The Flynn Effect and Its Critics: Rusty Linchpins and "Lookin' for g and Gf in Some of the Wrong Places"

Kevin S. McGrew

Journal of Psychoeducational Assessment 2010 28: 448 originally published online 7 July 2010
DOI: 10.1177/0734282910373347

Journal of Psychoeducational Assessment
28(5) 448–468

DOI: 10.1177/0734282910373347
http://jpa.sagepub.com

The Flynn Effect and Its Critics: Rusty Linchpins and "Lookin' for g and Gf in Some of the Wrong Places"

Kevin S. McGrew[1,2]

Abstract

The consensus of most intelligence scholars is that the Flynn effect (FE) is real, IQ test batteries are now routinely restandardized on a regular basis. A cornerstone in Flynn's explanation of the FE is his analysis of select Wechsler subtest scores across time. The featured articles by Kaufman and Zhou, Zhu, and Weiss question whether Flynn's arguments are grounded in the unproven assumption that similarly named Wechsler scores measure the same constructs across editions. Kaufman raises the issue by means of a detailed task analysis of changes in test administration and scoring directions for similarly named tests across different Wechsler editions. The author applauds Zhou et al. for bringing methodological rigor to the comparison of similarly named Wechsler Performance composite scores across time. Unfortunately, both Kaufman and Zhou et al. inadvertently perpetuate some of Flynn's incorrect interpretations of select Wechsler measures (Similarities and Performance tests) as measures of the novel abstract problem solving that characterizes fluid intelligence (Gf). The author presents empirical Wechsler subtest *g*-loadings based on seven Wechsler joint- or cross-battery factor analyses (with other cognitive batteries). The results suggest that the extant Wechsler FE data and its system of interpretations, hypotheses, and resultant theory are held together by multiple anchors, a number that, in the words of Kaufman, are "seriously coated in rust." The author briefly discusses the theory, tools, and technologies that currently exist to place a more reasonable degree of order in the house built by Flynn.

Keywords

Flynn effect, CHC theory, fluid intelligence, general intelligence, Wechsler batteries

> Progress in science depends on new techniques, new discoveries and new ideas, probably in that order.
>
> —Sydney Brenner (*Nature*, May 5, 1980)

[1]Woodcock-Muñoz Foundation, Olympia, WA, USA
[2]University of Minnesota, Minneapolis, MN, USA

Corresponding Author:
Kevin McGrew, 1313 Pondview Lane E., Minneapolis, MN 56374, USA
Email: iap@earthlink.net

The consensus of most (but not all) intelligence scholars is that the Flynn effect (FE), at the level of the global IQ score, is real (see Kaufman, 2010, Zhou, Zhu, & Weiss, 2010, for a definition of the FE). IQ tests require periodic updating of their norms to account for changes in the population over time. However, I also agree with Rodgers (1999) that " . . . the acceptance of the effect has been too quick. The proper explanations for the effect will not be meaningful until the nature of the effect is much better understood than it is now" (p. 337). Furthermore, Rodgers (1999) concluded that

> Flynn's arguments contain methodological weaknesses of which he was unaware, of which the community of researchers has not been sufficiently critical. Because his self-evaluation was also blind to these weaknesses, it tends to overstate the confidence we should have in the status of the Flynn Effect. (p. 338)

Because of space constraints, my response focuses on select logical and methodological "blind spots" of Flynn's analysis and arguments, particularly those unintentionally perpetuated by Kaufman and Zhou et al. I will not address hypothesized causal mechanisms of the FE "black box." Although my charge is to critique the Kaufman and Zhou et al. feature articles, it is impossible to perform this function without concurrently discussing significant pieces of Flynn's research.

Are Similarly Named Wechsler Subtests and Composites Measuring the Same Ability Across Different Editions?

The Kaufman (2010) and Zhou et al. (2010) articles both suggest that Flynn's arguments are grounded in the unproven assumption that similarly named Wechsler subtests measure the same ability constructs across editions. Kaufman correctly raises the subtest issue via a detailed task analysis of the changes in test administration and scoring directions for similarly named tests across different Wechsler editions. Zhou et al. attempt a more frontal attack on the thorny issue of measurement scale equivalence through the application of empirical equating of the Performance scale composites prior to their comparative statistical analyses. I applaud Zhou et al.'s attempt to bring methodological rigor (scale equating) to the comparison of similarly named Wechsler Performance composite scores across editions.

The Need to Look Under More Than One Lamplight

Unfortunately, both Kaufman and Zhou et al. perpetuate a fundamental flaw in Flynn's interpretation of the Wechsler scores across time. All three suffer from the incorrect assumption that various combinations of Wechsler Performance tests (until the most recent editions) and/or the Similarities test are good proxies of fluid intelligence (Gf). Contrary to clinical lore and many Wechsler interpretation books, the Wechsler batteries only began to include strong indicators of Gf with the publication of the WISC-IV (Wechsler Intelligence Scale for Children–Fourth Edition) and WAIS-III/IV (Wechsler Adult Intelligence Scale; see extant Wechsler Cattell–Horn–Carroll or CHC theory of intelligence joint factor studies; Flanagan, McGrew, & Ortiz, 2000; Flanagan, Ortiz, & Alfonso, 2007; McGrew & Flanagan, 1998; Woodcock, 1990).[1] According to Roberts, Markham, Matthews, and Zeidner (2005), when discussing the WAIS-III, "the test includes a measure of (pure) fluid intelligence *for the very first time (i.e., Matrices)*" [italics added] (p. 339). Even if scale equivalence of the Similarities and Performance tests across editions

can be empirically established, contemporary psychometric (i.e., CHC theory) research does not support the interpretation of these equated measures as strong indicators of Gf—a key foundation of Flynn's arguments.

Flynn (Flynn, 2007; Flynn & Weiss, 2007), Kaufman, and Zhou et al., and for that matter much of psychology, ground key Wechsler-based FE interpretations on *internal* structural validity factor analysis studies of the Wechsler tests. In contrast, *joint- or cross-battery* factor analysis includes tests from beyond the confines of a single intelligence battery, often using tests from at least one other intelligence battery (see McGrew & Flanagan, 1998) to provide external test construct validity evidence (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). Basing Wechsler subtest interpretation on internal factor analysis studies is analogous to the old story of the drunk looking for his keys under the lamplight. As the story is typically told, one night a drunk loses his keys a distance away from a nearby lamplight but only looks under the lamplight, because it is the only place where he can see. Test *g*-loadings and ability classifications (Gf, Gc, Gv, etc.) derived only from internal validity studies run the risk of misinterpretation of the abilities measured by the Wechsler subtests.[2] When the Wechsler tests have been examined under multiple lamplights, particularly enough lamplights (ability indicators) to shed light on the major CHC broad ability domains, some Wechsler test *g*-loadings and traditional ability classifications often change (see Flanagan et al., 2000; McGrew, 1997, 2005, 2009; McGrew & Flanagan, 1998; Woodcock, 1990).

Across the extant CHC-organized Wechsler joint-factor analysis studies, Similarities loaded 100% of the time on a robust Gc factor. To the best of my knowledge, Similarities has never loaded on Gf factors defined by strong Gf indicators (e.g., matrix reasoning tasks, deductive and/or inductive reasoning "learning" tasks). Even when examining the internal factor structures of the WISC-IV and WAIS-IV, the editions that include the greatest number of good Gf indicators (e.g., Matrix Reasoning, Figure Weights, Picture Concepts), Similarities consistently loads high (.70+) on the Verbal Comprehension (Gc) factor and displays no significant loading on the Perceptual Reasoning (Gf + Gv) factor (Wechsler, 2004, 2008). Furthermore, Similarities average correlation (approximately .50) with the most Raven-like test in the WISC-IV/WAIS-IV (Matrix Reasoning) indicates that Similarities shares only 25% common variance with a matrix reasoning Gf-type test. The Similarities test is no Raven's. "Quoth the Raven's, Nevermore—should Similarities be equated with me."[3]

Researchers and practitioners must resist the historical and traditional (and incorrect) Gf/abstract reasoning interpretation of Similarities that is often presented in seductive, face-valid, eloquent logical task analysis descriptions of the test. For example, Flynn's (2007) description of the demands of Similarities items may be hard to resist:

> All of the WISC scoring manuals show that most points are given for answers that are "abstract" rather than concrete. Classifying the world in terms of abstract rather than operational categories signals the spread of the scientific ethos. Saying that "dogs hunt rabbits" is a pre-scientific answer and gets no points. Saying that they are both "mammals" get full marks. Finding it natural to see the world through scientific spectacles is a prerequisite for success at Similarities. It is not just that biological, chemical, and astronomical terms are preferred, it is a matter of regarding the world as something to be classified rather than manipulated. So most items do not set a problem of logical inference so much as a problem of classification. They call for higher level abstract reasoning skills which emphasize scientific ways of thinking about the world. And these skills are molded by contemporary formal education. (p. 217)

Such descriptions, laced with reference to intellectually inspiring phrases such as "scientific ethos," "scientific spectacles," and "scientific ways of thinking about the world" are no substitute for empirical data. Flynn's (Flynn, 2007; Flynn & Weiss, 2007) treatment of Similarities as a strong Gf indicator of abstract problem-solving or on-the-spot-thinking (like the Raven's), which Kaufman (2010) appears to implicitly accept in his article, is wrong.[4] It is a fatal flaw in a key cornerstone in Flynn's arguments. As demonstrated in the series of analyses presented later in this article, multiple linchpins in Flynn's interpretation of the Wechsler subtests, which in turn serve as the foundation for his FE theorization, rest on nonempirically grounded, logically based traditional test interpretations. The arguments may be verbally and logically elegant, but they are data and theory poor.[5]

Wechsler Performance Tests Are Poor Indicators of Gf

Zhou et al. (2010) address one critical substantive issue (i.e., is the FE constant across levels of intelligence?) and one critical measurement issue—first establishing "Gf" construct measurement scale equivalence before interpreting FE-related findings. However, contrary to the clinical lore and many IQ test interpretation books, the Wechsler batteries only began to include strong indicators of Gf with the publication of the WISC-IV and WAIS-III/IV (McGrew, 2010; Roberts et al., 2005).

Table 1 summarizes the CHC composition of the different Wechsler Performance test-based Gf proxies used in the Zhou et al. analysis.[6] Inspection of Table 1 again suggests a comparison of apples and oranges; or at least different mixtures of apples and oranges. At one extreme is the Gv dominated (75%) WISC-III POI comparison with the WISC-IV PRI (a relatively stronger Gf proxy; 67% of the tests classified as medium-to-high indicators of Gf). Although not as extreme, the WAIS-R/WAIS-III PIQ comparison suffers from the same "different mixtures of fruit" flaw. The WAIS-R "Gf" proxy includes no strong Gf tests and is dominated (60%) by Gv abilities. The WAIS-III PIQ is only slightly improved via the inclusion of the single Gf Matrix Reasoning test. The WAIS-III POI/WAIS-IV PRI contrast comes closest to comparing similar composite construct indexes but, unfortunately, the similar composites are dominated by Gv (67%; only 33% measurement of Gf—the targeted construct of the various analyses). If the various flavors of equated PIQ composites were measuring the same ability construct, and if some degree of FE was operating, one would *not* expect the WAIS-III PIQ (in both the WAIS-R and WAIS-III, and WAIS-III and WAIS-IV comparisons) to be almost identical (approximately 103) over a period of 11 years (see Zhou et al., 2010, Table 2). I believe this finding, together with the PIQ differential ability composition comparison presented in Table1, fail to support Zhou et al.'s questionable assumption that the different versions of PIQ used across comparisons are measuring the same construct. Zhou et al. recognize this apples-and-oranges concern when, in their discussion of the changing mix of constructs included in different editions of the Wechsler batteries, they state "such change in test structure could also lead to variation from the expected rate of IQ change across time and/or among instruments." All FE analyses that have interpreted the WISC, WISC-R, WISC-III, WAIS, and WAIS-R Performance tests (or composites) as representing Gf (akin to that represented by the Raven's) were more likely drawing conclusions about Gv abilities (visual-spatial processing) flavored at times with some Gc and Gs, and minimal Gf-like abstract reasoning. Attempts to draw inferences about the changing nature of abstract problem solving or on-the-spot thinking (i.e., Gf) vis-à-vis different Wechsler Performance composites are questionable.

Table 1. CHC Composition of the Various Wechsler Performance Composites Analyzed by Zhou et al.

	WISC-III	WISC-V	WAIS-R	WAIS-III	WAIS-III	WAIS-IV
	POI	PRI	PIQ	PIQ	POI	PRI
Picture Completion	Gv/Gc		Gv/Gc	Gv	Gv	
Picture Arrangement	Gv/Gc		Gv/Gc	Gv/Gc		
Block Design	Gv	Gv	Gv	Gv	Gv	Gv
Object Assembly	Gv		Gv			
Matrix Reasoning		Gf		Gf	Gf	Gf
Picture Concepts		Gf				
Digit Sym/Coding			Gs	Gs		
Visual Puzzles						Gv
% CHC representation	POI	PRI	PIQ	PIQ	POI	PRI
Gf		66.7		20.0	33.3	33.3
Gv	75.0	33.3	60.0	50.0	66.7	66.7
Gc	25.0		20.0	10.0		
Gs			20.0	20.0		

Note: Information extracted from analysis by McGrew (2010).

"Lookin' for *g* and Gf in Several Wrong Places"

A test's "*g*-ness" has generally been accepted as one of the default currencies for designating certain tests (e.g., Raven's matrices) as the best indicators of general intelligence (*g*). The "*g* = degree of cognitive complexity" assumption, which is often at the core of many FE interpretations, is attributed to Jensen. Jensen (1998) proposed that cognitive complexity could be operationally quantified as per a test's loading on the first unrotated factor in factor or principal component analyses. The rationale is that performance on tests that are more cognitively complex involve abstract reasoning and problem solving and invoke a wider range of elementary cognitive processes (Jensen, 1998; Stankov, 2000, 2005), which in turn is reflected in high *g*-ness. Additionally, the strong relation between tests of fluid reasoning (Gf) and *g* has resulted in many intelligence scholars (Flynn included) developing a love affair with the Raven's test as one of (if not) the best single psychometric indicators of Gf or *g* (Stankov, 2005).

As highlighted in Kaufman's article, Flynn (Flynn, 2007; Flynn & Weiss, 2007) uses scores from the Raven's and Wechsler Similarities as key FE linchpins. The extant research literature generally supports the Raven's status (and other similar matrix tests) as a strong indicator of Gf (and possibly *g*). However, Flynn (and many others) makes the erroneous assumption that the Wechsler Similarities test should be accorded the same Gf status as the Raven's. Flynn (2007) also accords Gf-like status to the Wechsler Performance subtests when he makes much of an average gain of 17 points on the Performance tests. This historical interpretation of Wechsler Similarities and Performance tests ignores contemporary CHC-based factor analysis research. Unfortunately, both Kaufman (2010) and Zhou et al. (2010) inadvertently reinforce these key misinterpretations via their focus on similar interpretations and analysis of key Wechsler tests or composites.

Flynn's anointment of Gf status to Similarities is obvious. According to Flynn and Weiss (2007), the Similarities items "call for *higher level abstract reasoning skills* [italics added] which emphasize scientific ways of thinking about the world" (p. 217). Furthermore, "the two trends

that Similarities measures are reinforcing: the scientific ethos favors *abstract problem-solving* [italics added], learning to attack such problems renders the scientific ethos more and more relevant" (p. 218). Flynn equates the reported gains on the Raven's "as unambiguously ones of *on-the-spot problem-solving*" (italics added; p. 218). He binds Similarities and Raven's together when he states "both gains have a prerequisite in common: problems must be taken seriously even though they have *no obvious practical pay-off*" (italics added; p. 218). Flynn (Flynn & Weiss, 2007) clearly believes that both the Raven's and Similarities tests reflect the essence of Gf—novel problem solving (especially inductive and/or deductive reasoning) or what Flynn refers to as "on the spot problem solving" or "thinking on your feet." This is also reflected in Flynn's (2007) comment that "after all, Raven's and Similarities have little functional in common with subtests like Information and Vocabulary. The latter do not involve thinking on your feet" (p. 220).

Kaufman unfortunately appears to perpetuate this mistake by assuming that the Similarities subtest and other select Wechsler subtests are indicators of Gf-like abstract reasoning and problem solving (see Note 4). This is evident in Kaufman's (2010) statement that ". . . the problem is considerable for three subtests that measure *abstract reasoning* [italics added]: Similarities, Comprehension, and Picture Arrangement" (p. 390). In turn, Zhou et al. (2010) base their entire analysis on the assumption that the various Wechsler Performance composites are good proxies of Gf—"The PIQs are analyzed because, compared to the verbal scales, this composite is a *better measure of fluid intelligence*" (italics added; p. 401).

Wechsler Subtest g-Loadings: Corroded and Stressed FE Linchpin Indicators?[7]

Table 2 summarizes a series of joint-battery principal components analyses (completed for this article) of seven illustrative data sets, each of which included one edition of the WISC or WAIS together with other external validity ability indicators. As described previously, a variable's loading on the first unrotated component in principal component analysis is the most frequently used index of a variable's "*g*-ness."

The results from the two oldest data sets (1a, 1967; 1b, 1977) raise questions about the traditional Wechsler *g*-loading interpretations used by Flynn and other intelligence researchers.[8] Although the construct validity of the Illinois Test of Psycholinguistic Abilities (ITPA) theoretical model was a major concern during the early learning disability movement and its underlying test-based intervention program empirically refuted, Carroll (1993) considered the ITPA (for factor analysis purposes) a measure of a "child's lexical knowledge as well as ability to understand language of increasing *complexity*" (italics added; p. 152). It is interesting to note that the WISC tests that Flynn, Kaufman, and Zhou et al. interpret as measures of abstract reasoning or fluid intelligence are classified only as *medium* indicators of *g* when analyzed together with the ITPA subtests (see Table 1). In the WISC/ITPA grade one sample, Similarities is similar in *g*-loading to WISC Digit Span, Block Design, Mazes, Information, and Comprehension. The Similarities test is considerably lower in *g*-ness than WISC Vocabulary and Arithmetic. This pattern is even more distinct in the 37-variable Snow, Lohman, Marshalek, Yalow, and Webb (1977) analysis of high school data (Sample 1b) that included the WAIS and numerous classic tests used in factor analysis research conducted by Thurstone, Guilford, and the ETS work group that developed the Kit of Factor Referenced tests (see Marshalek, Lohman, & Snow, 1983, for a description of the sample and test variables). In Sample 1b, WAIS Similarities is near the bottom of the *medium* *g*-loading classification category, far below the *g*-loadings for WAIS Vocabulary, Arithmetic, and Information. More important is the large discrepancy between the Raven's (.78)

Table 2. Wechsler Subtest *g*-Loadings and Classifications From Joint or Cross-Battery Factor Analysis Studies With External Cognitive Validity Indicators in Seven Normal Samples

1a. WISC/ITPA, *n* = (unknown)			*g*	1b. WAIS/Snow Ref. Apt., *n* = 241			*g*
Grade 1 (1967)	*g*	h^2	Cls	High School (1977)	*g*	h^2	Cls
ITPA Auditory-Vocal Assn.	0.69	0.47	H	Terman Concept Mastery	0.81	0.66	H
WISC Vocabulary	**0.62**	**0.39**	**H**	Arithmetic Concepts	0.81	0.66	H
WISC Arithmetic	**0.59**	**0.35**	**H**	Reading Comprehension	0.79	0.62	H
ITPA Auditory-Vocal Auto.	0.56	0.31	M	Ravens Progressive Matrices	0.78	0.61	H
WISC Object Assembly	**0.56**	**0.31**	**M**	Reading Vocabulary	0.78	0.61	H
WISC Digit Span	**0.53**	**0.28**	**M**	Language Expression	0.77	0.60	H
WISC Block Design	**0.52**	**0.27**	**M**	Necessary Arithmetic Operations	0.77	0.60	H
WISC Similarities	***0.48***	***0.23***	***M***	Arithmetic Application	0.77	0.59	H
WISC Mazes	**0.47**	**0.22**	**M**	Arithmetic Computation	0.75	0.56	H
WISC Information	**0.47**	**0.22**	**M**	**WAIS Vocabulary**	**0.74**	**0.55**	**H**
WISC Comprehension	**0.45**	**0.20**	**M**	Word Transformations	0.74	0.54	H
WISC Picture Arrangement	**0.41**	**0.17**	**M**	Thurstone Letter Series	0.73	0.53	H
WISC Picture Completion	**0.39**	**0.15**	**M**	**WAIS Arithmetic**	**0.68**	**0.46**	**M**
ITPA Visual Decoding	0.32	0.10	L	**WAIS Information**	**0.67**	**0.45**	**M**
ITPA Motor Encoding	0.30	0.09	L	Surface Development	0.66	0.44	M
WISC Coding	**0.24**	**0.06**	**L**	Paper Folding	0.66	0.44	M
ITPA Visual-Motor Assn.	0.24	0.06	L	Language Mechanical	0.64	0.41	M
ITPA Visual-Motor Seq.	0.20	0.04	L	Hidden Figures	0.61	0.38	M
ITPA Visual Decoding	0.18	0.03	L	**WAIS Block Design**	**0.61**	**0.38**	**M**
ITPA Auditory Decoding	0.17	0.03	L	Language Spelling	0.61	0.37	M
ITPA Auditory-Vocal Seq.	0.15	0.02	L	Word Beginning Ending	0.60	0.36	M
				WAIS Comprehension	**0.57**	**0.33**	**M**
				Paper Form Board	0.56	0.31	M
				WAIS Similarities	***0.56***	***0.31***	***M***
				Visual Number Span	0.55	0.31	M
				Camouflage Words	0.52	0.27	L
				WAIS Object Assembly	**0.51**	**0.26**	**L**
				WAIS Digit Span	**0.50**	**0.25**	**L**
				Identical Pictures	0.50	0.25	L
				Finding A's	0.46	0.21	L
				WAIS Digit Symbol	**0.45**	**0.20**	**L**
				Auditory Letter Span	0.42	0.17	L
				WAIS Picture Completion	**0.40**	**0.16**	**L**
				Number Comparison	0.39	0.15	L
				WAIS Picture Arrangement	**0.37**	**0.14**	**L**
				Uses for Things	0.30	0.09	L
				Harshman Gestalt	0.29	0.08	L
				Street Gestalt	0.27	0.07	L
				Film Memory III	0.22	0.05	L

(continued)

Table 2. (continued)

1c. WJ/WAIS, $n = 78$			g	1d. WJ/WISC-R, $n = 167$			g
Grade 12 (1978)	g	h^2	Cls	Grades 3 and 5 (1978)	g	h^2	Cls
WJ Quantitative Concepts	0.85	0.72	H	WJ Antonyms-Synonyms	0.83	0.69	H
WAIS Vocabulary	**0.83**	**0.69**	**H**	WJ Quantitative Concepts	0.77	0.59	H
WJ Analogies	0.82	0.67	H	WJ Analogies	0.77	0.59	H
WAIS Information	**0.79**	**0.62**	**H**	**WISC-R Vocabulary**	**0.76**	**0.58**	**H**
WJ Antonyms-Synonyms	0.75	0.56	H	**WISC-R Information**	**0.72**	**0.52**	**H**
WAIS Arithmetic	**0.68**	**0.46**	**M**	**WISC-R Arithmetic**	**0.67**	**0.45**	**M**
WAIS Comprehension	**0.68**	**0.46**	**M**	**WISC-R Comprehension**	**0.62**	**0.38**	**M**
WAIS Picture Completion	**0.68**	**0.46**	**M**	WJ Picture Vocabulary	0.62	0.38	M
WJ Spatial Relations	0.65	0.42	M	***WISC-R Similarities***	***0.60***	***0.36***	***M***
WJ Memory for Sentences	0.63	0.40	M	WJ Numbers Reversed	0.60	0.36	M
WAIS Block Design	**0.62**	**0.38**	**M**	WJ Memory for Sentences	0.56	0.31	M
WAIS Picture Completion	**0.62**	**0.38**	**M**	**WISC-R Block Design**	**0.56**	0.31	M
WAIS Digit Span	**0.61**	**0.37**	**M**	WJ Analysis-Synthesis	0.55	0.30	M
WJ Analysis-Synthesis	0.60	0.36	M	WJ Visual-Auditory Learning	0.54	0.29	M
WJ Numbers Reversed	0.60	0.36	M	WJ Concept Formation	0.54	0.29	M
WJ Concept Formation	0.58	0.34	M	**WISC-R Digit Span**	**0.48**	**0.23**	**L**
WAIS Object Assembly	**0.56**	**0.31**	**M**	**WISC-R Object Assembly**	**0.47**	**0.22**	**L**
WAIS Similarities	***0.47***	***0.22***	***L***	**WISC-R Coding**	**0.47**	**0.22**	**L**
WJ Visual-Auditory Learning	0.43	0.18	L	WJ Spatial Relations	0.43	0.18	L
WAIS Coding	**0.41**	**0.17**	**L**	**WISC-R Picture Completion**	**0.42**	**0.18**	**L**
WJ Blending	0.41	0.17	L	WJ Blending	0.42	0.18	L
WAIS Picture Arrangement	**0.38**	**0.14**	**L**	**WISC-R Picture Arrangement**	**0.36**	**0.13**	**L**
WJ Visual Matching	0.26	0.07	L	**WISC-R Mazes**	**0.22**	**0.05**	**L**

1e. WJ-R/WISC-R, $n = 72$			g	1f. WJ III/WISC-III, $n = 148$			g
Grades 3-4 (1991)	g	h^2	Cls	Grades 3-5 (2001)	g	h^2	Cls
WJ-R Oral Vocabulary	0.83	0.69	H	WJ III Sound Awareness	0.85	0.72	H
WISC-R Vocabulary	**0.78**	**0.61**	**H**	WJ III Understanding Directions	0.81	0.66	H
WISC-R Similarities	***0.75***	***0.57***	***H***	WJ III Verbal Comprehension	0.76	0.58	H
WJ-R Concept Formation	0.75	0.56	H	WJ III Applied Problems	0.72	0.52	M
WJ-R Analogies	0.74	0.54	H	**WISC-III Vocabulary**	**0.70**	**0.49**	**M**
WISC-R Arithmetic	**0.73**	**0.54**	**H**	**WISC-III Arithmetic**	**0.70**	**0.49**	**M**
WISC-R Information	**0.72**	**0.52**	**M**	WJ III General Information	0.69	0.48	M

(continued)

Table 2. (continued)

I e. WJ-R/WISC-R, $n = 72$				If. WJ III/WISC-III, $n = 148$			
	g				*g*		
Grades 3-4 (1991)		h^2	Cls	Grades 3-5 (2001)		h^2	Cls
WISC-R Comprehension	**0.70**	**0.49**	**M**	***WISC-III Similarities***	***0.69***	***0.48***	***M***
WJ-R Picture Vocabulary	0.69	0.48	M	**WISC-III Information**	**0.68**	**0.46**	**M**
WJ-R Memory for Sentences	0.68	0.46	M	WJ III Concept Formation	0.65	0.42	M
WJ-R Listening Comprehension	0.66	0.44	M	WJ III Number Series	0.64	0.41	M
WISC-R Block Design	**0.65**	**0.42**	**M**	WJ III Number Matrices	0.61	0.37	M
WJ-R Analysis-Synthesis	0.65	0.42	M	**WISC-III Block Design**	**0.60**	**0.36**	**M**
WJ-R Spatial Relations	0.58	0.34	M	WJ III Calculation	0.59	0.35	M
WISC-R Digit Span	**0.57**	**0.32**	**M**	WJ III Story Recall	0.58	0.34	M
WISC-R Object Assembly	**0.57**	**0.32**	**M**	WJ III Analysis-Synthesis	0.58	0.34	M
WJ-R Numbers Reversed	0.55	0.30	M	WJ III Sound Patterns-Music	0.58	0.34	M
WJ-R Visual-Auditory Learning	0.53	0.28	M	WJ III Visual-Auditory Learning	0.57	0.32	M
WJ-R Memory for Words	0.48	0.23	L	WJ III Cross Out	0.55	0.30	L
WISC-R Picture Completion	**0.47**	**0.22**	**L**	**WISC-III Digit Span**	**0.55**	**0.30**	**L**
WJ-R Visual Matching	0.47	0.22	L	WJ III Auditory Working Memory	0.54	0.29	L
WJ-R Cross Out	0.41	0.17	L	WJ III Visual Matching	0.53	0.28	L
WJ-R Incomplete Words	0.41	0.17	L	WJ III Math Fluency	0.53	0.28	L
WJ-R Memory for Names	0.40	0.16	L	WJ III Numbers Reversed	0.52	0.27	L
WJ-R Sound Blending	0.39	0.15	L	**WISC-III Symbol Search**	**0.51**	**0.26**	**L**
WJ-R Picture Recognition	0.33	0.11	L	**WISC-III Comprehension**	**0.49**	**0.24**	**L**
WISC-R Picture Arrangement	**0.28**	**0.08**	**L**	WJ III Memory for Words	0.48	0.23	L
WJ-R Visual Closure	0.27	0.07	L	WJ III Pair Cancellation	0.48	0.23	L
WISC-R Coding	**0.05**	**0.00**	**L**	WJ III Memory for Sentences	0.47	0.22	L
				WJ III Memory for Names	0.46	0.21	L
				WISC-III Object Assembly	**0.44**	**0.19**	**L**
				WJ III Sound Patterns-Voice	0.44	0.19	L
				WJ III Decision Speed	0.43	0.18	L
				WJ III Planning	0.41	0.17	L
				WISC-III Coding	**0.41**	**0.17**	**L**
				WJ III Oral Comprehension	0.39	0.15	L

(continued)

Table 2. (continued)

1c. WJ/WAIS, $n = 78$			g	1d. WJ/WISC-R, $n = 167$			g
Grade 12 (1978)	g	h^2	CIs	Grades 3 and 5 (1978)	g	h^2	CIs
				WJ III Spatial Relations	0.37	0.14	L
				WISC-III Picture Completion	**0.36**	**0.13**	**L**
				WJ III Block Rotation	0.36	0.13	L
				WJ III Sound Blending	0.33	0.11	L
				WJ III Incomplete Words	0.31	0.10	L
				WJ III Auditory Attention	0.30	0.09	L
				WJ III Retrieval Fluency	0.23	0.05	L
				WJ III Picture Recognition	0.14	0.02	L
				WJ III Visual Closure	0.03	0.00	L

1g. WJ III/KAIT/WAIS-III, ($n = 149$)			g
University (2001)	g	h^2	CIs
WJ III Verbal Comprehension	0.76	0.58	H
WAIS-III Vocabulary	**0.72**	**0.51**	**H**
KAIT Definitions	0.70	0.49	H
KAIT Double Meanings	0.69	0.48	H
WJ III Auditory Working Memory	0.67	0.45	H
WAIS-III Arithmetic	**0.67**	**0.45**	**H**
KAIT Rebus Learning	0.65	0.42	H
KAIT Logical Steps	0.65	0.42	H
WAIS-III Information	**0.62**	**0.39**	**M**
WJ III Concept Formation	0.62	0.39	M
WAIS-III Matrix Reasoning	**0.62**	**0.38**	**M**
KAIT Auditory Comprehension	0.61	0.37	M
KAIT Mystery Codes	0.60	0.36	M
WAIS-III Block Design	**0.59**	**0.35**	**M**
WAIS-III Letter-Number Sequencing	**0.58**	**0.34**	**M**
WJ III Story Recall	0.58	0.34	M
WJ III Visual-Auditory Learning	0.57	0.33	M
WAIS-III Comprehension	**0.54**	**0.30**	**M**
WJ III Blending	0.53	0.28	M
WJ III Oral Comprehension	0.53	0.28	M
WJ III Spatial Relations	0.51	0.26	M

(continued)

Table 2. (continued)

1g. WJ III/KAIT/WAIS-III, (n = 149)			
			g
University (2001)	*g*	h^2	Cls
WJ III Numbers Reversed	0.49	0.24	M
WJ III Analysis-Synthesis	0.48	0.23	M
WAIS-III Symbol Search	**0.48**	**0.23**	**M**
WJ III Decision Speed	0.47	0.22	M
WAIS-III Similarities	***0.46***	***0.21***	***M***
WJ III Visual Matching	0.45	0.20	M
WAIS-III Digit Symbol/Coding	**0.41**	**0.16**	**L**
WJ III Retrieval Fluency	0.40	0.16	L
WJ III Memory for Words	0.37	0.14	L
WJ III Picture Recognition	0.36	0.13	L
WJ III Incomplete Words	0.36	0.13	L
WJ III Math Fluency	0.36	0.13	L
WAIS-III Picture Arrangement	**0.34**	**0.12**	**L**
WAIS-III Picture Completion	**0.31**	**0.10**	**L**
WJ III Auditory Attention	0.25	0.06	L
WJ III Rapid Picture Naming	0.24	0.06	L

Note: Descriptions of samples and measures for Samples 1c to 1g can be found in Woodcock (1978), McGrew, Werder, and Woodcock (1991), and McGrew and Woodcock (2001). Sample 1b information in Marshelek, Lohman, and Snow (1983). Information from Sample 1a (data set from Carroll, 1993) from unpublished doctoral dissertation (unable to secure copy). *g*-Loadings based on first unrotated component in principal component analyses of each sample's correlation matrix. Since the magnitude of *g*-loadings is influenced by variability within samples, an absolute level of *g*-loading could not be used to classify tests across samples. Instead, in each sample the highest h^2 (% of shared variance on first *g*-component factor) was identified. A point 15% lower was established as the lower limit of the H (high) *g* category. However, if a more "natural break" was near (in the direction of a more liberal cut point) it was used. For example, in Sample 1e the top h^2 was 69%; 15% lower was 57%. As noted above, WJ Antonyms-Synonyms had an h^2% of 56, 1% below the arbitrary cut-point, but was followed next by WAIS Arithmetic at 46% (10% less than Antonyms-Synonyms). Thus, the H category cut-point was moved to 56%; 20% was then subtracted from first test in the M (medium) category to define M/L (low), but was similarly modified (within each sample) to reflect similar natural breaks in the h^2 values. Bold font = Wechsler tests. Bold/italics = Similarities test.

and Similarities' (.56) *g*-loadings. When converted to the amount of shared variance with the underlying *g*-component (h^2), which theory and research suggests Gf indicators should possess in high quantities, Raven's (60.8%) has approximately *twice* the *g*-variance as Similarities (31.4%)! Also striking is the very low *g*-loading for Picture Arrangement (.37), a finding in direct contradiction with Flynn, Kaufman, and Zhou et al.'s interpretation of Picture Arrangement as a measure of Gf-like novel abstract reasoning.

The remaining *g*-loading analyses included in Table 1 (1c to 1g) include external validity test indicators from at least one edition of the Woodcock–Johnson battery (1977, 1989,

2001).[9] Again, Vocabulary, Information, and Arithmetic display the highest *g*-loadings within each respective Wechsler battery. An inconsistent pattern of Similarities's *g*-loadings is observed. Similarities is a *high* *g*-loading test in only one sample (1e; the elementary grade WISC-R/WJ-R sample of subjects). Similarities is classified a *medium* measure of *g* in the other two elementary grade school samples (1d, 1f). In samples with older subjects, the WAIS-III Similarities test is classified a *low* indicator of *g* (1c; Grade 12 sample) or *medium* in the university sample where it is surrounded by four processing speed (Gs) tests (i.e., WAIS-III Symbol Search and Digit Symbol-Coding; WJ III Decision Speed and Visual Matching).

A potentially interesting observation is a difference in Similarities' *g*-loadings between the young school-age WISC/WISC-R/WISC-III samples (1a, 1d-f) and the older adolescent and adult WAIS/WAIS-III samples (1b, 1c, 1g). Similarities' relative *g*-loading is in the *medium* to *high* classifications in the young samples but is consistently either *low* or a weak *medium* classification in the adolescent and adult samples (see Table 1). This suggests that the level of complex abstract reasoning measured by Similarities may differ across the childhood and adult Wechsler batteries. If true, this would suggest significant methodological problems when comparing Similarities scores in FE research across the different age-related Wechsler batteries. A possible explanation is that younger children need to pause and engage in more deliberate, voluntary, effortful "thinking" about abstract categories of similarity, whereas with increasing age-related learning and experience, this type of conceptual thinking "can reflect well-automatized conventions" (Sattler, 2001, p. 419), thus requiring less complex cognitive processing resources.

One limitation of the current analyses is that it is based on the unproven Flynn assumption that the various versions of the Similarities subtest (across editions) measure the same ability construct with each new revision. Kaufman (2010) presents convincing task analysis that the major changes in test administration and scoring between the 1949 WISC and 1979 WISC-R may have resulted in significant changes in the underlying construct measured by these two versions of the Similarities test. Given the lack of appropriately designed Similarities scale equating studies across the childhood and adult Wechsler batteries used in FE research, it is possible that the reason for the varying Similarities *g*-loadings in Table 2 is that the similarly named Similarities' subtests (across editions) are measuring different ability constructs.[10] In other words, there is no known answer to the question(s) "In what way is the ability measured by the WISC Similarities subtest similar to the ability measured by the WISC-R Similarities subtest?"; "In what way is the ability measured by the WISC-R Similarities subtest the same as the ability measured by the WISC-III Similarities subtest?"; and so on.

With only one exception (WISC/ITPA sample; 1a), Picture Arrangement consistently is in the basement of the *low* *g*-loading category (one of the five lowest *g*-loading tests in all analyses; see Table 2). This finding indicates that Picture Arrangement should not be considered an indicator of high level Gf-like abstract reasoning or on-the-spot-problem-solving. Picture Arrangement appears completely corroded by rust as a FE anchor.

The above joint test battery analyses indicate that the assumption of Flynn (Flynn & Weiss, 2007), as well as many others, that (with the exception of Coding) all Wechsler subtests demonstrate *very similar* g-*loadings*, is incorrect. A similar inaccurate statement is that "both Arithmetic and Similarities have *very high* *g*-loadings" (italics added; Flynn, 2007, p. 219). In the samples summarized in Table 2, the Wechsler subtests disperse widely across the high, medium, and low *g*-ness classification continuum. I conclude that many FE interpretations, conclusions, hypotheses, and theorizations that repeatedly allude to Wechsler subtest *g*-loadings are based on incorrect interpretations of multiple Wechsler-based FE cornerstones. The parochial assumption that Wechsler within-battery internal factor *g*-loadings generalize to the more complete domain of

Table 3. *g*-Loading and *g*-Variance Comparisons of Wechsler Similarities Test to Other Gc "Reasoning" Tests Across Five Samples

Sample/tests	*g*	h^2	Other Gc "Reasoning" Test/ Wechsler Similarities *g*-Loading Variance Difference
1c. WJ/WAIS (n = 78)			
WJ Analogies	0.82	0.67	45%
WJ Antonyms-Synonyms	0.75	0.56	34%
WAIS Similarities	**0.47**	**0.22**	
1d. WJ/WISC-R (n = 167)			
WJ Antonyms-Synonyms	0.83	0.69	33%
WJ Analogies	0.77	0.59	23%
WISC-R Similarities	**0.60**	**0.36**	
1e. WJ-R/WISC-R (n = 72)			
WJ-R Oral Vocabulary	0.83	0.69	15%
WISC-R Similarities	**0.75**	**0.57**	
WJ-R Analogies	0.74	0.54	–3%
1f. WJ III/WISC-III (n = 148)			
WJ III Verbal Comprehension	0.76	0.58	10%
WISC-III Similarities	**0.69**	**0.48**	
1g. WJ III/KAIT/WAIS-III (n = 149)			
WJ III Verbal Comprehension	0.76	0.58	37%
KAIT Definitions	0.70	0.49	28%
KAIT Double Meanings	0.69	0.48	27%
WAIS-III Similarities	**0.46**	**0.21**	
Average *g*-variance h^2 difference			Mean = 22.6%; Median = 27%

Note: Information was extracted from Table 1. h^2 is the *g*-loading squared and represents the amount of variance each test shares with the underlying *g*-component extracted via principal components analysis. Wechsler Similarities is in bold font.

human intelligence (see Carroll, 1993; McGrew, 2009) represents a serious oxidation of numerous Wechsler-based FE linchpins.

Wechsler Similarities: Comparison to Other Gc "Reasoning" Tests

My responses to this point might suggest that I believe that Gc tests like Similarities do not measure abstract problem solving or reasoning. This is not what I believe. I agree with Hunt (2000) who indicates that Gc includes the ability to "*apply culturally approved, previously acquired problem solving methods*" (italics added; p. 127). This differs from Gf, where problem solving is applied to *novel* tasks and situations influenced much less by culture. My central point is that Flynn (and others) tend to ignore this distinction and mistakenly suggest the problem solving and reasoning measured by Similarities and Ravens are the same, or very similar. *Similarities* (*and some other Gc tests*) *do measure some form of culturally approved problem solving or reasoning*. But even within the Gc domain Similarities is not a top notch indicator of this form of reasoning. To add an empirical perspective, a comparison of the complexity of Gc reasoning demanded by Similarities with other Gc reasoning tests is presented in Table 3.[11]

The WJ/WJ-R Analogies test is the classic "X is to Y as Z is to _____" test. The WJ/WJ-R Antonyms and Synonyms subtests (later renamed Oral Vocabulary in the WJ-R) require subjects to solve verbal items of the form: "Tell me a word that means the same as_____;" "Tell me

a word that means the opposite of _____." Finally, the WJ III Verbal Comprehension test represents the combination of Antonyms, Synonyms, Analogies, and Picture Vocabulary. The KAIT Definitions test requires examinees to figure out a word by studying the word that is presented with missing letters while hearing or reading a clue about the word's meaning. KAIT Double Meanings presents the examinee with two different sets of word clues, and the examinee must think of a single word with two meanings that relates closely to both sets of clues. It is important to note that WJ III and KAIT Gc tests only require an examinee to provide single word responses, reducing the subjectivity of scoring and eliminating the possibility that examinees may eventually secure a correct answer if they talk long enough. Verbosity is not rewarded.

As can be seen in Table 3, across five samples Similarities' *g*-loading variance (h^2) generally ranges from 23% to 45% less than the KAIT and WJ III Gc tests. On average, Similarities displays approximately ¼ less ($M = 22.6\%$; $Mdn = 27\%$) *g*-variance than the analogous KAIT and WJ III Gc reasoning tests, tests which also do not load on Gf in CHC joint-battery factor studies. These Gc test *g*-loading comparative analyses suggest, together with the extant CHC-organized joint factor research and the *g*-loading analysis presented previously, that *Wechsler Similarities is a relatively poor proxy for higher-level abstract reasoning* within the domain of Gc *and is even a poorer indicator of the on-the-spot problem solving that characterizes the abilities* within the domain Gf *(as traditionally and historically defined)*.

A possible explanation for the less than stellar Similarities within-Gc "reasoning" performance may be found in Cummins's (1979) distinction between Basic Interpersonal Communication Skills (*BICS*) and Cognitive Academic Language Proficiency (*CALP*). BICS is social language proficiency needed in everyday language contexts, such as conversational oral language. CALP is formal academic language proficiency needed in context-reduced settings, such as school. CALP is more cognitively demanding. Anyone familiar with Similarities knows that the test allows examinees to provide lengthy verbal responses in their own everyday language. In contrast, performance on the KAIT and WJ/WJ-R/WJ III Gc tests require one-word responses that focus more on cognitive processing involving language (e.g., antonyms, synonyms, verbal analogies). These findings suggest that Similarities may be less cognitively demanding (more BICS) than the comparable Gc scales from the WJ III and KAIT (more CALP) and less cognitively demanding than assumed by Flynn (Flynn, 2007; Flynn & Weiss, 2007) and Kaufman (2010).

If the Data Don't Fit (the FE): You May Need to Retrofit (the FE)

As noted by Zhou et al. (2010) and Weiss (2007), in 2006 Flynn suggested problems with the WAIS-III standardization norms given that studies comparing the WAIS-R/WAIS-III scores were not consistent with FE expectations. According to Zhou et al. and Weiss (2007), Flynn is ignoring data that do not fit his theory and instead is using theory to question data (and the integrity of the WAIS-III test batteries norms). According to Weiss (2007),

> The only evidence Flynn provides for this statement is that WAIS-III scores do not fit expectations made based on the FE. However, the progress of science demands that theories be modified based on new data. Adjusting data to fit theory is an inappropriate scientific method, regardless of how well supported the theory may have been in previous studies. (p. 1)

Three years later, Flynn again discounted a new set of data points inconsistent with FE expectations. Flynn (2009) stated:

> Just as I was about to exonerate the WAIS-III from the charge that its standardization sample was substandard, I received a copy of Floyd, Clark and Shadish (2008). A group of 148 college undergraduates scored 8.64 points higher (adjusted for dates of standardization) on the WAIS-III than on the Woodcock-Johnson III, and a group of 99 subjects scored 6.77 points higher (adjusted) on the Kaufman Adolescent and Adult Intelligence Scale. These results are very unsettling because cases are heard where WAIS-III IQs are on record. I strongly recommend simply setting the WAIS-III scores aside. (p. 103)

I concur with Weiss and Zhou et al. that the finding of new data (often from more contemporary versions of the Wechslers or other newer IQ batteries) inconsistent with FE expectations (and FE adjustments to the data) is insufficient evidence to discredit the norms of an intelligence battery or to make such strongly worded statements or pronouncements (e.g., "just as I was about to exonerate"; "results are unsettling"; "I strongly recommend"). Is this apparent "closed" FE system an indicator of the rigidity of Flynn's defense of the FE due to the multiple rusted FE linchpins discussed here? For the sake of scientific progress, I hope not. These "anomalous" (as per FE expectations) findings may unsettle Flynn, but I find them potentially important and requiring study and explanation. If the data don't fit, one may need to retrofit (the theory or hypothesis). "The purpose of models is not to fit the data but to sharpen the questions" (Samuel Karlin, 11th R. A. Fisher Memorial Lecture, Royal Society, April 20, 1983).

"New Versus Old Rules of Measurement" and the FE: Recommendations

The two themes of my response are that a significant portion of the Wechsler-based FE research is compromised by (a) the lack of empirical evidence for subtest or composite score evidence across key Wechsler FE linchpins and (b) the incorrect interpretation of key Wechsler measures as strong indicators of *g* or Gf—a crucial cornerstone of interpretation of the FE. Kaufman's (2010) apples-and-oranges analogy, albeit in the form of clinical and logical task-analysis-based arguments, reinforces the problem of unknown scale equivalence, particularly for the Similarities test. I applaud Zhou et al.'s (2010) attempt to grasp the slippery and complicated issue of measurement equivalence. We need more FE research that first establishes construct scale equivalence before interpreting average score changes across time. Unfortunately, the confidence placed in Zhou et al.'s admittedly tentative conclusions is weak given the variety of methodological complications (i.e., possible regression effects in the ANCOVA, inherent problems in classical test theory-based equal percentile equating [need for large samples with similar underlying ability distributions], comparison of Performance composites with different mixtures of CHC abilities). Zhou et al.'s inconsistent and contradictory results (across and within methodologies) are difficult to explain; and I believe these results render their IQ Ability Level × FE interaction findings inconclusive. Kaufman and Zhou et al.'s primary contribution is placing the critical issue of measure scale equivalence back on the FE research radar screen. A variety of statistical methods (linking, equating, and calibrating) are available for comparing scores from measures of the same underlying construct that do not contain the same exact set of items (Dorans, Pommerich, & Holland, 2007; Livingston, 2004).

I believe that a reasonably clear understanding of the extent of the FE and its causal mechanisms will emerge only when more FE researchers "freshen" their psychometric and theoretical normative knowledge in a manner similar to test publishers freshening (updating) IQ test norms. The extant psychometric intelligence research has improved the collective "normative" corpus of knowledge regarding our understanding, measurement, and statistical modeling of human cognitive abilities. Interpretation of FE research based on dated conceptualizations of human intelligence

and old statistical and measurement tools is analogous to using the WAIS-R FS IQ to estimate a person's level of intellectual functioning today—it will be of questionable accuracy and validity.

The "new rules of measurement" (Embretson, 1996; Embretson & Hershberger, 1999) item response theory (IRT) test development approach has led to improved and more flexible linking and equating techniques (see Beaujean & Osterlind, 2008, for a concise summary of the advantages of IRT over CTT in FE research). As summarized by Beaujean and Osterlind (2008),

> Unless it can be shown that the measures are invariant between the groups, CTT models are not sensitive to such sources of variance and can show a difference in true scores, even if there is no change in the underlying latent variable . . . IRT, on the other hand, can assess these properties for items on a given test, which, theoretically, allows the researcher to discern a difference between a true rise in intelligence (measured via a latent construct), changing item properties, or an interaction of the two. (p. 456)

Furthermore,

> if intelligence is actually rising, then the individuals who took the test at different time points can be placed on the same underlying θ (ability) distribution, which makes ability comparisons especially easy, as one can determine how many standard deviations one group's (average) cognitive ability is from another's. (p. 456)

The "new rules of measurement" generation of FE research has suggested that the extant FE research may be significantly compromised by the lack of established construct scale measurement equivalence and incorrect interpretations of various Wechsler measures. A good example is Beaujean and Osterlind's (2008) methodological investigation of the FE that used both CTT- and IRT-based scores from the Peabody Picture Vocabulary Test–Revised and Peabody Individual Achievement Test–Math embedded in the *National Longitudinal Study of Youth 79 Children and Young Adults*. These researchers found that when using raw or standardized scores, an increase in scores of the expected FE magnitude was evident. However, when using IRT-based scores, the magnitude of the FE decreased substantially, and in the case of the Peabody Picture Vocabulary Test–Revised, disappeared completely.

Wicherts and colleagues have examined the FE across a variety of samples and cognitive batteries often using IRT-based scales. In addition, they have used contemporary structural equation modeling based multiple group confirmatory factor analysis (MGCFA) to test the assumption of measurement invariance of different tests across cohorts (Wicherts et al., 2004). Wicherts et al. (2004) reported a lack of measurement invariance at the subtest level, a finding that raises concerns with Flynn's analyses and theorization based on Wechsler subtest scores. Wicherts recently combined the psychometric advantages of MGCFA- and IRT-equated test indicators across cohorts to disentangle the complex nuances and interactions of possible FE measurement artifacts and changes in latent constructs (at different levels of generality—specific, broad, and *g*) across the three editions of the WJ battery (1977 WJ, 1989 WJ-R, 2001 WJ III).[12] Wicherts presented his preliminary findings across all three WJ editions at an ISIR conference in 2005.[13] Wicherts most recently analyzed scores from 15 cognitive tests common to both the WJ-R and WJ III (which had been placed on the *same underlying IRT-based scale* via equating of common items across the same tests) that represented six broad CHC ability domains (Gsm, Gc, Ga, Gf, Gs, Grw) and general intelligence (*g*). According to Wicherts, "The picture emerging from our study is that the FE on the WJ is a rather heterogeneous phenomenon, as it depends on (1) the age-cohort, (2) the subtests, (3) and the type of broad abilities at hand" (personal communication, February 10, 2010).

The Beaujean and Osterlind and Wicherts research are illustrative examples of the contemporary IRT-based FE research recommended by Rodgers (1999). More important, the results from these contemporary programs of FE research reinforce Zhou et al.'s (2010) conclusion that the "FE is much more complicated phenomenon than a simple overall increase in IQs" (p. 409). I am particularly keen on Zhou et al.'s call for research "to go deeper into history—to include data from earlier versions of Wechsler, such as WPPSI, WISC-R, and the WAIS." I would further recommend that attempts to link historical and contemporary datasets apply IRT scaling methods across common Wechsler item sets (similar to the WJ data accessed by Wicherts) before cranking up the statistical comparison machinery and theorization.

Because the FE and its implications are no longer exclusive topics for intellectual debates among scholars and are having impact on critical life decisions (e.g., Atkins ID/MR death penalty cases, social security benefit eligibility), significant attention and resources should be pooled to form collaborative work groups that would specify standards and state-of-the-art methods for FE research and complete a systematic program of FE research. Yes, this would be time consuming. Yes, this would be expensive. Yes, I may be naive. However, no single researcher, publisher, or university-based group can tackle such a complex set of issues and massive quantities of data. Funding might be secured from organizations like the National Science Foundation or other private foundations. Resources and manpower could come from collaboration between publishers of the major intelligence test batteries and professional associations (ISIR, APA, AERA) as well as comparable international organizations. Furthermore, consistent with Kaufman's (2010) suggestion, I would recommend that attention also be focused on contemporary intelligence batteries that include a wider array of CHC construct indicators and that have undergone two or more revisions (e.g., DAS/DAS-II, K-ABC/KABC-II, SB-IV/SB-V, WJ/WJ-R/WJ-III).

Concluding Comments

The extant Wechsler FE data and its system of interpretations, hypotheses, and resultant theory are held together by multiple linchpins. These anchors need to be strong, shiny, and flexible enough to account for potentially new and anomalous research findings. A complex structure or framework such as the FE and FE theorization, which is currently held together at pivotal points by multiple rusted linchpins, runs the risk of becoming rigid and resistant to revision. It has been more than a decade since Rodgers (1999) provided one of the best point-by-point critiques and recommendations for improving FE research, yet implementation of his suggestions has been sparse. The theory, tools, and technology exist to place some reasonable degree of order in the house built by Flynn. Remodeling is in order.

So, what do I believe about the FE? As a coauthor of a major IQ and achievement battery (WJ III), who was intimately involved in the statistical analysis and calculation of norms for the WJ-R and WJ III revisions, I have stared the reality of parts of the FE (on test norms) in the face on my computer screen. My most recent encounter was during the calculation of the WJ III NU (normative update), where the WJ III 2001 norms, originally calculated based on 1996 U.S. Census *projections* for year 2000, were recalculated with the *final* 2000 census *statistics* (see McGrew, Dailey, & Schrank, 2007, for details), which resulted in a downward shift in global IQ scores of approximately 2.8 standard score points.[14,15] The WJ III-WJ III/NU subjects were identical. The tests were identical. The subjects' raw scores were identical. Only the U.S. Census subject weights applied to the norm data changed. A fundamental cause of the 2.8 average IQ score drop was the simple recalculation and application of new (NU) subject weights to more accurately reflect the final year 2000 population.[16] Clearly, shifts in the composition of the population result in shifts in IQ test norms. However, I agree with Rodgers that multiple methodological

issues, some which were addressed in my response, do not allow for a clear picture of the extent of the FE and has led to overconfidence and possible overstatement of the magnitude of the FE.

I believe we will eventually learn that the FE is a differential CHC-ability effect. This belief is consistent with Flynn's recent focus on scores beyond composite full scale scores, although as I have articulated in my response, I believe his differential-ability analysis suffers from significant methodological and measurement flaws. Thus, the meaning of reported global composite IQ FE findings may be of questionable value. When investigations are completed with newer methods, measures, and the application of newer methods to older measures, I predict that the global composite FE findings will be found to have masked differential CHC-ability changes across generations. However, as per the logic, data, and arguments I have presented here, I believe we are not yet at the place where the CHC-based FE research knowledge is sufficient to suggest differential FE modification of different IQ subcomponent scores in any context. Furthermore, as outlined in my response, I believe there are too many rust-coated FE linchpins to allow confidence to be placed in the limited studies that have suggested a differential FE score adjustment by levels of intelligence (e.g., as in Atkins ID/MR death penalty cases). As articulated in the lyrics from one of my favorite songs from the late 1960s (*For What It's Worth*; Buffalo Springfield), *There's something happening here, What it is ain't exactly clear*.

Finally, my critique may sound unduly harsh, critical, and at times naive (e.g., the collaborative work group recommendation). Critics always have 20/20 hindsight. But this does not minimize the professional respect I hold for James Flynn (as well as Alan Kaufman). Flynn has devoted most of his career to synthesizing mass quantities of FE intelligence research, debating and defending his findings and hypothesis, writing extensively, and weaving (at least on the surface) an internally consistent organized FE framework. As noted by Rodgers (1999), "given the complexity of the task he undertook, it is not surprising that some important points were treated quite well, while others were treated poorly or not at all" (p. 199). "It is the lone worker who makes the first advance in a subject: the details may be worked out by a team, but the prime idea is due to the enterpriser thought and perception of an individual" (Sir Alexander Fleming, in a speech at Edinburgh University, 1951).

Acknowledgement

I would like to thank Alex Beaujean, Beth Hope, Barbara Wendling, and Karen Salekin for their thoughtful insights, comments, and edits to early versions of this article.

Declaration of Conflicting Interests

The author(s) declared a potential conflict of interest (e.g., a financial relationship with the commercial organizations or products discussed in this article) as follows:

The author is a coauthor of, and thus has a royalty financial interest in, the Woodcock–Johnson III battery, a battery mentioned in this article. Any statements or positions included in this article do not necessarily reflect those of the other WJ III coauthors, the Woodcock-Muñoz Foundation, or Riverside Publishing.

Funding

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

Notes

1. Kaufman criticizes Flynn's failure to recognize, or even mention, intelligence tests grounded in the consensus contemporary psychometric Cattell–Horn–Carroll (CHC) taxonomy of human cognitive abilities (see Kaufman, 2010; McGrew, 1997, 2005, 2009; Roberts et al., 2005; Stankov, 2005). I agree with Kaufman's criticism.

2. Carroll (1993) acknowledged that factor loadings depend on what variables are included in the battery that is analyzed.
3. These data do not necessarily indicate that the Similarities test does not measure some form of abstract problem solving or reasoning. I clarify this point later in the article.
4. Kaufman inadvertently erred in his article by not addressing the extant CHC factor analysis research that suggests that the reasoning measured by Similarities is significantly different than that represented by fluid intelligence (Gf) and traditional Gf tests (e.g., Ravens). Kaufman's actual position is consistent with my position (i.e., Gc tests do not measure traditional Gf but can be designed to measure some form of culturally approved and acquired problem solving), which I discuss later in the article (Kaufman, personal communication, February 25 and March 8, 2010).
5. See Wilhelm (2005) for a thorough discussion of the cognitive processes involved in reasoning, the role of reasoning in different models of intelligence, methods for empirically classifying reasoning tests, and the relation of reasoning with other constructs (e.g., working memory).
6. Table 1 was constructed from information presented by McGrew (2010).
7. It is important to note that my use of critical phrases such as "corroded FE linchpin indicators" is not a generalized statement regarding the psychometric characteristics or utility of any Wechsler subtest in question. It is the use of certain Wechsler subtests or indexes as key FE linchpins that is being criticized.
8. The two historical data sets were obtained from the Woodcock-Muñoz Foundation (WMF) Human Cognitive Abilities Archive that can be accessed at the WMF webpage (http://www.woodcock-munoz-foundation.org/research/HCAProject.html).
9. The decision to use data sets that included markers from one of the editions of the WJ was due to: (a) my obvious easy access to the data files, (b) the fact that all three editions of the WJ include a greater diversity of cognitive domain indicators than most other intelligence tests, (c) the advantage of having some common test indicators across samples, and (d) given the crucial importance of Gf abilities in the analysis, the conclusion that

> The WJ III may actually oversample fluid (rather than crystallized) intelligence concepts, given that some of its supposed markers of SAR (e.g., Auditory Working Memory) and, arguably TSR [Glr] are thought to be highly related to reasoning at the first order and fluid intelligence at the second order of broad cognitive abilities. (Roberts et al., p. 344)

10. A possible other reason might be the lack of comparability of the different samples in Table 2, although the relative pattern for most of the Wechsler and WJ/WJ-R/WJ III tests appear relatively stable, a finding that suggests the relative magnitude of test *g*-loadings is not due to major sample differences.
11. Information was extracted from Table 2.
12. The Woodcock-Muñoz Foundation provided Wicherts data files where tests with common items across all three editions of the WJ battery were equated to a common IRT scale based on the common items across editions.
13. Wicherts's paper presentation earned him the ISIR 2005 Templeton Prize for Best Student Paper.
14. A copy of McGrew et al. (2007) can be viewed and downloaded at: http://www.iapsych.com/articles/asb9.pdf
15. The average 2.8 WJ III-to-WJ III NU IQ score change is my calculation of the mean General Intellectual Ability–Standard cluster standard scores based on the original WJ III 2001 and WJ III NU norms on the same norm subjects.
16. The WJ III/WJ III NU score differences may also be attributed to the use of new bootstrap resampling procedures in the calculation of the WJ III NU norms. However, these new procedures had the greatest impact at the very young and old ages (see McGrew et al., 2007, for details), which were excluded from the WJ III-to-WJ III NU score drop reported here for the first time. The mean score change reported here is for the WJ III-WJ III/NU General Intellectual Ability–Standard cluster only for subjects between the ages of 6 and 40 years.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

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.

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

Cummins, J. (1979). Cognitive/academic language proficiency, linguistic interdependence, the optimum age question and some other matters. *Working Papers on Bilingualism*, *19*, 121-129.

Dorans, N. J., Pommerich, M., & Holland, P. W. (2007). *Linking and aligning scores and scales.* New York, NY: Springer.

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment*, *8*, 341-349.

Embretson, S. E., & Hershberger, S. L. (1999). *The new rules of measurement*. Mahwah, NJ: Lawrence Erlbaum.

Flanagan, D. P., McGrew, K. S., & Ortiz, S. O. (2000). *The Wechsler Intelligence Scales and Gf-Gc theory: A contemporary approach to interpretation*. Boston, MA: Allyn & Bacon.

Flanagan, D. P., Ortiz, S. O., & Alfonso, V. C. (2007). *Essentials of cross-battery assessment* (2nd ed.). Hoboken, NJ: Wiley.

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

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

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.

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.

Hunt, E. (2000). Let's hear it for crystallized intelligence. *Learning and Individual Differences*, *12*, 123-129.

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

Kaufman, A. S. (2009). *IQ testing 101.* New York, NY: Springer.

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.

Livingston, S. (2004). *Equating test scores (without IRT).* Princeton, NJ: Educational Testing Service.

Marshalek, B., Lohman, D. F., & Snow, R. E. (1983). The complexity continuum in the Radex and hierarchical models of intelligence. *Intelligence*, *7*, 107-127.

McGrew, K. S. (1997). Analysis of the major intelligence batteries according to a proposed comprehensive Gf-Gc framework. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 151-179). New York, NY: Guilford Press.

McGrew, K. S. (2005). The Cattell-Horn-Carroll theory of cognitive abilities. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 136-181). New York, NY: Guilford Press.

McGrew, K. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research [Editorial]. *Intelligence*, *37*, 1-10.

McGrew, K. (2010, February). *AP101 Brief # 6b supplement: Summaries of Wechsler CHC test classifications*. Retrieved from http://www.iqscorner.com/2010/02/ap101-brief-6b-supplement-summaries-of.html

McGrew, K., Dailey, D., & Schrank, F. (2007). *Woodcock-Johnson III/Woodcock-Johnson III Normative Update score differences: What the user can expect and why* (Woodcock-Johnson III Assessment Service Bulletin No. 9). Rolling Meadows, IL: Riverside.

McGrew, K., & Flanagan, D. (1998). *The Intelligence Test Desk Reference (ITDR): Gf-Gc cross-battery assessment*. Boston, MA: Allyn & Bacon.

McGrew, K. S., Werder, J. K., & Woodcock, R. W. (1991). *The WJ-R technical manual.* Rolling Meadows, IL: Riverside.

McGrew, K. S., & Woodcock, R. W. (2001). *Technical manual. Woodcock-Johnson III*. Rolling Meadows, IL: Riverside.

Roberts, R. D., Markham, P. M., Matthews, G., & Zeidner, M. (2005). Assessing intelligence: Past, present, and future. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 333-360). Thousand Oaks, CA: SAGE.

Rodgers, J. L. (1999). A critique of the Flynn Effect: Massive IQ gains, methodological artifacts, or both? *Intelligence*, *26,* 337-356.

Sattler, J. (2001). *Assessment of Children: Cognitive Applications*- 4th edition. San Diego, CA: Jerome M. Sattler, Publisher, Inc.

Snow, R. E., Lohman, D. F., Marshalek, B., Yalow, E., & Webb, N. (1977). *Correlational analyses of reference aptitude constructs* (Technical Report No. 5). Stanford, CA: Aptitude Research Project, School of Education, Stanford University.

Stankov, L. (2000). Structural extensions of a hierarchical view on human cognitive abilities. *Learning and Individual Differences*, *12*, 35-51.

Stankov, L. (2005). *g* factor: Issues of design and interpretation. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 279-293). Thousand Oaks, CA: SAGE.

Wechsler, D. (2004). *WISC-IV integrated technical and interpretation manual*. San Antonio, TX: Pearson.

Wechsler, D. (2008). *WAIS-IV technical and interpretation manual*. San Antonio, TX: Pearson.

Weiss, L. G. (2007). *Response to Flynn. WAIS-III technical report*. San Antonio, TX: Harcourt Assessments.

Wicherts, J. M., Dolan, C. V., Hessen, D. J., Oosterveld, P., vanBaal, 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.

Wilhelm, O. (2005). Measuring reasoning ability. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 373-392). Thousand Oaks, CA: SAGE.

Woodcock, R. W. (1978). *Development and standardization of the Woodcock-Johnson Psycho-Educational Battery.* Hingham, MA: Teaching Resources.

Woodcock, R. W. (1990). Theoretical foundations of the WJ-R measures of cognitive ability. *Journal of Psychoeducational Assessment*, *8*, 231-258.

Woodcock, R. W., & Johnson, M. B. (1977). *Woodcock-Johnson Psycho-Educational Battery*. Rolling Meadows, IL: Riverside.

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

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III*. Rolling Meadows, IL: Riverside.

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*, 399-411.