Development*, *2*(1), 43–67.

Schmiedek, F., & Li, S.-C. (2004). Toward an alternative representation for disentangling age-associated differences in general and specific cognitive abilities. *Psychology and Aging*, *19*(1), 40–56. doi:10.1037/0882-7974.19.1.40

Shiu, W., Beaujean, A., Must, O., te Nijenhuis, J., & Must, A. (2013). An item-level examination of the Flynn effect on the National Intelligence Test in Estonia. *Intelligence*, *41*(6), 770–779. doi:10.1016/j.intell.2013.05.

Shiu, W., Beaujean, A. A., & Wells, K. (2015). *A meta-analysis of the Flynn effect in the United States*, United Kingdom, and Australia [Manuscript submitted for publication].

Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement*, *47*(2), 150–174. doi:10.1111/j.1745-3984.2010.00106.x

Sinharay, S., Puhan, G., & Haberman, S. J. (2011). An NCME instructional module on subscores. *Educational Measurement: Issues and Practice*, *30*(3), 29–40. doi:10.1111/j.1745-3992.2011.00208.x

Spearman, C. (1904). "General intelligence" objectively determined and measured. *American Journal of Psychology*, *15*, 201–293. doi:10.2307/1412107

Steinmetz, H. (2013). Analyzing observed composite differences across groups: Is partial measurement invariance enough? *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *9*(1), 1–12. doi:10.1027/1614-2241/a000049

te Nijenhuis, J., & van der Flier, H. (2013). Is the Flynn effect on *g*?: A meta-analysis. *Intelligence*, *41*(6), 802–807. doi:10.1016/j.intell.2013.03.001

te Nijenhuis, J., van Vianen, A. M., & van der Flier, H. (2007). Score gains on *g*-loaded tests: No *g*. *Intelligence*, *35*(3), 283–300. doi:10.1016/j.intell.2006.07.006

Wai, J., & Putallaz, M. (2011). The Flynn effect puzzle: A 30-year examination from the right tail of the ability distribution provides some missing pieces. *Intelligence*, *39*(6), 443–455. doi:10.1016/j.intell.2011.07.006

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

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

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

Weiss, L. G. (2010). Considerations on the Flynn Effect. *Journal of Psychoeducational Assessment*, *28*, 482–493. doi:10.1177/0734282910373572

Weiss, L. G., Keith, T. Z., Zhu, J., & Chen, H. (2013). Technical and practical issues in the structure and clinical invariance of the Wechsler scales: A rejoinder to commentaries. *Journal of Psychoeducational Assessment*, *31*(2), 235–243. doi:10.1177/0734282913478050

West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 209–231). New York, NY: Guilford.

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

Williams, R. L. (2013). Overview of the Flynn effect. *Intelligence*, *41*(6), 753–764. doi:10.1016/j.intell.2013.04.010

Woodley, M. A., te Nijenhuis, J., & Murphy, R. (2013). Were the Victorians cleverer than us? The decline in general intelligence estimated from a meta-analysis of the slowing of simple reaction time. *Intelligence*, *41*(6), 843–850. doi:10.1016/j.intell.2013.04.006

Zhou, X., Zhu, J., & Weiss, L. G. (2010). Peeking inside the "black box" of the Flynn effect: Evidence from three Wechsler instruments. *Journal of Psychoeducational Assessment*, *28*(5), 399–411. doi:10.1177/0734282910373340