INTERNATIONAL JOURNAL OF TESTING, *7*(2), 209–224
Copyright © 2007, Lawrence Erlbaum Associates, Inc.

American IQ Gains From 1932 to 2002: The WISC Subtests and Educational Progress

James R. Flynn
Department of Political Studies
University of Otago
Dunedin, New Zealand

Lawrence G. Weiss
Harcourt Assessment, Inc.
San Antonio, Texas, USA

Recent data from 12 pairs of tests representing eight standardization samples show that American IQ gains have occurred at a rate of 0.308 points per year from 1972 to 2002. Linked with earlier IQ gains, Americans have gained about 22 points over the 70 years between 1932 and 2002. Comparing the new WISC-IV (2002) and the old WISC-III (1989) shows a difference of only 2.5 points. However, they have only five subtests in common when full scale IQ is calculated. If one simulates a comparison of the WISC-III and WISC-IV standardization samples on the 10 subtests of the WISC-III, IQ gains over the intervening 12.75 years were no less than 3.83 points, yielding a minimum estimate of 0.300 points per year. Finally, WISC subtest trends taken in conjunction with "the Nation's Report Card" (NAEP test trends) provide a fascinating picture of the evolution of cognitive skills in America over the last two generations.

Key words: Flynn effect, IQ gains, education gains

IQ trends over time are the product of factors that fluctuate over time and therefore must be estimated anew whenever fresh data become available. We are not sure of the exact causes, but putative causes for the post-1932 era, at least in developed na-

Correspondence should be addressed to James R. Flynn, University of Otago, Department of Political Studies, P.O. Box 56, Dunedin, New Zealand. E-mail: jim.flynn@stonebow.otago.ac.nz

tions, include urbanization, ratio of adults to children in the home, more liberal parenting styles, emphasis on lateral thinking in schools, more leisure time spent on cognitively demanding pursuits, and so forth. It is intriguing that IQ gains seem to have stalled in Scandinavian nations who may have entered this phase earliest and may now be reaching a saturation point. Schneider (2006) gives an excellent summary of these trends. After all, eventually the rural population is virtually eliminated, the number of children in the typical family is unlikely to go below 1.5, more solo-parent homes can actually worsen the ratio of adults to children, and people must approach a point at which they want leisure to be relaxing rather than challenging.

Scandinavian conditions are far from those in developing nations and these may be just beginning the massive IQ gains the developed world has enjoyed during the 20th century, witness Kenya (Daley, Whaley, Sigman, Espinosa, & Neumann, 2003). America may or may not soon follow suit; for one thing, it has large minority populations who are gaining at an unusually high rate (Dickens & Flynn, 2006).

Fortunately, the recent standardizations of the Stanford-Binet 5 and WISC-IV give us a snapshot of trends from 1932 virtually up to the present day on Wechsler-Binet tests.

METHODS

As to methodology, standardization samples attempt to norm IQ tests on a representative sample of Americans. Someone who equals the average performance of the sample against which he or she is scored gets an IQ of 100. If IQ gains over time have occurred, as we go back into the past, the performance of representative samples deteriorates and the average performance is easier to beat. Therefore, if we give the same subjects the WISC (normed in 1947 to1948) and the WISC-R (normed in 1972), they may be dead average on the latter (and get an IQ of 100) but score 8 points better on the former (get an IQ of 108). All of the comparisons described herein follow this methodology: using groups of subjects who take two tests to determine whether IQ norms weaken as we go back into the past.

An example may help. Standards in the high jump have been rising over time. Therefore, someone who only matches the average performance of competitive high jumpers today would have qualified for the Olympics measured against the much lower standards of 50 years ago. The high jump has of course been unaltered over time, while some IQ tests have been revised to replace obsolete items. However, gains over time are similar on both tests that have been revised and those whose items have been unchanged (Flynn, 1987). Moreover, obsolete items would actually mean an underestimate of the rate of IQ gains. It is contemporary children who take both an up-to-date test and an older test with obsolete items. If those

items lower their performance, their scores on an earlier test will be deflated and rise above their scores on a later test less than they should.

Estimates for the period from 1932 to 1972 were based on a huge data-set of 73 studies covering almost 7,500 subjects as detailed in Flynn (1984b). A literature of 26 studies covering 2,266 subjects is available for one pair of tests that cover the period 1972 to 1989, namely, the WISC-R and WISC-III (Flynn, 1998b). But most recent studies consist of groups of 100 to 250 subjects selected by publishers to be typical (in terms of their distribution over the IQ curve) and therefore used to compare two tests normed some years apart. IQ test comparisons by independent scholars have gone out of fashion. The only antidote is to do what has been done here: to use a matrix of 12 comparisons of pairs of tests and average the results to cancel out anomalies resulting from the small number used in any given comparison.

The fact that these comparisons have been done by the publishers is a plus. Rodgers (1998) emphasized that estimates should be based on subjects from all IQ levels to see if they show roughly equivalent gains. The Psychological Corporation has accompanied its comparisons with tables that show that IQ gains have been roughly similar at all IQ levels, although shifting scoring conventions have caused problems at very low IQ levels. Flynn (2006b; in press, Table 4) summarizes this data. As for the magnitude of gains, the studies done by the Psychological Corporation have an excellent track record. Their comparison of the WISC-R and WISC-III used a sample of 206 subjects. These gave an estimate of 5.30 points for Full Scale IQ, 2.40 points for Verbal IQ, and 7.40 points for performance IQ. Almost a decade later, Zimmerman and Woo-Sam (1997) collected results from 26 studies involving 2266 subjects. Flynn (1998b) analyzed these with regard to a balanced representation of the whole IQ curve and this put Full Scale IQ at 5.12, Verbal IQ at 3.53, and Performance IQ at 5.91.

Our estimates for the WISC-III and WISC-IV are based on Psychological Corporation data from 244 subjects. We suspect it will take more than a decade for reasonable literature to accumulate and add confirmation. If it does accumulate, at least we will have good numbers for the WISC tests that span the whole period from 1948 to 2002. We wish to give advance warning that the literature will lose much of its value if comparisons by subtests are not given. Full Scale IQ would do for the practical purpose of comparing and evaluating the IQ scores of children who happen to take both tests. But for confirming our estimates of IQ gains between the WISC-III and the WISC-IV, for reasons to be given, studies would have to focus on the five subtests the two tests have in common.

Our matrix of 12 comparisons of pairs of tests that cover the 1972 to 2002 era is important beyond helping to make up the numbers. The major methodological problem that has emerged in recent years is the realization that it is much more difficult to get standardization samples that are representative of adults than it is to get good samples of schoolchildren. With schoolchildren, as long as you locate a good

sample of schools and test everyone, you are home-free. Adults are not gathered together in one institution and can be located only if you visit all of their workplaces or homes. It is likely that adult sample data can vary within a few points depending on the rigor of the sampling techniques involved.

Fortunately, the 12 comparisons presented herein include five combinations in which adult tests are compared to schoolchildren tests. There are also seven combinations in which a given adult test is the earlier test in one pair and the later test in another pair. Assume that an adult sample was substandard by say 2 IQ points. It would then inflate IQ scores whenever it was the earlier test; and this would lead to an overestimate of gains. But it would also inflate scores wherever it was the later test; and this would lead to an underestimate of IQ gains. So averaging the two comparisons would cancel out the bias. See Flynn (2006b) for a detailed discussion.

IQ GAINS FROM 1932 TO 2002

As for the history of measuring the IQ gains of American children on Wechsler-Binet tests, Flynn (1984b) chose the initial date of 1932 because the Stanford-Binet sample of that year was the first that could claim to be reasonably reliable. At that time, there was adequate data only through 1972. Flynn (1998b) calculated trends from 1972 to 1995, but he did so based on only three comparisons of pairs of standardization samples. Fortunately, the 12 comparisons now available are based on eight Wechsler-Binet standardization samples, all of which were selected between 1972 and 2002 (2001.75 to be precise) and all tested at least 6 years apart. For the first time, we can trace IQ trends in America all of the way from 1932 to 2002.

Table 1 gives the full names of the 12 tests whose standardization samples have been compared. The dates therein refer to the date the samples were tested rather than the date of publication. It provides a corrective to Flynn's analysis (1998b). At that time, he speculated that the historic rate of gain of just over 0.300 IQ points per year might have slowed, circa 1972, to about 0.250 points per year but advised patience until a wider array of data was available. Table 1 shows that the historic rate has not slowed. The 12 post-1972 comparisons now available yield an average rate of gain somewhere between 0.305 and 0.311 IQ points per year, depending on where one puts gains from the WISC-III to the WISC-IV. The reasons for a range of estimates for that pair of tests will be discussed in the next section.

Table 1 suggests 0.308 IQ points per year, the midpoint of the two averages, as a best estimate of the rate of gain between 1972 and 2002. This overlaps with the earlier period between 1932 and 1978 that showed a rate of 0.311 points per year (Flynn, 1987, Table 7). American IQ gains have been remarkably stable over seven decades. Between 1932 and 2002, the total gain comes to almost 22 IQ points.

TABLE 1
America From 1972 to 2002: Twelve Estimates of the Rate of IQ Gain

Tests Compared	*Gains*	*Period (Yrs)*	*Rate*
(1) WISC-R (1972) & WAIS-R (1978)	+0.90	6	+0.150
(2) SB-LM (1972) & SB-4 (1985)	+2.16	13	+0.166
(3) WISC-R (1972) & SB-4 (1985)	+2.95	13	+0.227
(4) WISC-R (1972) & WISC-III (1989)	+5.30	17	+0.312
(5) WAIS-R (1978) & SB-4 (1985)	+3.42	7	+0.489
(6) WAIS-R (1978) & WAIS-III(1995)	+2.90	17	+0.171
(7) SB-4 (1885) & SB-5 (2001)	+2.77	16	+0.173
(8) WISC-III (1989) & WAIS-III (1995)	–0.70	6	–0.117
(9) WISC-III (1989) & WISC-IV (2001.75)	+3.83/4.63	12.75	+0.300/0.363
(10) WISC-III (1989) & SB-5 (2001)	+5.00	12	+0.417
(11) WAIS-III (1995) & SB-5 (2001)	+5.50	6	+0.917
(12) WAIS-III (1995) & WISC-IV (2001.75)	+3.10	6.75	+0.459
Average of 12 comparisons:			0.305/0.311

Test names and sources:
(1) WISC-R & WAIS-R. Wechsler (1981), Table 18
(2) SB-LM & SB-4. Thorndike, Hagen, & Sattler (1986), Table 6.6
(3) WISC-R & SB-4. Thorndike et al. (1986), Table 6.7
(4) WISC-R & WISC-III. Flynn (1998b), Table 1
(5) WAIS-R & SB-4. Thorndike et al. (1986), Table 6.9
(6) WAIS-R & WAIS-III. Wechsler (1997), Table 4.1
(7) SB-4 & SB-5. Roid (2003), Table 4.1
(8) WISC-III & WAIS-III. Wechsler (1997), Table 4.3
(9) WISC-III & WISC-IV. Sources to be detailed below
(10) WISC-III & SB-5. Roid (2003), Table 4.6
(11) WAIS-III & SB-5. Roid (2003), Table 4.7
(12) WAIS-III & WISC-IV. Wechsler (2003), Table 5.12

Notes. Until the SB-5, the Stanford-Binet had an SD of 16; all of the above estimates were based on converting its scores using an SD of 15 to provide a common metric.

This table is adapted (with permission) from Table 1 of Flynn (2006b).

WISC GAINS FROM 1948 TO 2002

American gains on the WISC are of interest because they illuminate what is happening in American schools. However, before making good on that assertion, we must discuss gains from the WISC-III (1989) to the WISC-IV (2002). The dates in brackets refer to the years in which the standardization samples used to norm the tests were actually tested, a practice we will follow throughout. This pair of tests poses special problems because of altered content.

A huge array of international data show that the rate of IQ gains over time varies with the kind of test (Flynn, 1987). Raven's Progressive Matrices and the Wechsler

Similarities subtest have shown the largest gains (0.5 points per year or more) and WISC full scale IQ intermediate gains (about 0.3 points per year in the U.S.). The new WISC-IV is so different from its predecessors that it constitutes a new kind of test—a kind that may well show a lesser rate of gain. There is no harm in this: The purposes of the Psychological Corporation hardly include the goal of maximizing IQ gains over time.

However, we must not confuse a diminished rate of gain caused by a new test with one caused by a change in the real world, that is, a change in how much certain cognitive skills are being enhanced over time. The way to disentangle these two causes is to compare: (1) the score difference between the altered WISC-IV and the WISC-III with (2) the score difference between a (simulated) unaltered version of the WISC-IV and the WISC-III.

Of the 10 subtests used to calculate WISC-III full scale IQ only five are used to calculate WISC-IV IQ. Fortunately, the Psychological Corporation used its most recent standardization sample to establish current norms for not only those five but also three more WISC-III subtests—tests that were mainstream for the WISC-III but are only supplementary options for the WISC-IV. They also gave a sample of 244 children all eight subtests and compared their scores against the older WISC-III norms and the current norms. This means we can estimate performance trends on all of the old WISC-III subtests except Object Assembly and Picture Arrangement, an estimate to cover the period between 1989 (WISC III sample tested) and 2001.75 (WISC-IV sample tested). Moreover, based on the position of those two subtests in the hierarchy of subtest gains in the past, we can give good estimates of what would have happened had they been retained.

Table 2 assigns the WISC-IV standardization the date of 2001.75. The testing was done over the 16 months from September 2001 to December 2002 with a midpoint of April 22, 2002. The latest gains (1989 to 2001.75) are put at somewhere between 0.300 points per year and 0.363 points, the bounds being set by how one estimated values for the two missing subtests.

The estimated gains for Object Assembly and Picture Arrangement were calculated by two methods. First, it was assumed that those two subtests would have made up the same proportion of the total gain in the WISC-III to WISC-IV period as they did in the WISC-R to WISC-III period:

WISC-R to WISC-III period: 1.2 + 1.9 = 3.1; 3.1 divided by 7.9 = 39%

WISC-III to WISC-IV period: [0.93] + [1.47] = 2.4; 2.4 divided by 6.1 = 39%.

The bracketed values represent the estimates for Object Assembly and Picture Arrangement. Second, it might be argued that those two subtests made abnormally high gains in the WISC-R to WISC-III period and that a more modest estimate would be judicious. Therefore, it was assumed that their gains in the WISC-III to

TABLE 2
WISC Subtest and Full Scale IQ Gains: 1948 to 2002

	WISC to WISC-R 1947.5–72 *Gain 24.5 Yrs (SD = 3)*	*WISC-R to WISC-III 1972–89* *Gain 17 Yrs (SD = 3)*	*WISC-III to WISC-IV 1989–2001.75* *Gain 12.75 Yrs (SD = 3)*	*WISC to WISC-IV 1947.5–2001.75* *Gain 54.25 Yrs (SD = 3)*	*WISC to WISC-IV 1947.5–2001.75* *Gain 54.25 Yrs (SD = 15)*
Information	0.43	–0.3	0.3	0.43	2.15
Arithmetic	0.36	0.3	–0.2	0.46	2.30
Vocabulary	0.38	0.4	0.1	0.88	4.40
Comprehension	1.20	0.6	0.4	2.20	11.00
Picture completion	0.74	0.9	0.7	2.34	11.70
Block design	1.28	0.9	1.0	3.18	15.90
Object assembly	1.34	1.2	[0.93]	[3.47]	[17.35]
Coding	2.20	0.7	0.7	3.60	18.00
Picture arrangement	0.93	1.9	[1.47]	[4.30]	[21.50]
Similarities	2.77	1.3	0.7	4.77	23.85
SUM[a]	11.63	7.9	6.1	25.63	
SUM[b]	11.63	7.9	5.3	24.83	

(continued)

TABLE 2 *(Continued)*

	Subtest Sums	*Full Scale IQ*	*Gain*	*Rate/Year*
WISC	100.00	100.00	—	—
WISC-R	111.63	107.63	7.63	0.311
WISC-III	119.53	113.00	5.47	0.322
WISC-IVa	125.63	117.63	4.63	0.363
WISC-IVb	124.83	116.83	3.83	0.300
	Gains SD = 3	*Gains in SDs*		*Gains SD = 15*
Trends on selected subtests				
Similarities (1948–2002):	4.77	1.59		23.85 IQ
Digit Span (1972–2002):	0.20	0.07		1.00 IQ
Coding + Symbol Search (ave) (1989–2002):	0.95	0.32		4.74 IQ

Adapted from Flynn (2006a); used with permission: ArtMed Publishers. Sources: Flynn (2000, Table 1); Wechsler (2003, Table 5.8); Wechsler (1992, Table 6.8).

Notes.

[a]With values for OA and PA at those bracketed (see text).

[b]With values for OA and PA put at 0.80 for both (see text).

1. It is customary to score subtests on a scale in which the SD is 3, as opposed to IQ scores which are scaled with SD set at 15. To convert to IQ, just multiply subtest gains by 5, as was done to get the IQ gains in the last column.
2. As to how the full scale IQs at the bottom of the table were derived:
   a. The average member of the WISC sample (1947–48) was set at 100.
   b. The subtest gains by the WISC-R sample (1972) were summed and added to 100: 100 + 11.63 + 111.63.
   c. The appropriate conversion table was used to convert this sum into a Full Scale IQ score. The WISC-III table was chosen so that all samples would be scored against a common measure. That table equates 111.63 with an IQ of 107.63.
   d. Thus the IQ gain from WISC to WISC-R was 7.63 IQ points.
   e. Since the period between those two samples was 24.5 years, the rate of gain was 0.311 points per year (7.63 divided by 24.5 = 0.311).
   f. The subsequent gains are also calculated against the WISC sample, which is to say they are cumulative. By the time of the WISC-IV, closer to 2002 than 2001, you get a total IQ gain of somewhere between 16.83 and 17.63 IQ points over the whole period of 54.25 years. Taking the mid-point (17.23 points) gives an average rate of 0.318 points per year, with some minor variation (as the table shows) from one era to another.

WISC-IV period would have been the average gain of the three subtests that "surround" them in the subtest gains hierarchy, namely, Block Design, Coding, and Similarities. These three subtests had an average gain of 0.8 (1.0 + 0.7 + 0.7 = 2.4; 2.4 divided by 3 = 0.8). So the lesser value of 0.8 was assigned to both Object Assembly and Picture Arrangement.

Table 2 also records WISC gains all of the way from 1947–1948 (1947.5) to 2001.75. These virtually match both the historic rate of gain on Wechsler-Binet that has held since 1932 and the post-1972 estimate based on the Wechsler-Binet tests compared in Table 1. The latter indicated a consistent rate of 0.308 IQ points per year from 1972 to 2002. Taking the midpoint of the most recent estimates, the WISC gives a gain of 0.318 points for the period between 1948 and 2002 with little variation. As to how full scale IQ gains were derived from the scaled score gains on the various subtests, see Table 2. Essentially, the sums of the gains on the 10 subtests were converted using the appropriate table from the WISC-III to provide a common metric.

SIMILARITIES AND DIGIT SPAN

The subtest differences in Table 2 are illuminating. Looking at the whole period from 1948 to 2002, Similarities comes out on top (as usual) with a gain of 4.77 scaled score points. This gain amounts to 1.59 SDs and using 15 as SD, it equates to almost 24 IQ points.

Similarities contains a few items such as 'How are dawn and dusk alike?' In response, the children have to imagine alternatives and select the one that best captures an intrinsic similarity. The child's mental processes would run something like this: "You get up in the morning and go to bed at night but that makes no sense because I often sleep past dawn and go to bed after dark. They are alike in that the sky is half-lit and often very pretty but of course that is not always true. What they really have in common is that they are the beginning and end of both the day and the night. The right answer must be that they separate day and night."

On the other hand, most items are questions like "How are dogs and rabbits alike?" 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.

In other words, Similarities gains may be huge; indeed, today's children are literally at the 94th percentile of their grandparents' generation. But the skill gain is not identical to that signaled by the huge gains on Raven's Progressive Matrices that are found in every advanced nation (Flynn, 1998a). The Raven's gains are unambiguously ones of on-the-spot problem-solving. However, both gains have a prerequisite in common: problems must be taken seriously even though they have no obvious practical pay-off. Note that the other subtests that show large gains resemble it in this regard: the child must rearrange blocks so that the view from above duplicates a presented pattern, build an object out of its disassembled parts, arrange pictures to tell a story.

The essence of the Similarities gains is this: thanks to formal education, children have begun to view the world through the spectacles provided by science. This means that they not only are more familiar with the vocabulary of science but that they are also more likely to take seriously problems that are detached from everyday life. Therefore, the two trends that Similarities measures are reinforcing: the scientific ethos favors abstract problem-solving, learning to attack such problems renders the scientific ethos more and more relevant.

Even though Similarities items and items on the Performance subtests have no obvious practical pay-off, that does not mean that IQ gains have no practical effect. Over time America has demanded more and more people who can play professional and managerial work roles. That means people with a university education who can solve problems and make decisions without following a rulebook or taking orders from a superior. The dramatic expansion of the tertiary population would have been impossible without students with some grasp of scientific taxonomy. Performance of professional work roles would be difficult without what might be roughly labeled "innovative thinking."

Until recently, Digit Span was not one of the 10 core subtests of the WISC but what data exist in the literature that underpins Table 2 show almost no gain from 1972 to 2002. This test measures not only rote memory but also working memory, that is, our ability to manipulate what we call to mind. Perhaps society has not improved this cognitive skill because we need no greater store of memories or ability to manipulate memories to deal with the world today than in the past. On the other hand, two subtests that measure processing speed, Coding and Symbol Search, made substantial gains (equivalent to 4.75 IQ points) in the brief period from 1989 to 2002. Symbol search also became a core subtest only recently and therefore, the supporting data is less extensive than in the case of Coding. Perhaps the speeded-up tempo of events on the visual media that condition us today is having some effect on the speed with which we process information.

FACTOR INVARIANCE AND IQ GAINS

Wicherts et al. (2004) have shown that IQ gains from one time to another are often not factor-invariant. This means that when you analyze the gains on the various WISC subtests, the relative magnitude of the gains do not match the relative factor loadings of the subtests. For example, both Arithmetic and Similarities have very high *g*-loadings, but gains on the latter are huge and gains on the former very small. The implications have been discussed elsewhere and will soon receive book-length treatment (Flynn, 2003; 2006a; in press). Here, we will merely summarize the central point: factor analysis compares individuals in a static situation at a given time; gains on the WISC subtests reflect dynamic trends over time which lay bare how operational cognitive skills function in the real world. The real-word significance of these trends is not diminished by their lack of factor invariance.

The inevitable sports analogy may help. If we factor analyzed performances on the 10 events of the decathlon, a general factor or *g* would emerge and no doubt, subordinate factors representing speed (the sprints), spring (jumping events), and strength (throwing events). We would get a *g*(D) because at a given time and place, performance on the 10 events would be intercorrelated, that is, someone who tended to be superior on any one would tend to be above average on all. We would also get various *g*-loadings for the 10 events, that is, superior performers would tend to rise further above average on some of them than on the others. The 100 meters would have a much higher *g*-loading than the 1500 meters, which involves an endurance factor not very necessary in the other events.

Decathlon *g* might well have much utility in predicting performance differences between athletes of the same age cohort. However, if we used it to predict progress over time and forecast that trends on the 10 events would move in tandem, we would go astray. That is because *g*(D) cannot discriminate between pairs of events in terms of the extent to which they are functionally interrelated.

Let us assume that the 100 meters, the hurdles, and the high jump all had large and similar *g* loadings as they almost certainly would. A sprinter needs upper body strength as well as speed, a hurdler needs speed and spring, a high jumper needs spring and timing. We have no doubt that a good athlete would best the average athlete handily on all three at a given place and time. However, over time, social priorities change. People become obsessed with the 100 meters as the most spectacular spectator event (the world's fastest human). Young people find success in this event a secondary sex characteristic of great allure. Over 30 years, performance escalates by a full SD in the 100 meters, by half an SD in the hurdles, and not at all in the high jump.

We are mystified so we talk to some athletics coaches. They tell us that over the years, everyone has become focused on the 100 meters and it is hard to get people to take other events seriously. They point out that sprint speed may be highly corre-

lated with high jump performance but past a certain point, it is actually counterproductive. If you hurl yourself at the bar at maximum speed, your forward momentum cannot be converted into upward lift and you are likely to time your jump badly. They are not surprised that increased sprint speed has made some contribution to the hurdles because speed between the hurdles is important. But it is only half the story: you have to control your speed so that you take the same number of steps between hurdles and always jump off the same foot.

In sum, the trends do not mimic the relative *g* loadings of the "subtests." One pair of events highly correlated (sprint and hurdles) shows a modest trend for both to move in the same direction and another pair equally highly correlated (sprint and high jump) shows trends greatly at variance. But there is a good reason for this discrepancy: factor analysis cannot discriminate between operational skills that develop independently over time. At the end of the 30 years, we do another factor analysis of performance on the 10 events of the decathlon and lo and behold, *g*(D) is still there. Although average performance has risen "eccentrically" on various events, the following is still true: superior performers still do better than average on all 10 events and are about the same degree above average on various events as they were 30 years before.

Factor analysis captures the static beautifully but not the dynamic. The central point to grasp is this: the rise in 100 meter skills over time and the nil gain in high jump skills over time are simply facts. That they are not factor-invariant cannot alter those facts.

Similarly, even though the pattern of great progress plus little or no progress on various WISC subtests does not mimic factor loadings, this does not affect their real world significance. Factor analysis yields no factor called "looking at the world through scientific spectacles" or "freeing logic and the hypothetical from the concrete." Yet these have great social significance. Whether or not we develop larger everyday vocabularies and funds of information may be merely "test-specific" but they affect virtually everything that makes us human. What a pity that the pattern of subtests results do not get the blessing of factor analysis. However, we must not let that impede our understanding of score gains over time.

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. You give mainly automatic responses: you either know that Rome is the capital of Italy or you only know of Rome Georgia; you either know what "delectable" means or you do not. The latter do not involve freeing intelligence from the world of everyday life but rather, the words we use in ordinary conversation and the facts of the external world. The transition from reading to a visual culture and from concrete to abstract could easily enhance our ability to classify and to solve certain problems without being accompanied by any increase in general information and language skills.

"SCHOOL" SUBTESTS AND THE NAEP

We will supplement the above analysis by taking a more direct approach to establishing the real-world significance of gains on the various WISC subtests. I will attempt to show that trends on certain WISC subtests illuminate academic achievement trends in America's schools. The implication is that if trends are informative about real-word cognitive skills, they must themselves be grounded in the real world.

Table 2 traces trends on the three WISC subtests closest to core school subjects. Between 1948 and 2002, Information, Arithmetic, and Vocabulary gained very little. Let us compare their trends to those revealed by the Nation's Report Card from its inception until 2002, particularly trends on school-taught reading and mathematics.

The Nation's Report Card reports the performance of American students on tests designed by the National Association of Educational Progress (NAEP). The NAEP tests are administered to large representative samples of 4th, 8th, and 12th graders. From 1971 to 2002, 4th and 8th graders made a reading gain equivalent to 3.90 IQ points (SD = 15, Campbell, Voelkl, & Donahue, 2000, pp. 104 & 110; Grigg, Danne, Jin, & Campbell, 2003, p. 21). This puts them at the 60th percentile of their parents' generation. However, this happy result prepares us for disappointment. In the 12th grade, near high school graduation, the reading gain drops off to almost nothing (Campbell, et al., 2000, pp. 104 & 110; Grigg, et al., 2003, p. 21).

The IQ data suggest an interesting possibility. The WISC subtests show that from 1972 to 2002, schoolchildren made no gain in their store of general information and only minimal vocabulary gains (Table 2). Therefore, while today's children learn to master preadult literature at a younger age, they are no better prepared for reading more demanding adult literature. You cannot enjoy *War and Peace* if you have to run to the dictionary or encyclopedia every other paragraph. So by the time they leave secondary school, today's schoolchildren are no better off than the last generation. They open up an early lead but sometime after the age of 13 they hit a ceiling. And while they mark time between 13 and 17, the last generation catches up.

From 1973 to 2000, the Nation's Report Card shows 4th and 8th graders making mathematics gains equivalent to almost 7 IQ points. These put the young children of today at the 68th percentile of their parents' generation. But once again, the gain falls off at the 12th grade, this time to literally nothing (Braswell et al., 2001, p. 24; Campbell et al., 2000, pp. 54 & 60–61). And once again, a WISC subtest suggests why.

The Arithmetic subtest and the NAEP mathematics tests present a composite picture. An increasing percentage of young children have been mastering the computational skills the Nation's Report Card emphasizes at those ages. However, WISC Arithmetic measures both computational skills and something extra. The

questions are put verbally and often in a context that requires more than a times-table-type answer. For example, take an item like: 'If 4 toys cost 6 dollars, how much do 7 cost?' Many subjects who can do straight paper calculations cannot spontaneously see that you must first divide and then multiply. Others cannot do mental arithmetic involving fractions. In other words, WISC Arithmetic tests for the kind of mind that is likely to be able to reason mathematically.

My hypothesis is that during the period in which children mastered calculating skills at an earlier age, they made no progress in acquiring mathematical reasoning skills. Note the minimal gains registered on WISC Arithmetic (see Table 2: 1972 to 2002). Reasoning skills are essential for higher mathematics. Therefore, by the 12th grade, the failure to develop enhanced mathematical problem-solving strategies begins to bite. American schoolchildren cannot do algebra and geometry any better than the previous generation. Once again, although the previous generation was slower to master computational skills, they were no worse off at graduation.

In matching WISC subtests with the Nation's Report Card, we have assumed that score gains on the WISC subtests are relatively uniform from ages 6 to 16. Flynn (1984a) analyzed the literature on the WISC, WISC-R, and WAIS for ages 7 to 17, which cover the period from 1947 to 1972. Points gained rose slightly with age, that is, from 8.05 at age 7 to 8.81 at age 17. Sadly, no one has been sufficiently interested in subtests to accumulate a literature and no one has done a similar study since. Table 1 provides data for a rough comparison for the 1972 to 2002 period. The six comparisons that involve mainly schoolchildren of all ages give an average rate of gain of 0.272 points per year (numbers 2, 3, 4, 7, 9, and 10). The five comparisons that isolate those aged 16–17 give 0.380 points per year (numbers 1, 5, 8, 11, & 12). We doubt that either the pre-1972 or post-1972 era show a substantial escalation with age, and certainly recent data would have to be more robust and age-specific to reach such a conclusion. However, we can say this: what data exist show a pattern with age not in accord with the pattern of the Nation's Report Card. IQ gains have not declined as schoolchildren age.

EDUCATIONAL PROGRESS

The mixed record of skill trends among 12th graders does not mean that Americans have made no educational progress. Those who went on to complete either a BA or BS rose from about 13% in 1948, to 25% in 1972, and to about 33% in 2000 (Herrnstein & Murray, 1994; Neal, 2005). Presumably a university education means something in terms of mathematical skills and breadth of reading. Moreover, primary and secondary schools have helped enhance skills that provide the foundation on which universities build. Recall the huge gains on Raven's and Similarities that show that today's schoolchildren take on-the-spot problem-solving more seriously and are more comfortable with the vocabulary of science. These

advances prepare schoolchildren to cope with university education and no doubt, thanks to the schools, universities find it easier to strengthen these proclivities during the 4 years that lead toward graduation.

Looking to the future, analysis of IQ trends is beginning to pay-off in terms of reforming pedagogy. For example, primary schools have been teaching young children matrices under the delusion that on-the-spot problem-solving skills and arithmetic reasoning are functionally related. However, WISC subtest trends (Table 2) show that augmentation of the first skill (as shown by Similarities) has no direct effect on the second (as shown by Arithmetic). The relationship is actually correlational.

It looks very like what we concluded about the relationship between the 100-meter sprint and the high jump. Highly correlated at any given time but a minimal functional relationship in terms of the dynamics of actually performing in the two areas. Perhaps the greatest value of analyzing IQ trends will turn out to be diagnostic. They dissect the bundle of intercorrelated cognitive skills we call *g* into functionally autonomous components. The combination of IQ trends and achievement test trends is beginning to both write a history of the American mind and guide us toward a better future.

REFERENCES

Braswell, J.S., Lutkus, A., Grigg W., Santapau, S., Tay-Lim, B., & Johnson, M. (2001). *The nation's report card: Mathematics 2000* (NCES 2001–517), Washington, DC: U.S. Department of Education, Institute of Education Sciences.

Campbell, J.R., Voelkl, K.E., & Donahue, P.L. (2000). *NAEP 1996 trends in academic progress* (NCES 97–985r). Washington, DC: National Center for Education Statistics.

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

Dickens, W.T., & Flynn J.R. (2006). Black Americans reduce the racial IQ gap: Evidence from standardization samples. *Psychological Science*, *17*, 913–920.

Flynn, J.R. (1984a). IQ gains and the Binet decrements. *Journal of Educational Measurement, 21*, 283–290.

Flynn, J.R. (1984b). 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*(2), 171–191.

Flynn, J.R. (1998a). IQ gains over time: Towards 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. (1998b). WAIS-III and WISC-III: IQ gains in the United States from 1972 to 1995; how to compensate for obsolete norms. *Perceptual and Motor Skills, 86*, 1231–1239.

Flynn, J.R. (2000). IQ gains, WISC subtests, and fluid *g*: *g* theory and the relevance of Spearman's hypothesis to race (followed by Discussion). In G.R. Bock, J.A. Goode, & K. Webb (Eds.), *The nature of intelligence* (Novartis Foundation Symposium 233, pp. 202–227). New York: Wiley.

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

Flynn, J.R. (2006a). Efeito Flynn: Repensando a inteligência e seus efeitos [The Flynn Effect: Rethinking intelligence and what affects it]. In C. Flores-Mendoza & R. Colom (Eds.), *Introdução à Psicologia das Diferenças Individuais* [Introduction to the psychology of individual differences, J. Flynn, Trans.] (pp. 387–411). Porto Alegre, Brazil: ArtMed.

Flynn, J.R. (2006b). Tethering the elephant: Capital cases, IQ, and the Flynn effect. *Psychology, Public Policy, and Law*, *12*, 170–178.

Flynn, J.R. (in press). *What is intelligence? Beyond the Flynn effect.* Cambridge University Press.

Grigg, W.S., Daane, M.C., Jin, Y., & Campbell, J.R. (2003). *The nation's report card: Reading 2002* (NCES 2003–521). Washington, DC: U.S. Department of Education, Institute of Education Sciences.

Herrnstein, R.J., & Murray, C. (1994). *The Bell curve: Intelligence and class structure in American life*. New York: The Free Press.

Neal, D. (2005). *Why has Black-White skill convergence stopped?* NBER Working Paper 11090 (http://www.nber.org/papers/w110). Cambridge, MA: National Bureau of Economic Research.

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

Roid, G.H. (2003). *Stanford-Binet Intelligence Scales, fifth edition—Technical manual*. Itasca IL: Riverside.

Schneider, D. (2006). Smart as we can get? *American Scientist, 94,* 311–312.

Thorndike, R.L., Hagen, E.P., & Sattler, J.M. (1986). *The Stanford-Binet Intelligence Scale: fourth edition—Technical Manual*. Chicago: The Riverside Publishing Company.

Wechsler, D. (1981). *Wechsler Adult Intelligence Scale—Revised.* New York: The Psychological Corporation.

Wechsler, D. (1992). *Wechsler Intelligence Scale for Children—third edition.* (Australian adaptation). San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale—third edition*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2003). *The WISC-IV-Technical Manual*. San Antonio, TX, The Psychological Corporation.

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

Zimmerman, I.L., & Woo-Sam, J.M. (1997). Review of the criterion-related validity of the WISC-III: The first five years. *Perceptual and Motor Skills*, *85*, 531–546.

Copyright of International Journal of Testing is the property of Lawrence Erlbaum Associates and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.