"In What Way Are Apples and Oranges Alike?" A Critique of Flynn's Interpretation of the Flynn Effect

Alan S. Kaufman

Journal of Psychoeducational Assessment 2010 28: 382 originally published online 16 June 2010
DOI: 10.1177/0734282910373346

Journal of Psychoeducational Assessment
28(5) 382–398

DOI: 10.1177/0734282910373346
http://jpa.sagepub.com

"In What Way Are Apples and Oranges Alike?" A Critique of Flynn's Interpretation of the Flynn Effect

Alan S. Kaufman[1]

Abstract

Flynn wrote a book devoted to the Flynn effect, featuring his theoretical explanation of why the intelligence of worldwide populations has apparently increased from generation to generation. The essence of his theorizing is that because of the societal impact of scientific technology, people of today are much more guided by abstract, rather than concrete, approaches to problem solving. He bases his theory in large part on gains on specific tasks, most notably Raven's matrices, Wechsler Intelligence Scale for Children (WISC) Performance subtests, and, most important, on WISC Similarities. The gains on these separate tasks over more than half a century (1947-2002) are striking. However, Flynn failed to take into account the sweeping changes in test content, administration procedures, and scoring guidelines when the 1974 WISC-R was developed from the 1949 WISC. These substantial changes challenge the meaningfulness of comparing children's performance in 1947 with their performance in 2002 on Similarities and other WISC subtests—and therefore challenge Flynn's explanation of the effect that bears his name.

Keywords

Wechsler scales, IQ tests, Flynn effect, Similarities subtest, Raven's matrices

The Flynn effect is (FE) well known: Children and adults score higher on IQ tests now than they did in previous generations (Flynn, 1984, 2007, 2009b). The rate of increase in the United States has apparently remained a fairly constant 3 points per decade since the 1930s (Flynn, 1984, 1987, 1998; The Psychological Corporation, 2003, 2008), although that magnitude of steady gain pales in comparison with the gains observed in many other developed nations (e.g., 5-7 points in Belgium, The Netherlands, and Japan). The worldwide generational change has been observed most frequently using Wechsler's scales and Raven's (1938, 2000) matrices, but applies as well to data obtained on other tests such as the Stanford–Binet Intelligence Scale (Roid, 2003; Thorndike, Hagen, & Sattler, 1986) and the Kaufman Assessment Battery for Children (K-ABC; Kaufman & Kaufman, 1983, 2004).

[1]Yale University School of Medicine, New Haven, CT, USA

Corresponding Author:
Alan S. Kaufman, 3033 Curie Street, San Diego, CA 92122, USA
Email: alanskaufman@gmail.com

The FE directly affects clinical practice (Kanaya, Scullin, & Ceci, 2003) and has important societal implications (Flynn, 2006). By virtue of the steady gain in mean IQ, the norms for American IQ tests get outdated at the rate of 3 points every 10 years, impelling test publishers to revise IQ tests much more frequently than in times gone by. About 25 years elapsed before the Wechsler Intelligence Scale for Children (WISC; Wechsler, 1949) and the Wechsler Adult Intelligence Scale (WAIS; Wechsler, 1955) were revised; in contrast, the third and fourth editions of these widely used tests were separated by only about a dozen years. Even more dramatically, the FE can determine whether a convicted murderer lives or dies as a result of the Supreme Court decision that prohibits the death penalty for individuals diagnosed with mental retardation (*Atkins v. Virginia*, 2002); indeed, a subsequent California verdict (*People v. Superior Court*, 2005) stipulated that "the Flynn effect must be considered in determining a defendant's IQ" (Flynn, 2007, p. 131). Following Flynn's recommendations, "the formula of deducting 0.30 points per year is making headway at least among defense attorneys" (Flynn, 2007, p. 134).

Data in the WAIS-IV manual support the traditional FE of 3 points per decade (The Psychological Corporation, 2008, Table 5.5), although other recent data suggest that the generational gains in the United States vary considerably with ability level (Zhou, Zhu, & Weiss, 2010). Furthermore, 21st-century data from Norway (Sundet, Barlaug, & Torjussen, 2004) and Denmark (Teasdale & Owen, 2005, 2008) suggest that the FE has ended in those countries and may have "reversed" (i.e., decline in IQ). There is also evidence of a reverse FE in the United States for infants and toddlers (Yang, Zhu, Pinon, & Wilkins, 2006).

Research on the FE has abounded in the past quarter-century, with recent efforts shifting the focus from changes in global IQs to gains (or losses) on specific cognitive tasks (Flynn & Weiss, 2007; Teasdale & Owen, 2008). Accompanying the array of research studies has been the continued search for causality. What variables are most responsible for the FE? Using an airplane metaphor, Zhou and colleagues (Zhou & Zhu, 2007; Zhou et al., 2010) ask the probing question: What do you find when you peek inside the "black box" of the FE? The answers have been diverse, with some researchers attributing the IQ gains to environment in general (e.g., Dickens & Flynn, 2001), to specific aspects of environment such as nutrition (e.g., Colom, Lluis-Font, & Andres-Pueyo, 2005) or education (e.g., Teasdale & Owen, 2005), to genetics (e.g., Rodgers & Wanstrom, 2007), or to statistical artifacts stemming from methodological and psychometric issues (e.g., Beaujean & Osterlind, 2008; Rodgers, 1998). Books have been devoted to the "whys" of the FE (e.g., *The Rising Curve*, edited by Ulric Neisser, 1998), and the most recent book is the most fascinating of all because it was authored by Flynn himself and, therefore, demands serious attention.

Flynn's (2007, 2009b) explanation of the FE was driven to a considerable extent by differential gains on Wechsler's subtests (Flynn & Weiss, 2007). He was clearly impressed by differences among WISC Verbal subtests: "Americans gained 24 points on Similarities between 1947 and 2002 (1.6 *SD*s), 4 points on Vocabulary, and only 2 points on Arithmetic and Information (for an average of 3 points on these three subtests collectively)" (Flynn, 2007, p. 9). (The points he refers to for all subtests are "IQ" points, based on $SD = 15$ instead of the traditional subtest scaled-score points, based on $SD = 3$.) He was also impressed by huge gains on Raven's (1938, 2000) matrices, which he estimates to be 27.5 points, but, again, that estimate is "closely tied to U.S. gains on Similarities" (Flynn, 2007, p. 8).

Ultimately, Flynn developed an explanation for his effect that is deeply rooted in the humongous gain on the WISC Similarities subtest, a task whose *g* loading is comparable with the *g* loadings of Information and Arithmetic, but whose average gains over a half-century's time trumps the gains of these other highly saturated *g* subtests by a factor of 12. Why have Similarities gains been so astounding? According to Flynn (2007), the WISC subtests measure skills "that are functionally independent and responsive to changes in social priorities over time" (p. 10). Furthermore, "The

twentieth century saw some cognitive skills make great gains while others are in the doldrums" (p. 10). Why? To Flynn (2007), the intelligence of our forefathers

> was anchored in everyday reality. We differ from them in that we can use abstractions . . . to attack the formal problems that arise when science liberates thought from concrete situations. Since 1950, we have become more ingenious . . . to solve problems on the spot. (pp. 10-11)

When Flynn (2007) takes a stab at developing a theory of intelligence (variously known as a theory, a pretheory, and the Dickens/Flynn model), he asserts that "the rise of science engendered new habits of mind of enormous potency. It detached logic and the hypothetical from the concrete and today we use them to attack a whole range of new problems" (p. 53).

Flynn may be correct in his suppositions about the radical role that science and technology have played in turning human beings from concrete to abstract thinkers in the past half-century, but how much of his thesis is based on the "sticks-out-like-a-sore-thumb" Similarities gain of 24 points? Too much, I fear, and I believe that much of that alleged gain on Similarities is bogus. In fact, the magnitude of gains on several specific subtests is suspect when comparing data from the 1949 WISC to subsequent versions of the WISC and, similarly, when comparing gains on the 1955 WAIS with later editions of Wechsler's adult scale. I will explain.

Revision of the 1949 WISC

First of all, the revision of the 1949 WISC had two problems that no subsequent revisions had to face: (a) a shift in the age range of the test from 5-15 years to 6-16 years, necessitating substantial adjustments in item difficulty, and even item type, at the extremes of the range (age 5-6 was especially important because of the huge cognitive development that occurs at ages 5 and 5½) and (b) the administration and scoring guidelines in the old WISC were sketchy at best and certainly incomplete in that they provided a bare-bones minimum of instruction to the examiner to guide children properly through the nuances of each subtest (e.g., illustrations of verbal responses to be queried were sparse and inconsistent).[1] Flynn (2007) may have even been unaware of these two problems; at the least, he was probably unaware of the shift in age range, based on his incorrect statement that the WISC "has been administered since 1950 to children ages 6 to 16" (p. 5). Both sets of problems relate to substantive changes from the WISC to WISC-R (concerning item style and administration and scoring) that call into question the tacit assumption that a subtest measures the same skills in the same way on the WISC and the WISC-R simply because its name (e.g., Similarities, Block Design) has not changed.

The administration of Digit Span changed considerably from the WISC to the WISC-R (Wechsler, 1974): "Unlike the procedure used in the 1949 WISC, the examiner now administers *both* trials of each item, *even if the child passes the first trial*" (p. 13). Flynn (2007; Flynn & Weiss, 2007) was undoubtedly aware of that substantial change in administration because he only reports Digit Span gains from 1972 (WISC-R) through 2002 (WISC-IV), not from 1947 to 2002. However, he did not seem to be aware that other WISC subtests changed as much as Digit Span or even more, including the crucial Similarities subtest.

WISC Versus WISC-R Similarities

To some extent, the WISC Similarities' item style differed from that of the WISC-R. The first four WISC Similarities items were "Analogies" (e.g., *Boys grow up to be men and girls to be . . .*). In fact, the final Analogy item that immediately preceded the traditional Similarities item style (*In*

what way are a PLUM and a PEACH alike?) was *A knife and a piece of glass both . . .* Clearly, that item was intended as a transitional item, but it provided a mindset for children to respond with *concrete* (i.e., 1-point) answers. These types of Analogy items were eliminated from the WISC-R because they were only needed at age 5. Instead, three new very easy items of the Plum–Peach variety were added to the beginning of the test, and the first four items (the three new ones plus a modification of an existing item) were scored 0-1 instead of 0-1-2.

This shift in the Similarities subtest from the WISC to the WISC-R affected the difficulty level of items. The exact same item on both versions of the test could be easy or hard based on its placement in the sequence. A good example is Cat–Mouse. On the WISC it was the second item of its type administered whereas on the WISC-R it was Item 7. "Cat–Mouse was a much easier item on the WISC-R because it appeared after the individual had 'practiced' six times with comparable items" (Kaufman, 1990, p. 140). And, as I noted more than a generation ago (Kaufman, 1979), the two item types measure different cognitive abilities: "In Guilford terminology, the analogies require evaluation, whereas the conventional Similarities items demand cognition" (p. 125).

Just as important as the differences in the content of the Similarities subtest are the differences in the administration and scoring of this task on the WISC versus WISC-R:

- On the WISC, the items were introduced with the question, "In what way are a PLUM and a PEACH alike?" However, Boehm (1971) showed that basic concepts such as "alike" are not understood by young or culturally disadvantaged children, so the Similarities directions were changed on the WISC-R to increase the child's understanding of the aim of the task: "In what way are a PLUM and a PEACH alike? How are they the same?"
- The WISC never gave children a clue that abstract answers were worth more than concrete answers. The scoring system awarded 2 points for abstract categories and 1 point for functions or descriptive qualities, but children were never given appropriate feedback to shape their responses from concrete to abstract. That changed on the WISC-R. On Item 5 (which resembles PLUM–PEACH), examiners were instructed to tell children who gave a 1-point response (e.g., "You eat them both"), "That's right. You do eat them both. Also, they both are fruits." Similar feedback was given on Item 6 to help communicate the notion that an abstract, categorical answer is better than a concrete answer.
- The WISC general directions for administering and scoring all subtests said the following about questioning a child's verbal response: "To increase scoring accuracy and for better qualitative understanding of the subject's responses, the examiner may find it necessary to query his [or her] subject on some items" (Wechsler, 1949, p. 18). The manual does not state which Verbal subtests often require querying or what kinds of responses should be queried. In the scoring system for WISC Vocabulary, I counted a total of 16 illustrative responses that were listed with a "(Q)." For WISC Comprehension (the subtest was actually called General Comprehension), only 4 illustrative responses are shown with a "(Q)," and for WISC Similarities not a single illustrative response is followed by a "(Q)." The clear message to WISC examiners was to query Vocabulary items as needed, query Comprehension items occasionally, and do not query Similarities responses. All that changed with the WISC-R. The general directions explained when to query: "If a child's response to a Verbal item is ambiguous or incomplete, the examiner should ask him [or her] to clarify his [or her] answer" (Wechsler, 1974, p. 60). The WISC-R Vocabulary and Comprehension subtests include too many "(Q)"s to count (several illustrative responses per item), and for WISC-R Similarities about 50 illustrative responses are followed by a "(Q)."

Similarities changed so much from the WISC to the WISC-R that it is not possible to interpret gains on this task from 1947 to 1972, yet that is precisely what Flynn (2007) has done. And

Similarities is not the only WISC subtest to be revamped in meaningful ways when the 1949 WISC was revised.

WISC Versus WISC-R Picture Arrangement and Comprehension

Picture Arrangement, like Similarities, included a different item type for the first few items—items that were more like Object Assembly than Picture Arrangement. Items A, B, and C required the child to assemble puzzles of a dog, a mother, and a train. Item D was the first time sequence (story) item, followed by a Demonstration story item ("fight"), and then seven conventional story items. The Demonstration item, which was only for children who failed Item D, was accompanied by verbal directions to explain that the pictures told a story of a fight and that they had to be put in the right order to "make the story right" (Wechsler, 1949, p. 75). However, children merely watched the examiner arrange the pictures in the right order for the Demonstration item; they did not solve the item themselves. And the examiner was not permitted to give feedback to the child on any item. On the WISC-R, the puzzle items were eliminated and a Sample item was demonstrated to all children. Again, children watched the examiner tell the story of the "scale," but they did not solve it by themselves. However, the first four items became two-trial items during which examiners gave the children feedback and explanations each time they failed the first trial. The scoring system allotted 2 points for solving the item correctly on the first trial and 1 point for solving it on the second trial (after watching the examiner demonstrate the correct solution). Thus, the way children earned points on easy Picture Arrangement items was based on solving picture puzzles on the WISC and by solving story items (or by simply imitating the examiner's solution to a story item) on the WISC-R. In addition, points earned at the upper end of the ability spectrum were also different: WISC Picture Arrangement items allotted 3 bonus points per item for quick perfect performance in contrast to 2 bonus points for WISC-R items.

Comprehension is another WISC subtest that changed substantially from the WISC to the WISC-R. An array of WISC Comprehension items (8 out of 14) allotted 2 points for recognizing two different "ideas" and 1 point for recognizing only one idea—but children were never told to give a second response if they gave a single answer and then stopped responding, believing that they had given a satisfactory answer. All that changed on the WISC-R. Whenever children gave a single response and stopped answering the question when two ideas were called for (on 9 of 17 Comprehension items), they were told, "Tell me another reason why . . ." or "Tell me another thing to do . . ." Examiners who routinely administered the 1949 WISC (old-timers like me) remember just how common it was for children and adolescents to give a quick, correct (and often an articulate and high-level) reply to a Comprehension item, and then sit back and wait for the next item (often for reasons more related to their personality than their intellect). The WISC-R gave these individuals the opportunity to improve their scores and not penalize them for brevity of response.

WISC Versus WISC-R Performance Subtests

All the WISC-R Performance subtests that allotted bonus points for quick perfect performance included a modification in the examiner's directions that was analogous to the change in Similarities that gave an example of a 2-point response and the change in Comprehension that asked for a second reason why: Namely, they were told "Work as quickly as you can," "Put them together as quickly as you can," and so forth, to give them the clue that speed counts. On WISC-R Coding A and B the child was told, "Work as quickly as you can without making mistakes." On the WISC, the child was told nothing to suggest that it was important to work quickly on any Performance subtest (and numerous items gave 3 bonus points for solving an item in 1-5 or 1-10 seconds).

WISC-R Block Design included one other change that likely helped children understand what was expected of them. In the old WISC, the blocks included four colors, not just red and white. Also, "Another change was the insertion of a transitional item between the third and fourth designs of the 1949 WISC. This new item (Design 4) shows the child how two adjacent split-color blocks look *before* the black guidelines are removed" (Wechsler, 1974, p. 15). This transitional item facilitated some children's ability to "catch on" to the task—one that depends on learning the task during the early items—instead of discontinuing the subtest prematurely.

For WISC Picture Completion, "A maximum exposure of 15 seconds is allowed for each picture" (Wechsler, 1949, p. 72). That maximum exposure time was increased to 20 seconds on the WISC-R to conform to the time limit used for adults on the WAIS. In addition, on the WISC, children were given guidance if they failed either or both of the first two items. If they pointed to an unessential part of the picture they were told, "Yes, but what is the most important thing missing?" But that is the extent of the help they were given. On the WISC-R, they were also told "show me where you mean" if it was unclear where they pointed, and they were given two additional clues: (a) children who simply named a picture were told, "Yes, but what's *missing*?" and (b) children who named something that was off the card (like "the body of the person" when a man or woman was shown from the neck up) were told, "A part is missing *in* the picture. What is it that is missing?" (Wechsler, 1974, p. 71).

The Test Sequence

On the WISC, all Verbal subtests were administered first, followed by all Performance subtests. On the WISC-R, Verbal and Performance subtests were alternated. As Wechsler (1981) noted, "Experience has shown that varying the tasks in this way often helps maintain the subject's interest" (p. 12). Thompson (1987) and his colleagues (Thompson, Howard, & Anderson, 1986) put Wechsler's hypothesis to an empirical test. They administered Vocabulary and Block Design as a two-subtest short form, as part of a four-subtest short form, and as part of the complete WAIS-R (where they are administered in the middle of the test sequence). They found that maintenance of interest directly affected how well adult patients performed on the same subtests. Those who were administered *only* the two subtests (when they were "fresh") performed better than the patients who were given Vocabulary and Block Design as part of the complete battery (when they may have been tired or bored). Based on this study, it is possible that children who were administered Comprehension on the WISC (second in the test sequence) had an advantage over children who were administered the same subtest on the WISC-R (9th in sequence). Analogously, Similarities was administered fourth on the WAIS but last (11th) on the WAIS-R.

Revision of the 1955 WAIS

The preceding discussion has centered on key changes in the 1974 WISC-R, when the 1949 WISC was revised. However, substantial changes also occurred when the 1955 WAIS was revised and the WAIS-R (Wechsler, 1981) was published. WAIS Similarities and Picture Arrangement did not each comprise two separate item types, as on the WISC, but otherwise the substantial revisions in administration and scoring that were implemented in the WISC-R served as a blueprint for the development of the WAIS-R. In WAIS-R Similarities, adults who give a 1-point response to the first item (e.g., "They both have skins") are told, "Yes they both have skins. Also, they are both fruit." And like the WISC, the WAIS Similarities scoring system included no illustrative responses whatsoever that were followed by a "(Q)"; however, the WAIS-R Similarities scoring system was overhauled such that more than 20 responses needed to be queried.

WAIS Picture Arrangement allotted 3 bonus points for quick perfect performance on each the two hardest items; in contrast, WAIS-R Picture Arrangement awarded raw score points based only

on items solved correctly within the time limit, regardless of the exact response time. WAIS-R Comprehension incorporated the "Tell me another reason why . . ." procedure for the two items that required two ideas for full credit, and its scoring system is filled with "(Q)"s (unlike the WAIS Comprehension system that included only two illustrative responses to be queried). The directions for WAIS-R Block Design and Digit Symbol were modified to tell the person to work quickly (directions for WAIS Object Assembly already included the statement, "Put this together as quickly as you can"). Also, WAIS-R Digit Span required the administration of both trials of an item, even if the first trial was passed, unlike the rules that governed the early editions of Wechsler's scales.

And, as on the WISC, all WAIS Verbal subtests were administered first, followed by all Performance subtests. The WISC-R procedure of alternating Verbal and Performance subtests was adopted for the WAIS-R.

Foreign Versions of the WISC and WAIS

The substantial changes in administration that characterized the revisions of the WISC and WAIS do not only affect the interpretation of gains from one generation to the next in the United States but throughout the world. There are two common methods of adapting an American test for other countries or cultures (van de Vijver, Mylonas, Pavlopoulos, & Georgas, 2003): *applications* (close translations of items) and *adaptations* (changing item content to be more suitable for a particular culture). Adaptations are generally an ideal solution to address key cross-cultural issues. Historically, however, that has not happened, not even as recently as foreign versions of the WISC-III (Wechsler, 1991). As van de Vijver et al. (2003) state in their book devoted to the international use of the WISC-III:

> Aggregated across countries, the vast majority of the items in the WISC-III adaptations described in this book, about 90%, has been closely translated or simply copied (in the case of pictorial stimuli), while a small minority, about 10%, has been adapted. (p. 266)

Along with the actual items, directions for administration and scoring have likewise been directly translated. Hence, the worldwide WISCs and WAISs were virtually identical to their U.S. counterparts, and the same is true for the worldwide WISC-Rs and WAIS-Rs. Therefore, gains in Similarities, Comprehension, Picture Arrangement, and several other subtests are not only suspect in the United States but also in countries such as France, Germany, Austria, and Japan (all of which contributed Wechsler data to Flynn's, 1987, landmark analyses of IQ changes over time in 14 developed nations).

An Early "Flynn-Like" Study of WISC-WISC-R IQ Differences

When my colleague and I designed the first study of WISC/WISC-R differences (Doppelt & Kaufman, 1977), we did not have access to data from a sample of children tested in counterbalanced order on both the WISC and WISC-R (at that time, no such data existed). So we computed IQs based only on the single administration of the WISC-R to the 1972 standardization sample. We limited the analysis to the three Verbal subtests and the three Performance subtests that had changed the least and conducted regression analysis based only on the items that remained identical from WISC to WISC-R: 19 Information items, 8 Arithmetic items, 21 Vocabulary items, 2 Object Assembly puzzles, 8 Mazes, and the entire Coding subtest. We computed IQs separately with the WISC norms and the WISC-R norms, based on this common set of items, and discovered, to our surprise, that the WISC-R norms yielded lower IQs than the WISC norms at ages 6½ to 15½ on Verbal IQ (1.6 points), Performance IQ (6.0 points), and Full Scale IQ (4.1 points).

The lower IQs on the WISC-R meant that the new norms were *steeper* than the old norms—that is to say, that children in the early 1970s performed *better* than their counterparts in the late 1940s.

The better performance by children in the 1970s raised the bar, so to speak. Children tested on both the WISC and WISC-R at the same point in time (in counterbalanced order) earned higher scores on the WISC (because its norms were out of date) than on the WISC-R (with its new norms). Though the finding seemed counterintuitive at the time (it is now axiomatic because of the worldwide popularity of the FE), the lower IQs on the WISC-R than WISC reflected *gains* in children's abilities.

In that initial WISC/WISC-R study, we observed IQ gains of less than 1 point per decade on the Verbal Scale, 2.45 points on the Performance Scale, and 1.67 points on the Full Scale. We found age differences (2.37 Full Scale IQ points per decade gain at ages 6½-10½ vs. 0.94 points at ages 11½-15½), but we were clueless about the impact of our findings.

So even when the subtests that changed the most from the WISC to WISC-R were eliminated from consideration, we still observed the FE, even though the difference we found of about 1½ point for all age groups was about half of the size of the traditional FE in the United States. That reduced difference may have reflected the fact that we estimated Performance IQ from Mazes, Coding, and two Object Assembly puzzles. Regardless, the 3-point per decade finding in the United States seems to be robust. It has been identified for the WISC-III and WISC-IV for various combinations of Wechsler's and Binet's scales (Flynn, 2007, Table 3) and also for Wechsler/K-ABC comparisons for normal children and those diagnosed with learning disabilities, behavior disorders, mental retardation, and hearing impairment (Kaufman & Kaufman, 1983, Table 4.19). The 3-point gain between 1972 (WISC-R) and 1981 (K-ABC) for so many normal and clinical samples is especially noteworthy because the K-ABC intelligence scales exclude the kinds of language-oriented and fact-oriented tasks that populate Wechsler's Verbal scales.

Methodological Problems Resulting From Changes From WISC to WISC-R

So I am *not* claiming that the notable changes from the WISC to the WISC-R challenge either the existence of the FE or its magnitude. But I readily challenge Flynn's interpretation of gains on specific subtests such as Similarities and his reliance on those subtest differences to help formulate his explanation of the FE and to develop his theory of intelligence.

What the huge changes did do—especially in the administration and scoring of several WISC subtests—is make it unfeasible to compare the subtest performance from one generation to another when the 1949 WISC is used as the anchor for those comparisons. In contrast to the major changes from the WISC to WISC-R, within-subtest changes were relatively minor from the WISC-R to WISC-III and WISC-III to WISC-IV—even though several totally new subtests were introduced in each version. (The same is true for WAIS-R to WAIS-III and WAIS-III to WAIS-IV.)

Why is it incorrect to estimate generational changes on a subtest whose administration, scoring, and occasionally item type changed drastically? Because the changes directly affect the ability to conduct valid counterbalanced research. Remember, the size of the generational change is a function of research studies in which the old test battery and the new test battery are administered to the same person in counterbalanced order. The presumption is that counterbalancing the test sequence controls for the test order, thereby neutralizing the practice effect. However, that may not be the case for the WISC versus WISC-R (or the WAIS vs. WAIS-R) on certain subtests because of the dramatic changes that occurred in administration and scoring from the old test to the new test. In particular, Similarities, Comprehension, and Picture Arrangement changed to such a degree that counterbalancing of these subtests is compromised. There is likely to be a *differential practice effect* that gives an extra boost to the individuals tested on the older subtest (WISC) after being given the new subtest first (on the WISC-R). Those children are given all the advantages of the feedback and extra questioning and "shaping" that were built in to the WISC-R Similarities, Comprehension, and Picture Arrangement subtests when they are given those subtests

first. Then when they are given the subtest for a second time (on the WISC), they have already incorporated the rules of the game (e.g., giving abstract responses to Similarities items, giving two reasons for certain Comprehension items, being shaped to understand the notion of time sequencing on Picture Arrangement). The children who apply this learning on the WISC are being compared to a normative reference group (the standardization sample tested in 1947-1948) who did not have these benefits. Therefore, WISC scaled scores on these subtests will be inflated for the group tested in the order WISC-R/WISC. But not so for the group tested on the WISC first, because they were given the same limited feedback as everyone from the long-ago normative sample. In other words, the changes from the WISC to WISC-R may have created a methodological problem for several subtests that prevented control of the practice effect via counterbalancing and thereby overestimated scores on the older versions of these subtests.

Other subtests are not likely influenced too much by the administrative and scoring changes from WISC to WISC-R (and WAIS to WAIS-R). Adding the instruction to Performance subtests to "work quickly," for example, makes the subtests fairer for those who might dawdle, but many children and most adolescents and adults do not need that reminder; observing examiners take out a stopwatch, and clicking it on and off, is enough of a clue that time is of the essence. So I do not believe that counterbalanced studies are compromised for other subtests or for the global IQs, especially FS-IQ. There may be some inflation of the V-IQ, P-IQ, and FS-IQ due to differential practice effects, but it is likely mitigated by the other subtests that have limited changes in content or administration.

I do, however, believe that the problem is considerable for three subtests that measure abstract reasoning: Similarities, Comprehension, and Picture Arrangement. For this article, however, it is only the two Verbal subtests that are of interest. Picture Arrangement did not survive on the WISC-IV and, therefore, lacks continuity from 1947 to 2002; it seemed to play little or no role in Flynn's (2007) theory development.

A Reanalysis of Flynn's (2007) Data on Specific Subtests

Table 1, based on Flynn's (2007, pp. 180-181) data, presents generational gains on the subtests that were retained in all versions between the WISC and WISC-IV (except for Digit Span, which was also excluded by Flynn, 2007, in his analysis of half-century subtest gains). In developing this table, I relied on Flynn's approach and reported gain scores in IQ points ($SD = 15$) rather than scaled-score points ($SD = 3$). Table 1 separates gains from WISC to WISC-R (1947-1972) from gains on WISC-R to WISC-IV (1972-2002), which, I believe, is the only accurate way to report these gains. However, this table also shows the WISC to WISC-IV half-century gain scores (1947-2002) that Flynn (2007) emphasized when reporting his data and developing his theory to explain the FE. The monumental half-century gain of 23.8 points on Similarities breaks down to 13.8 points from WISC to WISC-R and 10.0 points from WISC-R to WISC-IV. In other words, nearly 60% of the gain occurred by comparing two Similarities subtests that were alike in name only. I contend that an indeterminate amount of this gain is related to the considerable changes in the task, most notably the WISC-R procedure that provides examples of 2-point responses on designated items, and the WISC-R scoring system that encourages, rather than discourages, querying children's incomplete or ambiguous responses. Likewise, the gain score for Comprehension should only be considered valid for 1972 to 2002; the value for 1947 to 1972 simply lacks interpretability.

For ease of comparison, I have converted gains on the WISC subtests to "IQ points per decade" for the two time intervals of interest (see Table 2). The values for Similarities indicate a monstrously large gain of 5.65 points per decade from WISC to WISC-R, when apples and oranges are literally being compared. In contrast, the gain of 3.36 points between WISC-R and WISC-IV, though still substantial, is similar to the value of 3.19 for Block Design over that same time period

Table 1. Gains on WISC Subtests Between 1947 and 2002

Subtest	1947-1972, WISC to WISC–R (24.5 Years)	1972-2002, WISC-R to WISC–IV (29.75 Years)	1947-2002, WISC to WISC–IV (54.25 Years)
Verbal			
Information	2.2	0.0	2.2
Arithmetic	1.8	0.5	2.3
Vocabulary	1.9	2.5	4.4
Comprehension	6.0	5.0	11.0
Similarities	13.8	10.0	23.8
Performance			
Picture	3.7	8.0	11.7
Completion			
Block Design	6.4	9.5	15.9
Coding	11.0	7.0	18.0

Note: WISC = Wechsler Intelligence Scale for Children. Gains are expressed in IQ points (*SD* = 15) instead of the more customary scaled-score points (*SD* = 3) to be consistent with Flynn's (2007) methodology. Technically, 1947 is 1947.5 and 2002 is 2001.75, accounting for the fractions of years that characterize the intervals. The WISC was standardized in 1947-1948, the WISC-R in 1972, the WISC-III in 1989, and the WISC-IV in 2001-2002. Performance subtests are included in the table only if they were included on all four versions of the WISC.

Table 2. Gains Per Decade on WISC Subtests from 1947-1972 and 1972-2002

Subtest	1947-1972, WISC to WISC–R (24.5 Years)	1972-2002, WISC-R to WISC–IV (29.75 Years)
Verbal		
Information	0.88	0.00
Arithmetic	0.73	0.17
Vocabulary	0.78	0.84
Comprehension	2.45	1.68
Similarities	5.65	3.36
Performance		
Picture	1.51	2.69
Completion		
Block Design	2.15	3.19
Coding	4.49	2.35

Note: See Note to Table 1.

and to the overall value of about 3 points per decade that has presumably characterized the FE in the United States for three-quarters of a century.

It is interesting that virtually all the changes from the WISC to WISC-R were made to promote fairness to children, that is to say, to help them catch on to the tasks more readily (e.g., by asking for a "second reason why" on certain Comprehension items) and to make sure they were clued into what is rewarded most heavily (like abstract, categorical responses on Similarities items). As discussed earlier, the large gain on Similarities and the substantial gain on Comprehension, between 1947 and 1972, are conceivably due to the methodological problems that resulted from the widespread modifications of these subtests. As noted, the many changes from the WISC to WISC-R on subtests like Similarities conceivably led to a differential practice effect that gave a relatively larger bonus to children who were tested first on the WISC-R and second on the WISC. Regardless of the mechanisms that caused such large apparent gains between 1947 and 1972 on Similarities and Comprehension, one point is clear: These gains do not reflect increased ability in the unique skills measured by these subtests. The *constructs* measured by the specific subtests did not change

from the WISC to WISC-R, but the way those constructs were measured did change based on the substantial modifications in administration and scoring. The methodological concerns introduced by the altered subtests (e.g., differential practice effects) do not permit interpretable comparisons of scores earned on Similarities and Comprehension based on the 1947 norms versus the 1972 norms.

Raven's Matrices

Flynn (2007, 2009b) did not base his explanation of the FE solely on Similarities or the Performance subtests. He was also influenced by large worldwide gains on Raven's (1938, 2000) matrices (Flynn, 1999, 2007, 2009a). However, Raven's matrices are quite different from Similarities. The item type used in Similarities resembles the age-old questions that teachers have asked children in schools for generations. In contrast, matrices-type items were totally unknown to children or adults of yesteryear and remained pretty atypical for years. Over time, however, this item type has become more familiar to people around the world, especially as tests of this sort have been increasingly used for nonbiased assessment, including for the identification of gifted individuals from ethnic minorities. And, because Raven's tests can be administered by nonpsychologists, these items tend to be more accessible to the public than are items on Wechsler's scales, which are closely guarded because of the clinical training that is a requisite for qualified examiners. But go to any major bookstore chain, or visit popular websites, and you can easily find entire puzzle books or pages of abstract matrix analogies.

It is, therefore, difficult to evaluate gains on matrices tasks without correcting these gains for time-of-measurement effects. The power of this "time lag" variable was demonstrated by Owens in his groundbreaking longitudinal study of aging and intelligence. Owens (1953) administered the Army Alpha test in 1950 to 127 men, age 50, who had been administered the same test in 1919 at age 19, when they were freshmen at Iowa State University (initial $N = 363$). The study continued in 1961 when 96 of these men were tested again, at age 61 (Owens, 1966).

The 96 men tested three times improved in verbal ability between ages 19 and 50 followed by a slight decline from age 50 to 61. On nonverbal reasoning ability, they displayed small increments from one test to the next. However, Owens had the insight to also test a random sample of 19-year-old Iowa State freshmen on the Army Alpha in 1961 to 1962 to permit a time-lag comparison. He was able to use the data from the 19-year-olds to estimate the impact of cultural change on the obtained test scores. When Owens corrected the data for cultural change, the Verbal scores continued to show gains between ages 19 and 61; but what had appeared to be small increments in Reasoning were actually steady decreases in performance.

The time-lag correction may reflect real differences in mental ability (i.e., FE) as well as changes in test-taking ability and familiarity with a particular kind of task. The mere fact of large gains on a test such as Raven's matrices over several generations, in and of itself, cannot be interpreted unequivocally as an increase in abstract reasoning ability without proper experimental controls. When Flynn has interpreted gain scores for groups of individuals tested generations apart on the identical Raven's matrices items (e.g., Flynn, 1999, 2009a), he has *not* controlled for time-of-measurement effects. And whereas the combination of large gains on matrices and WISC Similarities provided mutual support for Flynn's (2007) claim that society has increased dramatically over time in thinking abstractly instead of concretely, more than half of the purported gain on Similarities is illusory.

Some Final Words

The aim of this article is to challenge Flynn's (2007, 2009b) interpretation of the FE, not the existence of the effect itself, or even its magnitude. But the acceptance of the generational change in IQs applies only to global scores such as Wechsler Full Scale IQ. In that instance, the impact

on the magnitude of the difference in the IQ norms, due to changes from one version of a particular subtest to the next, is mitigated by other subtests with fewer changes. But as soon as the issue becomes subtest differences—and the focus is on presumed growth or decline in the abilities measured by that subtest—then changes in administration, scoring, and item style become potentially crucial. Then the subtest may be measuring the construct a bit differently at Time 1 and Time 2, and it is not reasonable to interpret apparent gains on that specific subtest. Flynn's mistake was in assuming that the subtests did not change in meaningful ways over time and that is wrong. It is also, therefore, wrong to build a theory based on the unique abilities supposedly measured by specific subtests.

Flynn's Theory of Generational changes in IQ

Flynn's explanation of a wholesale change in a society's orientation toward abstract thinking as a result of scientific technology may be correct but it is not supported by WISC data to the degree he has suggested. He relied to a considerable degree on the seemingly huge generational change in Similarities over a period of more than half a century. Throughout his book, Flynn (2007) continually emphasized the logic needed to solve Similarities items as well as the distinction between the concrete approaches (used by our parents) and the abstract approaches (used by our children) for solving these items. Yet the enormous gain in Similarities reported by Flynn was inflated considerably by his decision to use data from the 1949 WISC as his baseline. Indeed, the administration and scoring guidelines for *most* WISC subtests were modified in notable ways.

Nonetheless, Flynn (2007) emphasized WISC subtest changes, not Full Scale IQ changes, to explain the various paradoxes that he believes are created by the FE. For example, consider one paradox that he proposes—that based on the worldwide Flynn research, the intellectual gulf between generations should be self-evident from casual conversation between children and their elders, but, of course, there is no such gulf. To resolve this paradox, Flynn (2007) relies on half-century subtest changes on the subtests that I have shown are "tainted" from 1947 to 1972, especially Similarities and Comprehension:

> The solution to the paradox is to be found not by focusing on Full Scale IQ trends, but by focusing on the WISC subtests plus Raven's trends. . . . [B]etween 1947 (WISC) and 2002 (WISC-IV), Similarities and Raven's show huge gains of 24 to 27 points ($SD = 15$), the five Performance subtests show gains averaging 17 points, Comprehension shows 11 points, and the remaining Verbal subtests . . . show very limited gains. (p. 19)

But Flynn's contentions fall short on other grounds. Piaget spent much of his life investigating the transition from concrete to formal operations at about ages 11 to 12. Flynn (2007) virtually dismisses Piaget's conclusions by stating, "but today, there is general agreement that Piaget worked with an elite sample of children" (p. 31). Yet Piaget's finding that ages 11 to 12 correspond to the onset of formal operations is congruent with the ages that the prefrontal cortex starts to mature (Golden, 1981) and the ages that the volume of frontal gray matter "peaks" for females (age 11) and males (age 12½), based on analysis of 243 neuroimaging scans of individuals ages 4 to 22 years (Giedd et al., 2006, Figure 1). These are the key cortical areas needed for Luria's (1966) notion of abstract thinking and "Block 3" planning ability.

The 11- to 12-year level is also the age when the Guatemalan Indian infants and children studied by Kagan and Klein (1973) displayed cognitive development that was comparable with the development of an American control group. The infants within this culture are startlingly deprived of stimulation during the first 15 months of life; they display the affect of infants raised in orphanages (Spitz & Wolf, 1946) and appear intellectually disabled at ages 5 to 6 when compared with

an American control group. Yet by about age 11, the children of these Indian cultures appear normal in every way and are able to perform well on tests of basic cognitive competencies including conceptual inference, which requires abstract verbal reasoning and was the prototype for the Riddles subtest on the K-ABC (Kaufman & Kaufman, 1983, 2004), and tests that require the ability to deal with abstract symbols (perceptual analysis and perceptual inference).

If Flynn (2007, 2009b) is correct in the dramatic improvement over the past 50 to 75 years in on-the-spot problem solving that emphasizes abstract rather than concrete solutions, then data on adolescent and adult populations should demonstrate that improvement. But why should that skill be evident in children on the WISC or the Wechsler Preschool and Primary Scale of Intelligence (WPPSI; Wechsler, 1967, 1989, 2002)? Zhou et al. (2010) identified a FE on P-IQ, continuing into the 21st century, of about 2½ to 3 points per decade for preschool and primary grade children on the WPPSI, for 6 to 16-year-old children on the WISC, and for adults on the WAIS. For Flynn's (2007) explanation of the FE to hold for preschool and school-age children, the improvement in abstract thinking would have had to filter down to very young brains. Yet a half-century or a century is far too brief a time for evolutionary changes to take hold.

In developing his theory, Flynn (2007, pp. 48-82) stresses: (a) the brain, notably what he refers to as "neural decentralization"; (b) the interaction between genetics and environment—what he refers to as "general intelligence"—as illustrated by his statement, "if two people have the same opportunity, the better mind is likely to accumulate a wider range of information" (p. 57); and (c) the role of society, a concept that he calls "social utility," the driving force behind the fact that "[v]arious real-world cognitive skills show different trends over time as a result of shifting social priorities" (p. 57). His theory is clearly based on a broad foundation that extends well beyond gains on specific WISC subtests. But apart from the tie-in of his theory to WISC subtest data, Flynn's theory has other weak aspects. Even though he incorporates a wealth of research in diverse areas that relate in one way or another to intelligence, his view of IQ is remarkably narrow. The only IQ tests worth mentioning are Wechsler's scales and Raven's matrices with an occasional glimpse at the Stanford–Binet. The only theories worth featuring are Jensen's (1998) "*g*" theory and Sternberg's (1984) triarchic theory of successful human intelligence.

Missing from his text and Index is any mention of today's contemporary theory-based tests that are of top-notch psychometric quality (e.g., Elliott, 2007; Naglieri & Das, 1997; Kaufman & Kaufman, 2004; Woodcock, McGrew, & Mather, 2001). Even more astonishing is the omission of the pioneers whose theory forms the foundation of most of today's IQ tests and is the main theory used to interpret Wechsler's scales—Raymond Cattell, John Horn, and John Carroll (CHC theory; McGrew, 2005).

I understand that Flynn's (2007, 2009b) theory depends on Raven's matrices, shifts in the ages at which children achieve Piagetian formal operational thought (Flieller, 1999), genetics, brain research, experimental psychology research on IQ, and other factors as well. But in reading *What Is Intelligence?* I get the distinct feeling that the 1.6 *SD* improvement in WISC Similarities between 1947 and 2002 is the linchpin of Flynn's (2007) argument. And that linchpin is coated with rust.

Acknowledgements

The author is grateful to the following professionals for providing invaluable feedback on earlier versions of this article: Ms. Loretta van Iterson, Dr. James C. Kaufman, Dr. Nadeen L. Kaufman, Dr. Cecil R. Reynolds, Dr. Jennie Kaufman Singer, and Dr. Lawrence G. Weiss.

Declaration of Conflicting Interests

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

Funding

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

Note

1. As a graduate student in the late 1960s at Columbia University, I took my first course in clinical assessment. As I was learning to administer the WISC and WAIS, I was troubled by several aspects of test administration. The scoring system for Similarities rewarded abstract responses with 2 points, but never let the examinee know the rules of the game. I observed children and adults say, "You eat them both" or "They're both parts of our body," and I wanted to tell them that they needed to be a bit more conceptual to earn full credit. But I could not say it, because the rules did not permit it, and it struck me as grossly unfair. Sometimes I would "test the limits" after the test was completed to see whether a chronic 1-point responder could be shaped to give a 2-point response ("Yes, you drink them both, and they are also both alcoholic beverages"). Often, they could easily be shaped to respond abstractly. Comprehension also had a built-in problem. Numerous items gave 2 points to a person who responded with two different ideas to a question, but the examiner was not allowed to ask for "another reason why" when a person gave a short, crisp answer and saw no need to embellish. Again, by testing the limits I found that many individuals could talk at length about a social comprehension question, and easily earn 2 points on items, if prompted to do so. Finally, bonus points were rewarded for many Performance items—sometimes solving an item correctly in 1 to 5 seconds would add 3 bonus points to an item score—but, again, it was taboo to let examinees know that it was important to work quickly. Some children solved problems as if they were the fireman who smoked a cigar before going to put out a fire (a Verbal Absurdities item on the old Stanford–Binet), but they were able to respond quickly if they knew that speed was valued so highly.
   How fortunate I was, less than 2 years later, to be working at The Psychological Corporation with Dr. Wechsler, under Dr. Jerome Doppelt's tutelage, on the exciting (yet daunting) task of revising and restandardizing the WISC. Though Dr. Wechsler played the devil's advocate with my "radical" ideas for change, the fact that he was ultimately so flexible was a gift. As I once wrote (Kaufman, 1994), "Little did I realize that those battles with the Master would shape my own development as a test author and trainer of school psychologists, and would remain forever etched—fresh and vibrant and poignant—in my memory" (p. xv).

References

Atkins v. Virginia (2002). 536 U.S. 304, 122 S.CT 2242.

Beaujean, A. A., & Osterlind, S. J. (2008). Using item response theory to assess the Flynn effect in the National Longitudinal Study of Youth 79 children and young adults data. *Intelligence*, *36*, 455-463.

Boehm, A. E. (1971). *Manual for the Boehm Test of Basic Concepts*. New York, NY: Psychological Corporation.

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

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

Doppelt, J. E., & Kaufman, A. S. (1977). Estimation of the differences between WISC-R and WISC IQs. *Educational and Psychological Measurement*, *37*, 417-424.

Elliott, C. D. (2007). *Differential Ability Scales–Second Edition* (DAS-II). San Antonio, TX: Psychological Corporation.

Flieller, A. (1999). Comparison of the development of formal thought in adolescent cohorts aged 10 to 15 years (1967-1996 and 1972-1993). *Developmental Psychology*, *35*, 1048-1058.

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

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

Flynn, J. R. (1998). IQ gains over time: Toward finding the causes. In U. Neisser (Ed.), *The rising curve: Long-term gains in IQ and related measures* (pp. 25-66). Washington, DC: American Psychological Association.

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

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?* New York, NY: Cambridge University Press.

Flynn, J. R. (2009a). Requiem for nutrition as the cause of IQ gains: Raven's gains in Britain 1938-2008. *Economics and Human Biology*, *7*, 18-27.

Flynn, J. R. (2009b). *What is intelligence?* (Expanded Edition). New York, NY: Cambridge University Press.

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*, 1-16.

Giedd, J. N., Clasen, L. S., Lenroot, R., Greenstein, D., Wallace, G. L., Ordaz, S., . . . Chrousos, G. P.(2006). Puberty-related influences on brain development. *Molecular and Cellular Endocrinology, 254-255*, 154-162.

Golden, C. J. (1981). The Luria-Nebraska Children's Battery: Theory and formulation. In G. W. Hynd & J. E. Obrzut (Eds.), *Neuropsychological assessment and the school-age child: Issues and procedures* (pp. 277-302). New York, NY: Grune & Stratton.

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

Kagan, J., & Klein, R. E. (1973). Cross-cultural perspectives on early development. *American Psychologist*, *28*, 947-961.

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

Kaufman, A. S. (1979). *Intelligent testing with the WISC-R.* New York, NY: Wiley.

Kaufman, A. S. (1990). *Assessing adolescent and adult intelligence*. Boston, MA: Allyn & Bacon.

Kaufman, A. S. (1994). *Intelligent testing with the WISC-III.* New York, NY: Wiley.

Kaufman, A. S., & Kaufman, N. L. (1983). *Kaufman Assessment Battery for Children (K-ABC) interpretive manual*. Circle Pines, MN: American Guidance Service.

Kaufman, A. S., & Kaufman, N. L. (2004). *Kaufman Assessment Battery for Children–Second Edition (KABC-II)*. Circle Pines, MN: American Guidance Service.

Luria, A. R. (1966). *Higher cortical functions in man*. New York: Basic Books.

McGrew, K. S. (2005). The Cattell-Horn-Carroll theory of cognitive abilities: Past, present, and future. 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.

Naglieri, J. A., & Das, J. P. (1997). *Cognitive assessment system.* Itasca, IL: Riverside.

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

Owens, W. A. (1953). Age and mental abilities: A longitudinal study. *Genetic Psychology Monographs*, *48*, 3-54.

Owens, W. A. (1966). Age and mental ability: A second adult follow-up. *Journal of Educational Psychology*, *57,* 311-325.

People v. Superior Court (Vidal) (2005). 129 Cal. App. 4th 434, 28 Cal Rptr. 3d 529 (5th Ct. App. 2005), *vacated and later proceedings at People v. S.C.,* 2005 Cal. LEXIS 13290 (Cal., November 17, 2005).

The Psychological Corporation (2003). *WISC-IV technical and interpretive manual*. San Antonio, TX: Author.

The Psychological Corporation. (2008). *WAIS-IV technical and interpretive manual.* San Antonio, TX: Author.

Raven, J. C. (1938). *Progressive matrices.* London, England: Lewis.

Raven, J. C. (2000). *Raven manual research supplement 3; neuropsychological applications*. Oxford, England: Oxford Psychologists Press.

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

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

Roid, G. (2003). *Stanford–Binet Intelligence Scales* (5th Ed.). Itasca, IL: Riverside.

Spitz, R. A., & Wolf, K. M. (1946). Anaclitic depression; an inquiry into the genesis of psychiatric conditions in early childhood. In A. Freud (Ed.), *The psychoanalytic study of the child* (Vol. II, pp. 313-342). New York, NY: International Universities Press.

Sternberg, R. J. (1984). *Beyond IQ: A triarchic theory of human intelligence.* New York, NY: Cambridge University Press.

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

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

Teasdale, T. W., & Owen, D. R. (2008). Secular declines in cognitive test scores: A reversal of the Flynn effect. *Intelligence*, *36*, 121-126.

Thorndike, R. L., Hagen, E. P., & Sattler, J. M. (1986). *Stanford-Binet Intelligence Scale* (4th ed.). Chicago, IL: Riverside.

Thompson, A. P. (1987). Methodological issues in the clinical evaluation of two- and four-subtest short forms of the WAIS-R. *Journal of Clinical Psychology, 43*, 142-144.

Thompson, A. P., Howard, D., & Anderson, J. (1986). Two- and four-subtest short forms of the WAIS-R: Validity in a psychiatric sample. *Canadian Journal of Behavioral Science*, *18*, 287-293.

van de Vijver, F. J. R., Mylonas, K., Pavlopoulos, V., & Georgas, J. (2003). Methodology of combining the WISC-III data sets. In J. Georgas, L. G. Weiss, F. J. R. van de Vijver, & D. H. Saklofske (Eds.), *Culture and children's intelligence: Cross-cultural analysis of the WISC-III* (pp. 265-276). San Diego, CA: Academic Press.

Wechsler, D. (1949). *Manual for the Wechsler Intelligence Scale for Children (WISC).* New York, NY: Psychological Corporation.

Wechsler, D. (1955). *Manual for the Wechsler Adult Intelligence Scale (WAIS).* San Antonio, TX: Psychological Corporation.

Wechsler, D. (1967). *Manual for the Wechsler Preschool and Primary Scale of Intelligence (WPPSI).* San Antonio, TX: Psychological Corporation.

Wechsler, D. (1974). *Manual for the Wechsler Intelligence Scale for Children-Revised Edition (WISC-R).* San Antonio, TX: Psychological Corporation.

Wechsler, D. (1981). *Manual for the Wechsler Adult Intelligence Scale–Revised (WAIS-R).* San Antonio, TX: Psychological Corporation.

Wechsler, D. (1989). *Manual for the Wechsler Preschool and Primary Scale of Intelligence–Revised (WPPSI-R).* San Antonio, TX: Psychological Corporation.

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

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

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

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

Zhou, X., & Zhu, J. (2007, August). *Peeking inside the "blackbox" of Flynn effect: Evidence from three Wechsler instruments.* Paper presented at 115th annual convention of American Psychological, San Francisco, CA.

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