Understanding the Flynn Effect

by Bob Williams

In order to make this material more readable, I have abandoned the scientific paper format that I initially tried. It simply became too tedious for a topic that must necessarily cover a lot of material. Where I have given the names of researchers, I have referenced only the first author, so please consider that an "et al." might be appropriate. Although I have removed the direct reference links (dates have been included where they have bearing on the sequence of events), there is a list of the referenced material.

Note that abbreviations are defined at the end of the text.

Background

The secular rise in IQ scores has presented a challenge to intelligence researchers since it was first noticed. Smith (1942) recorded a gain over a 14 year span. Later, Tuddenham (1948) found an increased intelligence when he compared inductee scores from World War I and World War II and proposed that the gains might be due to increased familiarity with tests; public health and nutrition; and education. [The gains from 1932 to 1943 were 4.4 points per decade.] He cited a high correlation (about .75) between years of education and the Army Alpha and Wells Alpha tests that he was studying.

The secular gain remained relatively dormant until it was rediscovered by Richard Lynn (1982) while working on a comparison of Japanese and U.S. data. It was then rediscovered again by James Flynn (1984). The raw score gains did not have a name until Herrnstein and Murray coined the term Flynn Effect in their book *The Bell Curve* (page 307). A rather sizable group of researchers choose to refer to the secular gain as the Lynn-Flynn Effect for the obvious reason that they feel Lynn has been somewhat slighted by not using his name. I recall hearing one paper presented on the topic in which the researcher creatively opted for the spelling "Flynn Effect".

Since the early 80s, researchers have looked and found the FE in virtually every group they have examined. They have published a huge number of papers (well over 100) on the gains and possible causes, but the results have been contradictory. In the following pages, I will review the characteristics of the FE, examine its possible causes, and address the important question as to whether there are real gains in intelligence or only score artifacts.

Gains

FE gains vary from country to country and over different time intervals, but the gains are usually a fraction of a point per year. As a matter of convenience, the gains are usually given as the number of points gained over a decade and written "ΔIQ". A few typical national gains:

U.S.	ΔIQ = 3
Estonia	ΔIQ = 1.65
Japan	ΔIQ = 7.7 (for those born from 1940 to 1965)

South Koreans gained at about the same rate as did the Japanese, but for those born between 1970 and 1990. A large number of researchers have reported FE gains in countries throughout the world, including both industrialized and third world nations. The number of countries showing a FE cannot be stated, since additions are frequently reported. Kanaya reported 20 nations, but that number is certainly low today.

Teasdale and Owen examined two samples of Danish draftees, consisting of 32,862 and 6,757 males. They found that the gains were concentrated mostly among the lower IQ levels and concluded that changes in the educational system were driving the score gains. They performed an interesting test, using Monte Carlo simulations to demonstrate that the FE gain distribution (low-end gains) were not caused by a ceiling effect in the test.

Other researchers, including Lynn and Colom, have found FE gains that were mainly concentrated in the lower IQ levels. This pattern suggests that the gains are related to improved environmental conditions associated with non-industrialized countries, rural areas, and low income.

Although it has now been 13 years since Jensen published *The g Factor*, his discussion of the FE remains current with respect to the items he considered. He reported these gains:

Raven's	ΔIQ = 5.69	
Wechsler	ΔIQ = 5.2	
	Performance	ΔIQ = 7.8
	Verbal	ΔIQ = 4.2

These show greater gains on the most abstract tests and subtests, although it is surprising to see the Wechsler as close to the Raven as the above numbers indicate — both being above the usually cited U.S. rate (ΔIQ = 3).

When Jensen examined subtests more closely, he found that nonscholastic test items showed increases at the same time (same test data sets) that scholastic items were decreasing. He noted that this is not what one would expect to see, but this is indeed what other researchers have reported. Jensen examined the SAT for the period 1952-1990 and found the well known decline. The usual explanation for the decline is that each year more students took the test and most of the additions to the pool of test takers were added below (lower intelligence) the prior group, leading to a decline at the mean. But Jensen corrected for the changes in demographics and showed that ¾ of the decline was due to the addition of more lower IQ testees, while the remaining ¼ was a real decline in scores. The ΔIQ loss for the SAT was -5 for the time period in question, while the FE gain was +3. This strongly suggests that the IQ test scores were not reflecting real world gains in intelligence.

The Raven's Progressive Matrices

The Raven tests have been cited frequently in the FE literature because most samples show particularly large gains on these tests (there are three: Colored, Standard, and Advanced progressive matrices). Intelligence researchers have been particularly drawn to the Raven tests because they are relatively easy to give, are as close to culture free as any, and do not load on the traditional group factors. Their high *g* loading makes them a de facto test of psychometric *g*. The Raven tests have shown gains of 18-20 IQ points per generation in many industrialized countries. Dutch gains were 21 points over 30 years (ΔIQ = 7). Urban Chinese gained 22 points between 1936 and 1986.

Hiscock reported a higher rate of FE gains for the Raven's Matrices than for the Wechsler and Stanford-Binet tests. He also showed that Raven scores for birth years from 1877 to 1967 increase steadily, but roll off over that time span to a possibly flat (no effect) rate for the last 10 year interval. Lynn has argued that the Raven tests are being inflated as a result of mathematical education. [The Raven's requires some use of addition, subtraction, progression and the distribution of values.] However, the applications of simple math does not seem to be a factor in the Colored or Standard tests.

Academic performance down

While IQ test scores have been rising (in some cases soaring), academic performance has done the opposite. As Jensen pointed out, when he observed that the SAT and subtests of scholastic test items have declined, real world academic performance has done the same.

Philip Adey (King's College) studied the test scores of 25,000 children across both state and private schools and concluded: "The intelligence of 11-year-olds has fallen by three years' worth in the past two decades. In 1976 a third of boys and a quarter of girls scored highly in the tests overall; by 2004, the figures had plummeted to just 6% of boys and 5% of girls. These children were on average two to three years behind those who were tested in the mid-1990s."

For an assessment of how well U.S. students are doing, this URL leads to a well written, if depressing, description of the state of teaching, education, and students: www.lhup.edu/~dsimanek/decline1.htm

Estonia

Thanks to the work done by Olev and Aasa Must, there is a good bit of information about the FE as it has manifested itself in Estonia. The messages from their studies are that the FE gains follow different trajectories in different countries and the factors driving those changes are also different.

In the Estonian studies, subtests that needed computation skills and mathematical thinking were unchanged over 60 years. The information subtest declined; verbal subtests showed moderate gains; but there were impressive gains in symbol-number and comparison subtests.

Must examined data over a 72 year span and found a relatively small ΔIQ of 1.65. But when the eight years from 1998-2006 are examined separately, the ΔIQ almost doubles to 3 points. The *g* factor loadings were different at the subtest level for each of the three birth cohort groups examined, but the greatest difference was between the oldest cohorts compared to the other two relatively recent cohorts.

Large WISC gains were observed in arithmetic, information, and vocabulary. These gains are opposite from score changes seen in the U.S. and Britain. The authors identified several possible causes: greatly improved education, better nutrition, better health care, and changes in demographic behavior (smaller families).

South Africa

ΔIQ = 3.63	Whites	(same group took two different test batteries)
ΔIQ = 1.57	Indians	(same group took two different test batteries)

The FE score gain is stronger for the Afrikaans speakers than for the English speakers. [te Nijenhuis]

Gains seen in young children

British children aged 6 and 18 months displayed large gains over the period from 1949 to 1985. When measured on the Griffiths Test, developmental quotients (DQ) gained 2.45 points per decade. Similar studies (using the Bayley Mental Scales) were done by other researchers in the U.S. and Australia and show gains of 2.9 DQ points per decade.

Flynn reported ΔIQ = 3.9 as the mean of 14 studies of children from age 4 to 6. These DQ and IQ gains show a FE that is as large in preschool children as in adults, making education an unlikely explanation for the cause (at least in the data sets examined). Similarly, Kanaya reported that elementary school children (WISC test) show FE gains that are similar to adult gains on the WAIS.

As is already apparent, FE findings in one place do not generalize globally. Cotton reported no FE, using the Raven's Colored Progressive Matrices, for a group of Australian children ages 6-11 from 1975 to 2003.

Low-end versus high-end gains

As previously mentioned, Teasdale and Owen found that FE gains for Danish draftees were concentrated in the lower end of the intelligence spectrum, suggesting a cause or causes such as improved nutrition, better health care, or increased education. One example of lower end gains can be seen in the following (from Colom, 2005):

Percentiles	1970 Raw scores	1999 Raw scores	Difference
01	30	39	9
05	37	45	8
15	42	49	7
25	45	51	6
35	48	53	5
45	50	54	4
65	53	57	4
85	58	61	3
95	62	64	2
99	66	67	1

The last column (score difference), shows that the raw scores decrease monotonically with increasing percentiles. The gains are obviously greater at the lower end. Colom also noted that FE gains were much greater on the Raven's Standard Progressive Matrices (19.2 points over 28 years) than on the Advanced Progressive Matrices (6.75 points over 28 years). He concluded that the cause of the increases probably had a greater impact in the low and medium segments of the intelligence distribution.

Lynn reported a low-end gain that was about double the high-end gain, for a British group over the period 1932 to 1982. Similarly, Kağıtçıbaşı found greater gains at the low end, over the period from 1977 to 2010. The differences were particularly large (23 points) for remote villages. Within urban locations, the lower SES groups also showed more gains (7.4 points) than higher SES groups, but these were less than in the remote villages.

The FE is so specific that for every finding, there seems to be an opposite finding. Flynn claimed IQ gains at "every level" based on his observation that "score variance remains unchanged over time". Colom examined data for Brazilian children covering a span of 72 years. He found that the FE gains were greater for urban samples than for rural samples and concluded: "whatever the causes of the increase, they act more intensively for more intelligent children"[1]. This finding for the Brazilian samples is opposite of Colom's finding for Spanish samples.

Ang computed FE gains from the National Longitudinal Study of Youth (NLSY) data. This is particularly interesting because of the high quality of that source

[1] There has been a consistent finding in studies of rural and urban samples that the urban samples have higher IQ. In fact, both Colom and Kağıtçıbaşı and Biricik found that IQ decreased as a function of the distance from the nearest city.

(includes multiple generations). The data include scores from the Peabody Individual Achievement Test (PAIT); the math portion was deemed to be closest to fluid intelligence. In this instance, the gains were skewed towards higher educated and higher income families (these are higher IQ). Only the PAIT-math showed FE gains, which the authors believe is difficult to explain by a nutrition hypothesis. This study showed no race or sex related differences in FE gains.

Right tail gains

Only one study examined the FE in a data set that is limited to very high IQ individuals. Wai examined the huge (1.7 million scores) data set of 7th grade students who took the SAT and ACT and 5th and 6th grade students who took the EXPLORE test. These tests are given to students who have scored in the top 5% for their grade on a standardized test (composite or subtest), and are part of the Duke TIP 7th Grade Talent Search.

As previously discussed, the FE has sometimes been shown to be skewed to the lower half and sometimes skewed to the upper half. Flynn (1986) argued that the gains were present at all levels (see previous comment), but did not have data specific to the high range that is usually considered as gifted. Wai found the following generational IQ gains in the top 5%:

5.1	SAT-M
13.5	ACT-M
11.1	EXPLORE-M

The gains were concentrated on math and nonverbal subtests (see earlier comments on Ang, 2010).

Wai also examined SAT-M scores of 500 and above (top 0.5%) and equivalent scores for the ACT, with the following results:

SAT-M 1981-1985,	7.7%	at or above 500
2006-2010,	22.7%	at or above 500
ACT-M 1990-1995,	17.7%	at or above a similar level
2006-2010,	29.3%	at or above a similar level

The obvious conclusion is that either there are a lot more truly bright children in the 2006-2010 set, or the test results are showing a significant score inflation that is not merited. Wai also used multigroup confirmatory factor analysis to determine whether the data sets were invariant with respect to cohort; they were

not. Consequently, it can be concluded that something changed in the test construct from one cohort to the other.

Hypothetical causes

Among the causes that have been proposed to explain the FE are these:

Education	Decreased family size
Increased exposure to testing	Heterosis
Exposure to artificial light	More complex visual environment
Nutrition	Child rearing practices

and the use of Classical Test Theory versus Item Response Theory.

Education

Since FE gains have been observed in preschool children, education is unlikely to be a cause in all data sets. As previously discussed, FE gains have been more pronounced on non-scholastic items and scholastic subtests have even demonstrated lower scores at the same time and within the same tests that show FE gains in non-scholastic subtests. Direct measures of academic performance (see previous discussion) have also shown secular declines while FE gains were evident in IQ tests. Lynn attributes gains in the Raven as the result of mathematical education.

But as previously noted, the simple math involved does not seem to have any influence on Raven's scores.

As has already been shown, FE gains are inconsistent from one place to another. It is possible, and even likely, that gains in less developed countries have been at least partially driven by improved education, even if education is not a factor in industrialized nations.

Increased exposure to testing

There is little doubt that testing has increased over the past years. Tuddenheim listed it as one possible explanation for the secular gains he found between WWI and WWII cohorts. There are two mechanisms that have been proposed. Brand suggested that the use of timed tests has caused students to work faster by guessing more frequently (multiple choice). While this may be a factor, FE gains are seen on tests that are untimed and on tests that do not use multiple choice.

Jensen mentioned "increasing test wiseness from more frequent use of tests". His point was that frequent testing may have the same sort of impact on test scores as the increase associated with test-retest. This is the same process that is associated with learning and shows up in situations where test training has been used (as is common with the SAT). When this happens, the test *g* loading decreases and its *s* loading[2] increases.

Both Brand's and Jensen's ideas would presumably cause test scores to increase without showing gains on *g*. As will be seen later, numerous studies have shown that FE gains that are not *g* loaded.

Nutrition and medical care

Both nutrition and medical care have improved over the past century and have been accompanied by a large number of gains that appear to be caused by these improvements: increased mean height, increased head size, faster growth, earlier maturation, etc. Lynn argues that gains in developmental quotients (DQs — hold up head, sit up, stand, walk, jump, etc.) are indicators of gains in IQ. DQs have gained 3.7 points per decade, while IQ gains of 3.9 points per decade have been seen in preschool children (age 4-6). Using the Griffiths Test, British children at age 6 months showed an average DQ gain of 2.8 points per decade and children, age 18 months, showed an average gain of 2.1 points per decade. Flynn and Bocerean have reported IQ gains that are similar to the DQ gains for preschool children.

Lynn cites various studies that show poor nutrition in the early part of the 20th century in the U.S., Britain, Spain, and Sweden. Those indications of poor nutrition disappeared over the course of that century. Three nutrients that are known to be related to the development of intelligence are iron, folate, and iodine. Lynn presented references showing insufficient intake of these in various countries in the early part of the 20th century.

The studies that have shown greater FE gains in the lower part of the IQ distribution are consistent with the nutrition argument. Presumably the people affected most by poor nutrition were those at the lower end of the intelligence spectrum because this group typically shows lower income and lower SES. Lynn gives this example of FE gains (averages of 9 age groups) from 1979 to 2008 on the Raven's Standard Progressive Matrices:

[2] Spearman's *s* is the specificity of a test. It is a factor that is unique to the test. Consequently, the *s*-loading is simply the correlation between a factor and the test specificity.

Percentile	5	25	50	75	95
Gain	5	4	3	2	0

And, for the Raven's Colored Progressive Matrices from 1982 to 2007:

Percentile	5	25	50	75	95
Gain	3.5	3.2	3.1	2.5	1.6

Birth weights

One factor influencing birth weight is pre-natal nutrition. Birth weight correlates positively with IQ and with DQs. Brazelton reported that when birth weights reached 3500 grams, infants were advanced by approximately 15 DQ points at age 28 days (compared with lower birth weight babies). Low birth weights show the opposite; Drillien reported DQ score depressions of 12 points for infants with birth weights under 2000 grams, compared to those with birth weights over 2500 grams (ages 6 months through 2 years). Various other studies have reported similar findings. In general, improved pre-natal nutrition increases birth weights and head size. [Birth weight is correlated with head size at r = 0.75.] It is head size that is directly linked to higher cognitive performance.

Jensen found that head size is mostly correlated with *g* (as opposed to group factors) and notes that the reason for the correlation is that head size is a proxy for brain size. When measured with MRI, the correlation between brain size and IQ is about 0.40. Larger brain size means more neurons and is logically consistent with the correlations between head and brain measurements versus IQ. Lynn cited numerous sources that have reported head size increases of about one standard deviation over the past 50-plus years.

Height

Lynn attributes the change in height and in DQs as being caused by nutritional improvements. Both measures increased by about one SD over 50 years. Flynn, however, countered that gains in height have not happened at the same times as gains in IQ. This argument seems to imply a degree of data tracking, with respect to time, that is not necessary for the argument to hold.

Head size

In Britain, the head circumference of 1 year olds has increased by approximately 1.5 cm from 1930 to 1985. Head circumference, DQs, IQs, and height, over that

time span, have all shown gains of about 1 SD. Head size is an approximate measure of brain size; the two correlate at r = 0.8.

The correlation between brain volume and IQ is presumably due to the larger number of neurons in larger brains, although E. Miller has suggested that it may be due to higher levels of myelination in larger brains. In any case, increases in brain size should be direct contributors to higher intelligence.

Not nutrition

For most proposed causes of the FE, there is both supporting and opposing data:

- *The Rising Curve* (American Psychological Association) notes that studies of nutrition have shown that neither vitamins nor supplements have had any impact on intelligence.
- Nutrition is unlikely to have declined over the past 20 years in those countries that have a negative FE. Height did not decline in those countries.
- Contrary to the intelligence gains seen in Norway, height gains from 1969 to 2002 were mostly in the upper half of the intelligence range.

Exposure to artificial light

This hypothesis is not seen often in the literature and might have been omitted in this review, except that it did not come from a weak source, but was one of the items listed by Jensen in *The g Factor*. The idea is based on the response of the pineal gland in animals to artificial light. The pineal gland appears to play a major role in sexual development, hibernation, metabolism, and seasonal breeding. The effect of stimulating growth is used by poultry farmers to increase their output. A quick search of the Internet yields numerous papers on the optimal use of artificial light for this purpose.

There does not seem to be any data available for whether this effect happens in humans, but the speculation is that it might. There has been an obvious increase in the use of electric lighting by humans over much of the time that the FE has been observed. Besides lighting, people have been increasingly exposed to artificial light from television and computer screens, even during early childhood.

Decreasing family size

It has been known for some time that the mean IQs of families decreases as family size increases. There are two factors that contribute (presumably independently)

to this effect:

1. Maternal IQ correlates negatively with fertility. This is the underlying factor behind Richard Lynn's papers and book relating to global dysgenics and has been shown for numerous data sets from various countries. Low IQ people statistically have more children than high IQ people. The high heritability of intelligence, therefore, is a source of dysgenic pressure. If the average family size decreases, the reduced numbers of low IQ children should produce a net increase in the mean, which would show up as a FE gain.
2. Dating as far back as Sir Francis Galton, it was believed that IQ declined as a function of birth order. That belief was disputed by J. L. Rodgers after he examined the NLSY and did not find a birth order effect. His argument seemed strong and held until Kristensen published papers based on the very large data set of Norwegian conscripts, which showed the birth order effect. The mechanism of the effect has not been resolved. The hypotheses that have been advanced include prenatal gestational factors and social factors. The former seem more consistent with the general finding that social factors have little, if any effect on intelligence. Causation of the birth order effect does not matter when discussing the FE. If family size is declining in various groups, there must be a positive contribution to mean IQ due to fewer low IQ children being born. [It is important to note that the birth order effect is robust, but not of high magnitude.]

Heterosis

Mingroni has argued that since the effects of the environment[3] (on intelligence) are so small, the possibility of a genetic effect should be investigated. If environmental factors were significant, between-family variance would cause MZA twins to be less alike and siblings to be more alike.

Besides IQ, there have been secular trends in such things as height, growth rate, myopia, asthma, autism, ADHD, and head circumference. It may, therefore, seem reasonable to argue that there is a global change that is affecting some or all of these factors (possibly consistent with Lynn's nutrition hypothesis). If selective breeding were involved, Jensen claimed that, in order to produce the magnitudes seen in the FE, breeding would have to be restricted to only those

[3] See Loehlin, et al. (1989) and Scarr, et al. (1978).

people in the upper half of the IQ distribution. As previously discussed, it is the bottom half that has the higher fertility.

Lynn argued that heterosis is unlikely for three reasons:

1. There was little immigration in Europe before 1950 (the FE was present before that date).
2. The FE for IQs and DQs are just as large in Europe as in other places.
3. Studies of heterosis have shown little positive effect on IQ.

Perhaps the most important consideration in determining whether there is a heterosis effect was pointed out by Mingroni: If the FE is found within-families, the cause is not genetic. Sundet found that the FE operates within sibships. His analysis draws on two findings: the birth order effect (previously discussed), and the up and down direction of the FE in Norway. Given those observations, the following findings show that the FE operates within sibships:

FE increasing Intelligence difference between brothers decreased with increasing age differences.

FE decreasing Difference between the later-born and the earlier-born brother increased across age differences.

FE absent No change as a function of age difference.

Sundet: "Despite the Flynn Effect, later-born brothers show lower IQ, establishing a birth order effect in Norway."

Unless Sundet's finding cannot be extended beyond Norway, the heterosis hypothesis must be dropped. *The Rising Curve*: "[There is] No evidence that within family environmental influences have anything but a minimal impact on intelligence. What environmental influences there may be are between families."

Enriched visual environment

Greenfield and others suggested that the FE gains are caused by the ever increasing shift from verbal communication to visual and interactive media. This is seen globally in the increased presence of movies, television, photography, video games, computers, puzzles, mazes, exploded views, etc. Advertising has become ubiquitous and is saturated with images, graphs, charts, and rapid sequence visuals.

The mechanism for this hypothesis is that the shift towards visual representations removes some of the novelty from tests, especially the culture reduced tests that have shown about double the FE gains as found in other tests. This is particularly convincing for tests such as the Raven's, which presents abstract figures in a matrix. Several decades ago these figures may have been more baffling than they are today. If this factor is operative, it most likely falls in the category of learning, which causes test *g* loadings to decrease (see Jensen, 1998).

Child rearing practices

The FE has been seen throughout the world, in both developed and undeveloped countries where child rearing practices certainly vary greatly. It is unlikely that this hypothesis is a significant factor, not only because of the cultural variation in child rearing practices, but also because the shared environment has essentially no impact on adult intelligence (per prior discussion). To some extent, this category overlaps the increased visual environment and education. In that regard, it may contribute to the FE in some instances.

Is the FE invariant?

As previously noted, some researchers have tested for invariance and have found that the data sets they were examining were not invariant (see Wai and Must). Jelte Wicherts did a study of five data sets to determine if they were invariant. These included the Must and Teasdale studies previously mentioned. Multigroup confirmatory factor analyses of these data sets showed that they were not invariant, meaning that FE gains were not gains on the latent variables that the tests were supposed to measure. Besides providing insight as to the nature of the FE gains, the rejection of factorial invariance demonstrates that subtest score interpretations are necessarily different over time.

This finding, and its confirmation from other independent studies may well be one of the most important aspects of understanding the FE. The things that are tested at one time cannot be regarded as presenting the same construct to an earlier or later group.

When multi-group confirmatory factor analyses has been used to test for invariance, racial differences typically show invariance, meaning that the scores have similar meanings for the different groups (see Dolan). But in the case of the FE, invariance is typically absent when age cohorts are compared. This means that the tests have different meanings for the cohort groups (see Wicherts).

Classical Test Theory versus Item Response Theory

Alex Beaujean did an analysis that is related to the Wicherts analysis of invariance, in that it examined the nature of the test construct itself. Most studies in the literature are based on CTT and are presented without passing along the test item data. This practice hides some of the information that could be extracted from a data set. Test scores are given, but the latent constructs they are designed to measure cannot be examined. IRT, on the other hand, allows the researcher to examine the changes in underlying latent ability. Thus, CTT can show differences in scores, even when there is no change in the latent variable. An increase may be due to a general gain in real intelligence, or a decrease in the levels of difficulty of test items.

IRT is generally considered to be the better methodology, despite its relatively infrequent use. It is particularly useful in FE studies because it reveals changes in item properties between two groups measured at different times. CTT requires groups that are being compared to have similar ability distributions, but this is not a requirement when IRT is used. In IRT, the item parameters do not depend on the ability level of the testees.

Beaujean's results using CCT and IRT to measure FE gains:

PPVT-R		
	CCT	0.44 points per year
	IRT	0.06 points per year
PIAT-M		
	CCT	0.27 points per year
	IRT	0.13 points per year

Rodgers and Wanstrom (2007) replicated Beaujean findings (with NLSY data).

The results clearly show that the FE essentially vanishes for the PPVT-R when IRT is used. The PIAT-M gains are cut to half using IRT. Ergo, the FE gains are determined by the methodology, leading to the concern that much of the literature has reported findings that might be quite different if IRT had been used.

Real or hollow gains?

When David Wechsler studied his WAIS, he gave the old 1953 version and the new 1978 version (WAIS-R) to the same group. That group averaged 103.8 on the new version and 111.3 on the old version (ΔIQ = 3).

If children of 1997 took the 1932 test, ¼ would score above IQ 130 (an increase of 10x). Or, if children of 1932 took the 1997 test, the mean would be about 80! ¼ would be "deficient". [see *The Rising Curve*]

Vroon made a similar observation about Dutch men: When scored against 1982 norms, men in 1952 would have had a mean IQ of 79.

Flynn initially questioned the reality that intelligence has increased:

> "Has the average person in The Netherlands ever been near mental retardation?" "Does it make sense to assume that at one time almost 40% of Dutch men lacked the capacity to understand soccer, their most favored national sport?" He noted that there are not more gifted Dutch school children now and that patented inventions have shown a sharp decline.
>
> The U.S. mean in 1918 would have been 75, if scored against today's norms.
>
> If the score gains were real intelligence gains, real-life consequences would be conspicuous.

Is the Flynn Effect a Jensen Effect?

[A Jensen effect is one that loads on *g*. It was named by Rushton.]

Roberto Colom (paper title): The secular increase in test scores is a "Jensen Effect".

Olev Must (paper title): The secular rise in IQs: In Estonia, the Flynn effect is not a Jensen effect.

Rushton and Jensen (from paper): The Flynn Effect is not a Jensen Effect (because it does not occur on *g*).

Not a Jensen Effect

In a related study, Jante te Nijenhuis did a meta-analysis of 64 test-retest studies using IQ batteries (total N=26,990). He found a correlation between *g* loadings and score gains of -1.00; a similar finding was reported by an independent meta-analysis by van Bloois. Olev Must found (in Estonia) a correlation of -0.40 between *g* and FE gains. These all show that the gains were not on *g* and were, therefore, hollow.

Rushton and Jensen showed that heritabilities calculated from twins also correlate with the *g* loadings, r =0.99, P<0.001 (for the estimated true correlation),

providing biological evidence for a true genetic *g*. The importance of this is that if the FE is being driven by environmental factors, it is unlikely that the gains would load on *g*. If the cause is genetic (as in the Mingroni hypothesis), the gains should show a Jensen effect.

They also pointed out that *g* loadings and inbreeding depression scores on the 11 subtests of the WISC correlate significantly positively with racial differences and significantly negatively (or not at all) with the secular gains. This is further evidence that the FE is caused by environmental factors.

Perhaps the strongest argument that the FE does not load on *g* came from Rushton (1999). He used principal components analysis to show the independence of the FE from known genetic effects.

1. The IQ gains on the WISC-R and WISC-III form a cluster. This means that the secular trend is a reliable phenomenon.
2. This cluster is independent of the cluster formed by racial differences, inbreeding depression scores (purely genetic), and *g* factor loadings (largely genetic). The secular increase is, therefore, unrelated to *g* and other heritable measures.

Although it is possible that these findings apply in all cases, they may not. As has already been shown, different data sets (from different countries or different times) often produce opposite results.

For the previously mentioned Estonian study, Must used the Method of Correlated Vectors (see Jensen, 1998) to test the FE gains for *g* loading. Rank order correlations between the various subtests and the rank of those subtests on the *g* factor were negative and nonsignificant: r= -.40 (one-tailed P=.13). Subtests with the lowest *g* loadings showed the greatest FE gains. Must concluded: "In Estonia, the Flynn effect is not a Jensen effect."

Yes, it is a Jensen Effect

Roberto Colom believes the data he has examined shows gains on *g*. Colom: "Not a 'Jensen effect' is true for crystallized tests but not for fluid tests." I had the opportunity to ask Colom about his position that there was a gain in *g*. His explanation was centered on his finding that the gains on G_c were smaller than the gains on G_f and his belief that G_f is essentially the same as *g*. But his argument was not entirely subjective. Using the DAT, Colom showed that subtest gains increased as their rank order of *g* loading increased. [The subtests in the

DAT are (in order of g loading from lowest) numerical ability, verbal reasoning, mechanical reasoning, abstract reasoning, and spatial relations.]

Predictive bias

Jensen commented that the definitive test of whether FE gains are hollow or not is to apply the predictive bias test. This means that two points in time would be compared on the basis of an external criterion (real world measurement, such as school grades). If the gains are hollow, the later time point would show underprediction, relative to the earlier time. This assumes that the later group has not been renormed. [Earlier IQ points would exceed the performance of the later generation for the same IQ.] In actual practice tests are periodically renormed so that the mean remains at 100. The result of this recentering is that the tests maintain their predictive validity, indicating that the FE gains are indeed hollow. If the gains were real and the tests were renormed, people at a given IQ would be getting smarter and this would show up in the predictive validity.

Which explanations work?

Most of the mechanisms that have been proposed as causes of the FE are plausible under some circumstances. Even when one is ruled out by a specific study, it may apply elsewhere. As has been shown in the foregoing material, the most consistent aspect of the FE is that it is inconsistent from one time or place to another. Sometimes the gains have been mostly in abstract reasoning (as in the U.S.), but elsewhere the gains have been strongly tilted towards scholastic subtests (Estonia). Gains have been strong, weak, flat, or have reversed (even within the same country when measured at different times — Norway and Denmark). Of the items discussed, the ones that might be most questionable are exposure to artificial light (lacking human data), heterosis (because a within-family FE has been reported), and child rearing practices (no data showing that there is an independent effect here and the established finding that the family environment has no impact on adult IQ).

While in a three-way conversation with Jim Flynn and Ted Nettlebeck, I asked them what they thought of Jensen's suggestion that the observed FE is caused by small contributions from several components. Both men immediately agreed that this must be true.

Finally, there is the large issue of lacking invariance and methodological incon-

sistency when IRT is used instead of CCT. The instances in which confirmatory factor analysis have failed to show invariance (every case so far) tells us that the meaning of IQ tests is not constant over time, leading to score changes that are not *g* loaded. The reduction in FE magnitude (to near zero in some cases) when IRT is applied suggests that the test vehicle is contributing 50 to 100% of the gains and that those gains are methodological artifacts and carry no *g* loading.

Real or hollow?

Most of the tests for *g* loading have shown little or no *g* saturation. Most of the researchers who have addressed the issue have argued that the gains are hollow, with the exception of Lynn and Colom, both of whom have made strong arguments that there is at least some genuine gain in intelligence. This inconsistency may be due in part to different data sets and may be due in part to their failure to use IRT methods. The best guess is that most of the FE gains that have been reported are hollow. If this were not true, renorming would cause predictive validity to change, but there are no reports that this has happened.

Summary

- The FE exists between birth cohorts.
- It is found within sibships.
- It appears early in life (before school age).
- There are presumably multiple causes.
- The gains are all or mostly hollow (not Jensen Effects).
- There are serious methodological issues to be resolved and which may be a major cause of the gains.
- The FE is not invariant over time.

Abbreviations / Definitions

CTT Classical Test Theory
DAT Differential Aptitude Test
DQ Developmental quotient
Duke TIP Talent Identification Program
FE Flynn Effect
g Psychometric *g* (the general factor that emerges from a factor analysis)
G_c Crystalized intelligence (as a second stratum factor)
G_f Fluid intelligence (as a second stratum factor)
IRT Item Response Theory
MZA Monozygotic twins reared apart
NLSY National Longitudinal Study of Youth
PIAT-M Peabody Individual Achievement Test-Math
PPVT-R Peabody Picture Vocabulary Test-Revised
SES Socioeconomic status
SD Standard deviation
WAIS Wechsler Adult Intelligence Scale
WISC Wechsler Intelligence Scale for Children

References

Ang, S., Rodgers, J., & Wänström, L. (2010). The Flynn Effect within subgroups in the U.S.: Gender, race, income, education, and urbanization differences in the NLSY-Children data. *Intelligence*, 38-4, 367-384.

Bayley, N. (1993). *Bayley Scales of Infant Development.* San Antonio, TX: Psychological Corporation.

Black, M. M., Hess, C. R., & Berenson-Howard, J. (2000). Toddlers from low-income families have below normal mental, motor and behavior scores on the revised Bayley Scales. *Journal of Applied Developmental Psychology*, 21, 655-666.

Bocerean, C., Fischer, J. -P., & Flieller, A. (2003). Long term comparison (1921–2001) of numerical knowledge in 3 to five and a half year old children. *European Journal of Psychology of Education*, 18, 405-424.

Brand, C. (1996). *The g Factor: General Intelligence and Its Implications.* Chichester, England: Wiley.

Brazelton, T. B., Tronik, E., Lechtig, A., Lasky, R. E., & Klein, R. E. (1977). The behavior of nutritionally deprived Guatemalan infants. *Developmental Medicine and Child Neurology*, 19, 364-372.

Broman, S. H., Nichols, P. L., & Kennedy, W. A. (1975). *Preschool IQ: Prenatal and Developmental Correlate.* Hillsale, NJ: Wiley.

Campbell, S. K., Siegel, E., Parr, C. A., & Ramey, C. T. (1986). Evidence for the need to renorm the Bayley Scales of Infant Development based on the performance of a population-based sample of 12 month old infants. *Topics in Early Childhood Education*, 6, 83-96.

Carlson, J. S., & Jensen, C. M. (1980). The factorial structure of the Raven Coloured Progressive Matrices Test: A reanalysis. *Educational and Psychological Measurement*, 40, 1111–1116.

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.

Colom, R., Flores-Mendoza, C. E., & Francisco J. Abad, F. J. (2007). Generational Changes On The Draw-A-Man Test: A Comparison Of Brazilian Urban And Rural Children Tested In 1930, 2002 AND 2004. *Journal of Biosocial Science*, 39, 79–89.

Colom, R., Juan-Espinosa, M., & Garcia, L. F. (2001). The secular increase in test scores is a "Jensen effect." *Personality and Individual Differences*, 30 553-559.

Cotton, S. M., Kiely, P. M., Crewther, D. P., Thomson, B., Laycock, R., & Crewther, S. G. (2005). A normative and reliability study for the Raven's Colored Progressive Matrices for primary school aged children in Australia. *Personality and Individual Differences*, 39, 647-660.

Dolan, C. V., & Hamaker, E. L. (2001). Investigating Black–White differences in psychometric IQ: Multi-group confirmatory factor analyses of the WISC-R and K-ABC and a critique of the method of correlated vectors. In F. Columbus (Ed.), *Advances in Psychology Research*, vol. 6. (pp. 31-59) Huntington, NY: Nova Science Publishers.

Drillien, C. M. (1969). School disposal and performance for children of different birthweight born 1953–1960. *Archives of Diseases in Childhood*, 44, 562-570.

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, 171- 191.

Greenfield, P. M. (1998). The cultural evolution of IQ. In U. Neisser (Ed.), *The rising curve: Long-term gains in IQ and related measures* (pp. 81–123). Washington, DC: American Psychological Association.

Hanson, R., Smith, J. A., & Hume, W. (1985). Achievements of infants on items of the Griffiths scales: 1980 compared with 1950. *Child: Care, Health and Development*, 11, 91-104.

Herrnstein, R. J., & Murray, C. (1994). *The Bell Curve: Intelligence and Class Structure in American Life.* New York: Free Press.

Hiscock, M. (2007) 'The Flynn effect and its relevance to neuropsychology', *Journal of Clinical and Experimental Neuropsychology*, 29:5, 514- 529.

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

Kağıtçıbaşı, C., & Biricik, D. (2011). Generational gains on the Draw-a-Person IQ

scores: A three-decade comparison from Turkey. *Intelligence* 39, 351–356.

Kanaya, T., Ceci, S. J., & Scullin, M. H. (2005). Age differences within secular IQ trends: An individual growth modeling approach. *Intelligence* 33, 613-621.

Loehlin, J. C., Horn, J. M., & Willerman, L. (1989). Modeling IQ change: Evidence from the Texas Adoption Project. *Child Development*, 60, 993–1004.

Lynn, R. (1982). IQ in Japan and the United States shows a growing disparity. *Nature*, 297, 222–223.

Lynn, R., & Hampson, S. (1986). The rise of national intelligence. Evidence from Britain, Japan and the United States. *Personality and Individual Differences*, 7, 23–32.

Lynn, R., & Hampson, S. (1989). Secular increases in reasoning and mathematical abilities in Britain, 1974–1982. School. *Psychology International*, 10, 301–304.

Lynn, R. (1990). The role of nutrition in secular increases in intelligence. *Personality and Individual Differences*, 11(3), 273–285.

Lynn, R. (1993). Nutrition and intelligence. In P. A. Vernon (Ed.), *Biological approaches to the study of intelligence*. Norwood, NJ: Ablex.

Lynn, R. (1996). *Dysgenics: Genetic Deterioration in Modern Populations*. Praeger Publishers.

Lynn, R. (1998). In support of nutrition theory. In U. Neisser (Ed.), *The rising curve*. Washington, DC: American Psychological Association.

Lynn, R., & Harvey, J. (2008). The decline of the world's IQ. *Intelligence*, 36 112 – 120.

Lynn, R. (2009a). What has caused the Flynn effect? Secular increases in the Development Quotients of infants. *Intelligence*, 37 (2009a) 16–24.

Lynn, R. (2009b). Fluid intelligence but not vocabulary has increased in Britain, 1979–2008. *Intelligence*, 37, 249–255.

Miller, E. M. (1994). Intelligence and brain myelination: A hypothesis. *Personality and Individual Differences*, 17, 803-832.

Miller, G. F., and Penke, L. (2007) The evolution of human intelligence and the coefficient of additive genetic variance in human brain size. *Intelligence*, 35, 97-114.

Mingroni, M. A. (2004). The secular rise in IQ: Giving heterosis a closer look. *Intelligence*, 32 65–83.

Must, O., Must, A. & Raudik, V. (2003). The secular rise in IQs: In Estonia, the Flynn effect is not a Jensen effect. *Intelligence* 31, 461–471.

Must, O., te Nijenhuis, J., Must, A., & van Vianen, A. E. M. (2009). Comparability of IQ scores over time. *Intelligence* 37, 25-33.

Neisser, U. (1998). *The rising curve.* Washington 7 American Psychological Association.

Nijenhuis, J. T., van Vianen, A., & van der Flier, H. (2007). Score gains on g-loaded tests: No g. *Intelligence*, 35 283-300.

Nijenhuis, J. T., Murphy, R. & van Eeden, R. (2011; in press). The Flynn effect in South Africa. *Intelligence.*

Nijenhuis, J. T., Cho, S. H., Murphy, R., Lee., K. H. (2011; in press). The Flynn effect in Korea: Large gains. *Personality and Individual Differences.*

Read more: Family Size - Child, Intelligence, Siblings, and Relationships - JRank Articles social.jrank.org/pages/251/Family-Size.html#ixzz1cSlJTIK2

Rodgers, J. L., H. H. Cleveland, E. van den Oord, and D. C. Rowe (2000). Resolving the Debate over Birth Order, Family Size, and Intelligence. *American Psychologist*, 55 599-612.

Rushton, J. P. (1999). Secular gains in IQ not related to the g factor and inbreeding depression—unlike Black–White differences: A reply to Flynn. *Personality and Individual Differences*, 26, 381-389.

Rushton, J. P. (2000). Flynn effects not genetic and unrelated to race differences. *American Psychologist*, 55, 542 - 543.

Rushton, J. P. & Jensen, A. R. ((2010). The rise and fall of the Flynn Effect as a reason to expect a narrowing of the Black-White IQ gap. *Intelligence*, 38 213-219.

Scarr, S., & Weinberg, R. A. (1978). The influence of ''family background'' in intellectual attainment. *American Sociological Review*, 43, 674–692.

Smith, S. (1942). Language and nonverbal test performance of racial groups in Honolulu before and after a 14-year interval. *Journal of General Psychology*, 26, 51 - 92.

www.timesonline.co.uk/tol/news/article721863.ece *The Sunday Times*, January 29, 2006.

Sundet, J. M., Barlaug, D. G., & 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*, 33, 349 - 362.

Tasbihsazan, R., Nettlebeck, T., & Kirby, N. (1997). Increasing mental development index in Australian children: A comparative study of two versions of the Bayley Mental Scale. *Australian Psychologist*, 32, 120-125.

Teasdale, T. W., & Owen, D. R. (1989). Continuing secular increases in intelligence and a stable prevalence of high intelligence levels. *Intelligence*, 13, 255–262.

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

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

Wai, J., Putallaz, M. (2011). The Flynn effect puzzle: A 30-year examination from the right tail of the ability distribution provides some missing pieces. *Intelligence*, In Press, 2011.

Wicherts, J. M., Dolan, C. V., Hessen, D., Oosterveld, P., Baal, G. C. M., van Boomsma, D. I., et al. (2004). Are intelligence tests measurement invariant over time? Investigating the nature of the Flynn Effect. *Intelligence*, 32, 509 - 537.

Wicherts, J. M. (2005, December). Flynn effect in the Woodcock – Johnson Cognitive Ability and Achievement Tests 1976 – 1999. Paper presentation at the annual meeting of the International Society of Intelligence Research, Albuquerque, NM.