A Critique of the Flynn Effect: Massive IQ Gains, Methodological Artifacts, or Both?

JOSEPH L. RODGERS
University of Oklahoma, Norman, OK, USA

The Flynn Effect proposed by Flynn (1984;1987) is reviewed and evaluated. Even in the presence of a skeptical and critical scrutiny of the effect, it appears that there is more than just methodological artifact to be explained. But the acceptance of the effect has been too quick. The proper explanations for the effect will not be meaningful until the nature of the effect is much better understood than it is now. Six questions are raised that have not been adequately answered. Two criticisms of the logic underlying the Flynn Effect are presented — one showing that even if IQ and SAT are highly correlated, their secular means will not necessarily track one another; the second showing that results by Flynn (1984) are as consistent with a changing IQ variance as with a changing mean. The second of these is empirically evaluated with a re-analysis of a subset of the sources of Flynn's original 1984 data. Finally, 10 research strategies and designs are suggested that would help us better understand the effect. The critique is developed with the goal of clarifying the nature, meaning and causes of the Flynn Effect. The author hopes that this critique will stimulate both healthy skepticism about the Flynn Effect and careful research into its actual causes.

The purpose of this article is to present a critique of the Flynn Effect, with particular attention to the original two articles that established the existence of the effect (Flynn, 1984, 1987). The critique will be organized around six questions, two criticisms and 10 research proposals. The six questions are ones that need to be answered to clarify the nature and meaning of the Flynn Effect. The two criticisms will point out methodological and logical weaknesses in the original source articles and in the response of the community of intelligence researchers to the proposed effect. The 10 proposals will suggest research strategies and designs that could be (and hopefully will be) used to clarify the nature, meaning and cause of the effect.

Direct all correspondence to: Joe L. Rodgers, Department of Psychology, University of Oklahoma, Norman, OK 73019, USA. E-mail: JRODGERS@OU.EDU

INTELLIGENCE 26(4): 337–356
ISSN: 0160-2896

Background

The Flynn Effect is one of the most surprising, most intriguing — and potentially most important — findings in the recent psychology research literature. Flynn (1984) used patterns in 73 studies to suggest the existence of "massive gains" in IQ in the US. He calibrated the level of gain at around 0.33 IQ points per year in the US from 1932 to 1978, an overall increase of around 15 IQ points over this period. Flynn (1987) followed with an analysis of IQ scores from 14 economically developed countries around the world, and found similar patterns that also supported IQ gains. These gains appear to reflect abstract problem solving ability more than other intellectual abilities that involve learning material; patterns from the Ravens Progressive Matrices, a relatively culture-free IQ test that involves a great deal of problem solving, provide the strongest support for the Flynn Effect across countries. In fact, Flynn (1987) suggested that there may have been *declines* in abilities related to learned content, and that these have been suppressing rather than contributing to IQ gains. Thus, he suggests that gains in specific problem solving abilities may be, if anything, even greater than those he documented for general IQ.

The Flynn Effect has been tested and replicated (Flynn, 1987; Teasdale & Owen, 1987, 1989; Lynn, 1990; Lynn & Hampson, 1986), and treated as input to related theoretical argument (Herrnstein & Murray, 1994; Stelzl, Merz, Ehlers, & Remer, 1995). The effect has been given attention in the popular press, which suggested that the cause "baffles intelligence experts" (Horgan, 1995). In April of 1996, a group of social scientist met at Emory University to "discuss possible explanations" for the Flynn Effect, a meeting that also drew the attention of the press (Azar, 1996). Articles by social scientists are beginning to emerge in journals and edited books (see, in particular, the articles by various scholars in Detterman, 1996 and the article by Neisser, 1997). Several causes of the Flynn Effect have been suggested. Lynn (1990) proposed that nutritional changes underlie the Flynn Effect, Brand (1996) pointed to the increased use of speeded tests, Mahlberg (1997) suggested a collective memory interpretation, and other suggested causes include educational innovation and television. Jensen (1996) proposed that the cause of the Flynn Effect may lie in "the summation of a great many possible causes, each of which alone has but a very small but real effect on mental development" (p. 149).

But before the effect is taken seriously by the community of social science researchers, its very existence should not be questionable. In other words, research addressing the *legitimacy* and *meaning* of the effect should precede research *testing* for and evaluating *causes* of the effect. Surely, Flynn's analysis was complete and painstaking, and his writing is clear and appropriately self-critical. In fact, he was very careful to self-evaluate the strength of his arguments; for example, in Flynn (1987) he classified how much support data from each country offered to the "Massive Gains" hypothesis, and also classified the legitimacy of his interpretations. However, I will argue that Flynn's arguments contain methodological weaknesses of which he was unaware, of which the community of researchers has not been sufficiently critical. Because his self-evaluation was also blind to these weaknesses, it tends to overstate the confidence we should have in the status of the Flynn Effect. Certainly, scrutiny by independent investigators of the logic leading to the claim is necessary before we will be able to understand what the Flynn Effect is, and ultimately to identify what cause or causes lie behind it.

Thus, the purpose of this critique is not to resolve the issue of the meaningfulness of the Flynn Effect or to specify the causes of the effect. Neither is the purpose to present extensive empirical analysis to provide further data or evidence concerning its legitimacy (although one suggestive empirical study will be presented). Rather, the purpose is to frame an approach to studying the Flynn Effect by defining a set of questions, criticisms and a research agenda. This critique opens discussion over what the nature of the Flynn Effect is, and of whether the Flynn Effect is real or a methodological artifact (or some combination). Other interpretations besides Flynn's and the ones presented here certainly exist as well, and should also be subjected to logical and empirical scrutiny.

Six Questions

Having read all of the literature I can find pertaining to the Flynn Effect, I am still not sure what the Flynn Effect really is. I will explore this uncertainty in the following questions, and answer them as best I can with reference to Flynn's articles and other intelligence literature.

1. Does the IQ gain to which the Flynn Effect refers operate within the individual?

In other words, does Flynn propose that an individual's IQ increases systematically over time? The answer to this question is almost certainly "No." In fact, the literature on human abilities discusses whether cognitive abilities really *decline* with age (Horn & Donaldson, 1976), and the large longitudinal study of Schaie (1994) documents slight increases in some mental abilities during ages in the 20s and 30s, with substantial declines following. There is no evidence — or suggestions from Flynn — that within-individual IQ increases are causing the effect Flynn observed in the studies he collected to establish the existence of the effect. In fact, most of the studies he used were based on cross-sectional data, and did not even contain the information necessary to evaluate within-individual change. One of the best data sources reviewed in Flynn (1987) came from The Netherlands (deLeeuw & Meester, 1984; cited in Flynn, 1987), which contained Raven scores allowing comparison of sons to fathers. No longitudinal studies containing within-individual information were even evaluated.

2. Is the gain to which the Flynn Effect refers supposed to operate within the family?

In other words, should later-born children have higher IQs than earlier-born children? Are the effects to which Flynn refers occurring within the family? The Dutch data mentioned above showed substantial gains by the sons when compared to their fathers' scores on the same test at the same age. This provides a powerful design, and eliminates some (but not all) of the confounds caused by using cross-sectional data. There is a large birth order literature studying IQ. Most empirical birth order studies have used cross-sectional data. Large national studies have showed *declining* IQ/achievement with increasing birth order (e.g., Breland, 1974 and Steelman & Mercy, 1980, using US data; Velandia, Grandon, & Page, 1978, using data from Columbia; Belmont & Marolla, 1983, using Dutch data), or in a few other cases, a flat relationship (e.g., Zajonc, 1976, reviewing previous research on French and Scottish data). Several theories have been proposed to explain the relationship of IQ scores to birth order and family size (e.g., Zajonc & Markus, 1975; Page & Grandon, 1979; Blake, 1981). Further, studies that have used within-family data have typically found approximately flat patterns, suggesting little or no within-family correlation between IQ and birth order (see Outhit, 1933; Berbaum & Moreland, 1980; Galbraith, 1982; Rodgers, 1984; Retherford & Sewell, 1991). An excellent review of this literature was presented by Ernst and Angst (1983), who were skeptical of the existence of any systematic birth order effects. These patterns suggest the opposite of the Flynn Effect, with IQ score decreasing or staying flat across birth order, which necessarily increases with time. If the Flynn Effect operated within families (but not within individuals), then there should be systematic increases in IQ across birth order in within-family data, and these would be translated into positive IQ–birth order relationships in both cross-sectional and longitudinal data. The absence of these effects can be used to infer that the Flynn Effect is not operating within the family. Further, some of the other sources of the Flynn Effect (discussed below) should be translated into within-family patterns, so that the absence of birth order patterns within families is suggestive in helping to rule out some interpretations.

3. Is the Flynn Effect an age, cohort, or period effect?

Psychologists are quite aware that changes across time — most effectively observed in longitudinal data — may be caused by either age effects, cohort effects, or period effects, and that only two of these types of effects operate independently (e.g., Adam, 1978; Costa & McCrae, 1982). I have already discussed and dismissed the possibility of an age effect in treating question #1 above, which leaves the period and cohort effects. Period effects would be ones caused by one or more social innovations that act equally at a point in time on all individuals, regardless of age. A particular educational innovation leading to improved educational quality experienced by all school-age individuals would be an example of a period effect. A cohort effect would be one acting on children or adults of a particular age, persisting across time. A particular educational innovation experienced only by high school students during a restricted period would be an example of a cohort effect; those students would potentially carry the value of the innovation with them throughout their lives, but students outside of the high school cohort would not. Theoretically, either a period or cohort effect or both could be what the Flynn Effect is resting on.

In Flynn's treatment, he seems to lean in the direction of interpreting the effect as a period effect. His calibration is applied to changes per unit of time, rather than to changes per cohort. But in cross-sectional data, period and cohort effects are perfectly confounded, and cannot be distinguished. Flynn does not, however, use this language, but rather refers consistently to "between-generation IQ differences." Because he does not make any claims about the cause of the effect, we could very properly excuse his treatment for not resolving this issue. However, a deeper consideration of the nature of the Flynn Effect suggests that a cohort vs. period interpretation needs to be resolved, since each has different implications for what the Flynn Effect means and what causes it. Research designs that can distinguish between these two causal sources are needed in this investigation.

4. Does the Flynn Effect operate equally within race categories?

Flynn (1987, p. 189) suggested that "IQ differences cannot, at present, be used as evidence" for group differences because "differences on IQ tests may not be equivalent to intelligence differences." This conclusion further derives from his suggestion that "Until the causal problem of what factors engender between-generation IQ differences is solved, no one knows what cultural variables are relevant" (p. 189). Thus, he reserves treatment of this question until some more preliminary questions can be answered. However, the *empirical* question of whether the IQ gains he observed would be the same within each race is a meaningful and important question. Statistical paradoxes exist (e.g., Simpsons Paradox; Simpson, 1951; Nunnally & Bernstein, 1994, p. 181) in which aggregate behavior is different than that of the individual groups contributing to the aggregate. Consistent Flynn Effects within each race would strengthen the legitimacy of the claim that the effect is real and meaningful. Inconsistent patterns would help point to the root causes. If the effect disappeared within races, the paradoxical finding would still require explanation, but more from a methodological perspective than a substantive one.

5. Does the Flynn Effect operate across the whole ability distribution?

In other words, is the effect operating equally on those of low, medium and high ability, so that the whole ability distribution is being improved at a constant rate? Or is the mean change observed repeatedly in all of the studies compiled by Flynn caused by increases in the IQ of a restricted part of the distribution? This is an extremely important methodological question that will be treated in more detail in the section on Criticisms. Teasdale and Owen (1989) found that the Flynn Effect gains in a Danish dataset were driven by gains in the bottom part of the intelligence distribution. If only a part of the distribution is driving the overall effect, the increases must be even more remarkable for that subgroup than if it applies to the overall distribution.

6. Does the Flynn Effect operate on all cognitive abilities, or only on certain abilities? Is the effect one that applies to latent intelligence itself, or to artifacts of the measurement of intelligence?

These two related questions have received a great deal of careful attention from Flynn (and others). Particularly, Flynn's (1987) own self-critique of his results is careful and thoughtful in addressing this issue. The major IQ improvements, if they are real, clearly derive from the domain of abstract problem solving ability on relatively culture-free tests. This finding all by itself rules out many causes that have strong cultural and educational content bases. Others have also raised and treated this problem. In particular, Jensen (1991) suggested separating superficial measurement processes from the underlying distribution of ability by using chronometric methods to anchor the test scores. Others (e.g., Loehlin, 1996) have also treated this problem. Flynn's early work questioned whether the substantial increases in IQ actually reflected gains in intelligence (Flynn, 1987). His recent article (Flynn, 1996) takes an even stronger position on this question than his earlier work, suggesting that "the portion of IQ gains over time which represents an intelligence gain is very small indeed; . . . paradoxically, the exciting thing to explain is the huge non-intelligence gain" (p. 27). Flynn (1996) repeatedly refers to "ersatz intelligence gains" as those caused by some artifactual process not related to actual intelligence. Our goal, he suggests, is to find measures of intelligence that will not register the artifactual gain but more validly measure actual intelligence levels and changes. One of the biggest puzzles growing out of the identification and research on the Flynn Effect is the elusive status of causal explanations for the phenomenon. One of the main points of this article is to emphasize the importance of understanding the nature of the Flynn Effect as a starting point for explaining the causes — artifactual or real — of the Flynn Effect.

In summary, Flynn's treatment strongly addresses question #6, can be used along with other literature and through implication to provide probable answers to #1, #2 and #3, and is relatively silent with respect to #4 and #5. Before meaning and causes can be addressed, however, better answers to these questions are prerequisite.

Two Criticisms

Flynn has presented the research community with a fascinating intellectual puzzle that has important theoretical and policy implications. Many others have shared his interests, either independently or in response to his work. Anyone working in this area must recognize that Flynn (1984) and Flynn (1987) are superb examples of synthetic research, in which a large corpus of existing empirical research was integrated into a coherent framework. Of course, given the complexity of the task he undertook, it is not surprising that some important points were treated quite well, while others were treated poorly or not at all. Two methodological mistakes will be discussed in this section.

1. Flynn (1984) was perplexed that IQ was increasing at the same time that Scholastic Aptitude Test (SAT) scores were declining. There is nothing perplexing at all about this finding. Flynn's "partitioning" method to justify his confusion was neither necessary nor correct.

This first criticism involves Flynn's concern over the relationship between rising IQ scores and falling SAT scores in the US during the 1960s and 1970s. His error does no damage to the Flynn Effect per se, but rather to the implications of the effect as discussed in Flynn (1984). He noted that "Between 1963 and 1981 . . . , American high school students who took the SAT showed a sharp decline in their average performance, particularly on the SAT-Verbal, the test most significant as a predictor of college grades" (p. 36). Following, he noted that past data showed correlations of around 0.5 to 0.8 between IQ tests and SAT scores, suggesting that "IQ and aptitude tests measure general intelligence to about the same degree and are functionally more or less equivalent" (p. 37). He concluded: "It seemed impossible that tests correlated at the 0.80 level and measuring much in common would permit the following: that over a period of 18 years, performance on the two kinds of tests had diverged by something between 0.288 SDU [Standard Deviation Units] (safe estimate) and 0.648 SDU (speculative estimate)" (p. 46). Throughout his article, Flynn refers to this combination of patterns as "unpalatable" (p. 38), "baffling" (p. 38) and "inexplicable" (p. 48). I will show that a pattern of increasing IQ, decreasing SAT and high correlation between them is, in fact, palatable, understandable and explicable.

It is a well-known result that means and correlations are statistically independent of one another. This fact suggests that knowledge of mean structure in the population gives us

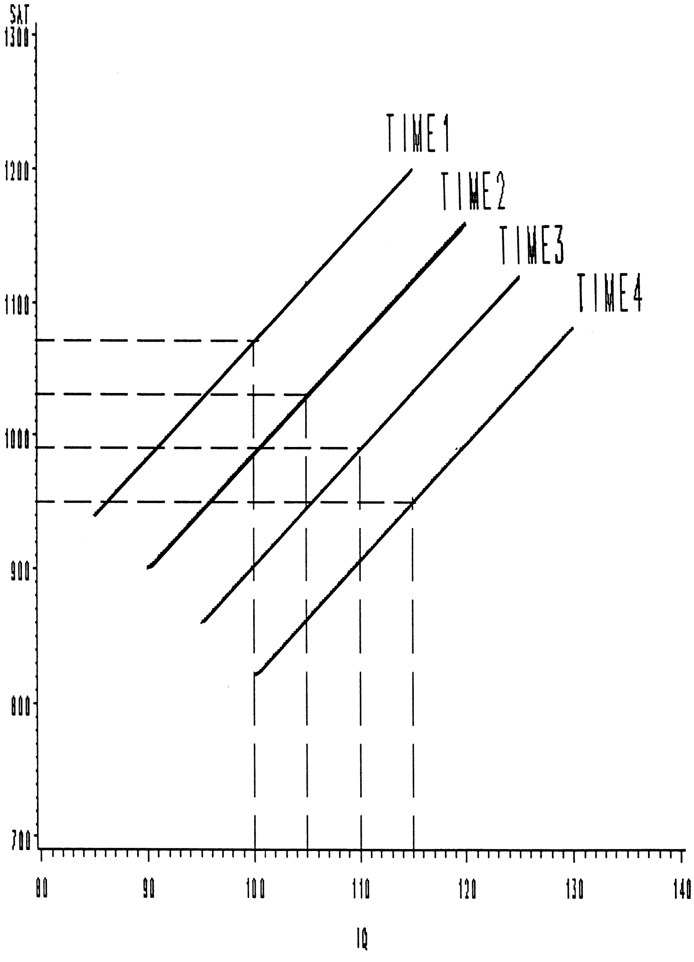


Figure 1. Presentation of increasing IQ, decreasing SAT, with $r_{\text{IQ,SAT}} = 1.0$.

no information about what correlation structure exists, or vice versa. (It is easy to observe in the correlation formula how means are "taken out" of each score in the computation of Pearson's *r*.)

Flynn found it surprising that two variables correlated around $r = 0.8$ could exist, with one of them showing a mean increase across time and the other a mean decrease. In fact, this pattern could occur even if the correlation were $r = 1.0$. Figure 1 shows a speculative finding in which mean SAT is decreasing over time (shown by projections on the *Y* axis), mean IQ is increasing over time (shown by projections on the *X* axis), and the within-year correlation between IQ and SAT is perfect.

How would we interpret such a finding substantively? A high correlation implies that IQ scores above the mean occur in individuals who also have SAT scores above the mean, and low IQ scores in individuals with low SAT scores, *within a given year*. Clearly, this link could occur year after year after year, while the population means *across years* are changing in any possible pattern. The IQ and SAT means and their relationships to $r_{\mathrm{IQ,SAT}}$ is not at all baffling.

Flynn spent considerable effort (pp. 36–39) trying to interpret the part of IQ *not* measured by SAT (i.e., the "pure" intellectual ability, uncontaminated by motivation and effort) as creating an even bigger quandary than was apparent: If the IQ gains are ones required to *overcome* the substantial parts of IQ that are measured by SAT that have been declining, then the gains in IQ means must be even greater in their domain than they appeared. Flynn (1984) developed a calculus of gains and losses in those means based on the incorrect belief that high correlations imply trends in means. But the calculus is neither necessary nor theoretically correct. Rather than conceptualizing the means as containing overlapping and non-overlapping parts, a technically more correct approach would be to use partial correlations or regression models to partition the explained variance. SAT scores can be declining and IQ scores can be increasing simultaneously. At the completion of his development, he noted "It is precisely at this point that one's head begins to spin" (p. 38) and "Going beyond simple models to speculate about ultimate causes makes no sense whatsoever of the trends in question" (p. 38).

He continued to use this "mean partitioning calculus" in Flynn (1987, p. 189), but concluded that "those entering American high schools were getting more and more intelligent, and yet they were leaving high school with worse and worse academic skills." Exactly! At this point he shifted to interpreting the patterns as substantively meaningful (rather than baffling), and simply reported what the observation implies. Both of those processes can be happening exactly as described, and yet the high positive correlation can be maintained. As stated above, this problem does no damage whatsoever to the existence of the Flynn Effect, but resolves the suggested paradox as only apparent. Explanations for simultaneous increasing IQ and decreasing SAT occur at the aggregate mean level across years. Explanations for the correlation between IQ and SAT occur at the individual level within a year. The two types of explanations may or may not overlap.

2. The methodology used in Flynn (1984) does not necessarily lead to the conclusion that IQ is increasing. Other interpretations are plausible and meaningful.

It is important that Flynn (1984) did not directly observe the gains he reported. Rather, he inferred their existence from a logical argument, applied to a certain type of data. The conclusion of IQ increase proposed in Flynn (1984) is presented as though this inference is not only plausible (which it is), but also automatic (which it is not). This second criticism could potentially undermine Flynn's whole argument. (Although, as it turns out, it apparently does not; Flynn's original article by itself leaves the proper conclusion unresolved, but his later work helps strengthen the legitimacy of his particular interpretation as probably correct.) He assumed that only rising IQ scores could have produced results like those in the datasets he analyzed. In fact, there are other population changes that could cause these effects as well.

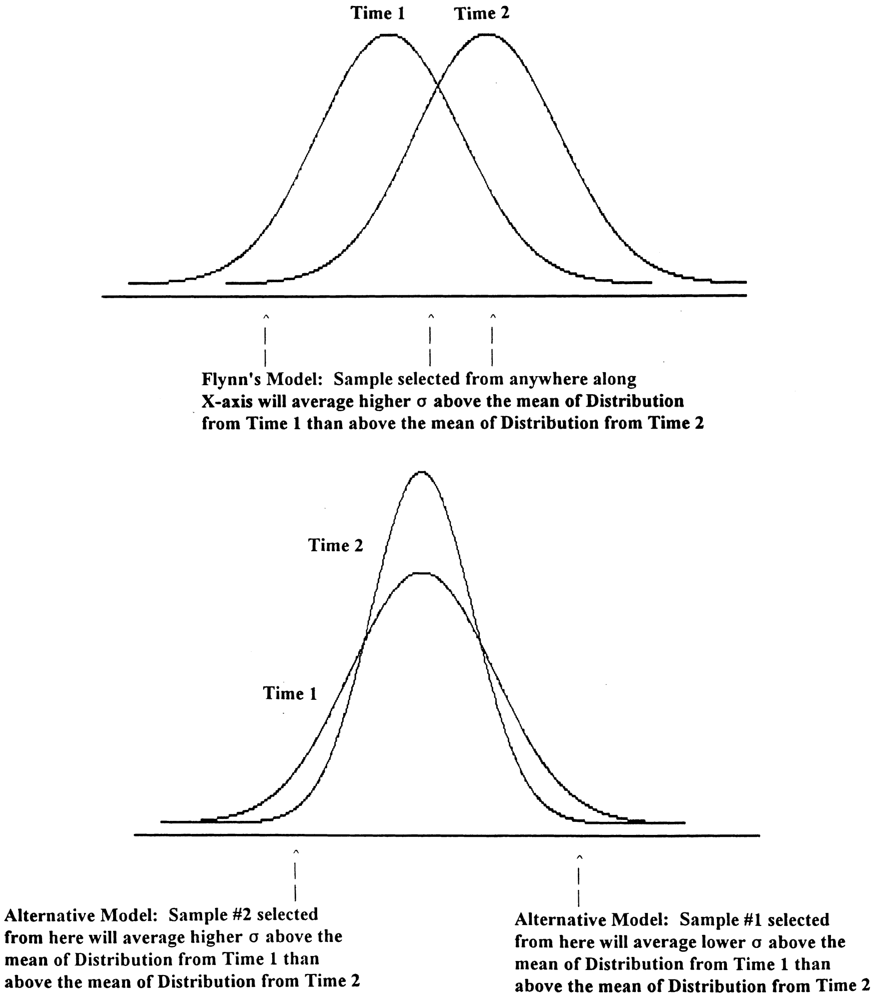
The original argument of Flynn (1984) was based on evaluating a certain type of data structure. Many studies have been run in which a single sample was administered two IQ tests, one from a recently normed instrument and one from an older instrument. If norming is done properly, then an individual's score can be interpreted in relationship to the population mean and standard deviation (usually set to be $\mu = 100$ and $\sigma = 15$ or 16 for IQ instruments). Because the population values may potentially change over time, test instruments must be re-normed periodically.

Flynn (1984, Table 2) found 73 studies based on a total of 7431 subjects that had compared scores from different versions (i.e., versions scaled using different norming samples) of the Stanford Binet and the Wechsler IQ tests. These comparisons accounted for 18 different pairs of IQ scales, with replications ranging from 1 for seven of the pairs, to as many as 17 for two of the pairs (for the 1932 SB and the 1947 WISC, and the 1947 WISC and the 1972 WISC-R). Generally, the mean across replications of the mean IQ scores for the two forms showed higher scores for the earlier IQ instrument. There was only one category inconsistent with this pattern, based on two studies; all other 17 pairs of IQ scales showed a higher mean for the earlier instrument. (Note that several other individual studies were inconsistent with the pattern, although averaging across studies in each category showed the consistent result.) For example, the 17 studies that compared the 1932 SB to the 1947 WISC showed a mean *gain* (indicated by a lower score on the later IQ test) of 5.49 IQ points. Collecting these 73 studies and identifying this remarkably consistent pattern is what led Flynn to conclude that IQ had been increasing at a fairly constant rate of around 1/3 IQ point per year for at least the 46 years between the first and last norming samples.

Between the identification of the patterns and the conclusion, however, exists a logical argument that is plausible and reasonable. Others exist as well, however. This argument presumes that subjects performing better on earlier IQ forms is prima facie evidence that an increase in IQ has occurred. Suppose a sample in 1947 took the WISC 1947 IQ test and scored an average of 100. They would be exactly average compared to the 1947 population. If they also scored 105 on the 1932 WB IQ test (consistent with Flynn's empirical patterns), however, they would be around 1/3 standard deviations above the population mean in 1932, suggesting that this average 1947 sample would be have been above average in 1932. Another way of stating the same logic suggests that if a test must become increasingly more difficult to reflect a normatively average score, then the latent ability that is being measured must be increasing as well.

Flynn — as well as both his critics and supporters in the Psychological community — have treated this argument as though it were unassailable. While it is plausible, it is certainly not the only possible interpretation. I re-emphasize that Flynn (1984) results did not demonstrate the existence of IQ gains. His data showed consistent differences within a single sample (i.e., on tests administered to a single group at one point in time) across pairs of tests in which one had been normed earlier in time than the other. The existence of IQ gains over time was his logical inference based on these data patterns.

Without loss of generality (and simply to provide a reasonable graphical tool), assume a normal distribution of intelligence in the population. In Figure 2 (top), I present the pattern that Flynn assumed to be underlying his findings. If (latent) intelligence increases, mean (observed) IQ increases as well. This pattern does indeed explain his results. If mean intelligence is increasing systematically across time, then the gain scores found in his samples would indeed obtain.



Flynn's Model: Sample selected from anywhere along X-axis will average higher σ above the mean of Distribution from Time 1 than above the mean of Distribution from Time 2

Alternative Model: Sample #2 selected from here will average higher σ above the mean of Distribution from Time 1 than above the mean of Distribution from Time 2

Alternative Model: Sample #1 selected from here will average lower σ above the mean of Distribution from Time 1 than above the mean of Distribution from Time 2

Figure 2. IQ distributions normed at two different times, with samples drawn to take both IQ forms.

A change in population standard deviation of intelligence across time — combined with either fixed or changing population mean — could also produce the results that Flynn obtained. Before changing σ's could masquerade as Flynn's changing μ's, however, the sample taking the two tests must be a select sample (i.e., its sample mean must be different from the population mean). It is important to note that only slight selection — even at a level caused by accidents of random sampling — could lead to misinterpreting changing variability for changing means.

Suppose that population mean intelligence remains fixed, but the population standard deviation systematically decreases. Now draw a sample of subjects (Sample #1) that is positively select (i.e., their average intelligence is above the fixed population mean). This situation is shown in Figure 2 (bottom). Obviously, the sample will be expected to score further out in the tail of the newer intelligence distribution; the newer IQ test — normed on this more recent and less variable population — will respond by assigning a higher score to the later test than to the earlier test. The reverse effect will occur if the sample is negatively selected, if a sample (Sample #2) has a lower mean intelligence than that of the population. Then IQ scores from later tests will be systematically lower than those from earlier tests. This is exactly the pattern that Flynn observed, which is as consistent with an interpretation of negatively selected samples and reduced variance as with generally increasing means. Alternatively, if the population standard deviation in intelligence systematically increases over time, the opposite patterns will be observed. Samples that are positively selected will score lower on the newer IQ test, and negatively selected samples will score higher on the newer IQ test.

Thus, Flynn's findings are consistent with increased variability in the upper half of the intelligence distribution, and/or decreased variability in the lower half of the distribution. While the simple models implied above suggest changing variability at the overall distributional level, different processes can in theory occur in the different halves of the intelligence distribution. For example, increased variability in the upper half of the intelligence distribution with no change in the lower half would result in a Flynn Effect, detectable in positively selected samples taking two different IQ tests. Decreased variability in the lower half of the intelligence distribution would result in a Flynn Effect, detectable in negatively selected samples. Finally, if variability in intelligence increased over time in the upper half and simultaneously decreased over time in the lower half (possibly caused by two entirely different processes), this would also produce a Flynn Effect that would be detectable in either positively or negatively selected samples.

These types of changes in variability — decreasing variance in the lower half of the intelligence, or increasing variance in the upper half — will cause the mean to increase automatically. But the dynamics underlying these changes would be interpreted differently under these different causal models. For example, a nutritional improvement that affected all children's intelligence positively and equally would be reflected in an overall mean IQ change (reflected in the need to renorm the IQ tests to adjust for this mean change), without changing the shape of the IQ distribution. A nutritional program for low SES children that reduced variability in the lower half of the distribution by improving intelligence for underprivileged children would also show an overall mean IQ change, but the intelligence in the upper half of the distribution would not change at all. The process driving the change could be interpreted as either a variance-reduction or mean-increasing process in the lower half of the distribution (and the overall shape of the distribution would change substantially). While these two nutritional effects could each produce the patterns that Flynn observed, distinguishing them is an important step in understanding the factors causing the Flynn Effect. These arguments suggest that researchers should attend to the overall structure of the IQ distribution, rather than just to means or standard deviations.

Previous empirical evidence exists supporting the contraction of variability in the bottom half of the distribution (e.g., Herrnstein & Murray, 1994, p. 308). Teasdale and Owen (1989) studied a representative sample of Danish draftees and concluded that "gains appeared to be concentrated among lower intelligence levels, and we find no evidence of gains at the higher levels" (p. 255). Their conclusion was to support educational system changes as driving the change. Lynn and Hampson (1986) noted that "in our studies of the rise of mean IQ in Britain over the period 1932–1982, we have found that the rise in the lower half of the intelligence distribution has been about double that in the upper half." On the other hand, Flynn (1996) suggested that "IQ gains extend to every IQ level" because "score variance remains unchanged over time" (p. 25).

This issue is obviously of considerable concern, yet Flynn's (1984) original article provided no discussion of the select nature of his samples. He did spend considerable effort accounting for the representativeness of the norming samples used in the different IQ forms (which is indeed prerequisite for the types of conclusions he wanted to draw, and in fact for the validity of the test score procedures in the first place). But virtually no reference is made to the nature of the samples. Two different empirical results will be presented here that begin to address this critical question. First, the means presented in Table 2 of Flynn (1984) will be evaluated in relationship to the arguments presented above (i.e., to account for positively and negatively selected samples in relation to possible changes in the variabilities). Second, the variability of several of the studies summarized in Flynn (1984) will be presented to help in this evaluation.

Flynn (1984, p. 33, Table 2) presented 18 sets of means for "Test 1" (the earlier IQ test) and "Test 2" (the later IQ test). These means were weighted means derived from multiple studies that he identified that evaluated comparisons between combinations of IQ test forms. To evaluate the select nature of each sample, it would be ideal to have an IQ score from a test normed to the population in the same year that the test was taken. Although this did not occur, in each case, "Test 2" scores were closer to defining the "contemporary IQ" than the "Test 1" scores. In the column labeled "Test 2," six of the sample combinations had IQ means above 105, five had IQ means below 95, and the remaining seven had IQ means between 95 and 105 (three above 100, and four below 100). The correlation between these IQ means and the gain from Test 1 to Test 2 was $r = -0.14$, a non-significant correlation. The conclusion from this evaluation is that there is no particular pattern to the selected nature of the samples. If the overall variability were the source of the Flynn Effect, it should be reflected in a relationship between the gain scores and the select nature of the sample, as described above, and such a pattern was not observed. However, it should be noted that these means were weighted means based on from 1 to 17 studies, and the selection issue described above applies to the individual study and not to combinations of studies. It is certain that many of the categories included both positively and negatively selected samples. Only by re-evaluating the original studies from Flynn (1984) can this issue be address directly.

A complete evaluation of all of the individual studies in Flynn's (1984) Table 2 should be undertaken in future research. A highly suggestive smaller evaluation of the original studies will be presented here. I obtained the original studies from six of the 18 IQ form combinations. These six were chosen on the following grounds. The first choice was the two studies (Knopf, Murfett, & Milstein, 1954; Price & Thorne, 1955) from the single category that contributed to the only anomalous finding in Flynn's (1984) Table 2, the only finding inconsistent with an IQ increase (a comparison of the Wechsler–Bellevue Form I — normed in 1936, with the WISC — normed in 1947). The additional five categories were chosen as all of those that Flynn rated as indicating an IQ rate gain of 0.44 per year or higher (i.e., they were the five highest IQ gain categories, including rates of 1.13 per year, 0.63 per year, 0.57 per year, 0.50 per year and 0.44 per year.) The logic in this choice was that the basic structure (as well as the causes) of the Flynn Effect should be best revealed in settings in which the effect is magnified. The nine original studies within each of these categories were Triggs and Cartee (1953), Quereshi (1968), Querishi and Miller (1970), Hannon and Kicklighter (1970), Simpson (1970), Brooks (1977), Rasbury, McCoy, & Perry (1977), Sewell (1977) and Wechsler (1974). The last of these (Wechsler, 1974) was contained in a testing manual that was not available. The means from Wechsler (1974) were estimated, however, by using the other study in this category to infer the means (although the standard deviations were not available). The reason for obtaining these original studies was to observe the mean and standard deviation patterns in each; only category means averaged across studies, and no standard deviations, were reported in Flynn (1984).

In the two studies showing a higher mean on the later test, Knopf et al.'s sample had WISC IQ means ranging from 103 for Verbal IQ to 105 for Performance IQ (with full scale IQ of 104.030), and Wechsler–Bellevue scores ranging from 99 on Verbal IQ to 103 on Performance IQ (with full scale IQ of 100.63); Price and Thorne's sample had mean WISC IQs ranging from 103 to 109 (with a composite Full-Scale IQ of 106.1) and Wechsler–Bellevue IQs ranging from 95 to 110 (with a composite Full-Scale IQ of 104.5). In both cases, subjects were slightly positively selected compared to the most recent norming population. The combination of positive selection and higher IQ scores on the later form is consistent with either an IQ loss (based on Flynn's mean argument) or on an upper-tail reduced variance interpretation (based on the argument from the right-hand side of Figure 2b). Thus, for example, if the other studies reviewed by Flynn were based on negatively selected samples (or, in fact, if the "mean study" in each category of Table 2 were negatively selected), this finding would be as consistent with a generally decreasing variance in the IQ distribution as with an increase in the mean, and would furthermore resolve the confusion resulting from the two anomalous studies. However, inspection of the other studies showed that this was not the case.

The means and standard deviations from the eleven studies are presented in Table 1. In the nine studies from categories Flynn classified as indicating the most extreme rates of IQ gain, four were positively selected (IQ2 mean greater than 105), three were negatively selected (IQ2 mean less than 95) and two were unselected (IQ2 mean between 95 and 105). By category, the Stanford–Binet (Form M) vs. WISC comparison was positively selected; the WISC vs. WAIS on average had no selection; the WISC vs. Stanford–Binet comparison was negatively selected; the WPSSI vs. Stanford–Binet comparison was negatively selected; and the WPSSI vs. WISC comparison was positively selected. Further, inspection of the standard deviations showed that in half of the positively selected samples, the sample standard deviation was lower for the earlier test than for the later test, while in the other half it was higher for the earlier than for the later test. Among the three negatively selected samples, the later standard deviation was lower than the earlier two times, and the earlier was lower than the later one time. Thus, there appears to be no pattern relating the select nature of the samples and the overall standard deviations. These results suggest strongly that the cause of the Flynn Effect is *not* an overall change in the variance (like that shown in Figure 2).

Table 1. Means and Standard Deviations from Studies Used in Flynn (1984) that Came from Categories at the Positive and Negative Extremes of the IQ-Change Continuum

Test 1	*Test 2*	*Study*[a]	$\overline{X}_1$[b]	$\overline{X}_2$[b]	S_1[b]	S_2[b]
SB-M, 1932	WISC, 1947.5	Triggs (1953)	124.1	107.6	9.7	13.2
WBI, 1936.5	WISC, 1947.5	Knopf (1954)	100.6	104.0	10.1	10.0
		Price (1955)	104.5	106.1	13.8	12.3
WISC, 1947.5	WAIS, 1953.5	Hannon (1970)	104.1	103.2	24.1	18.7
		Quereshi (1968)	110.9	107.0	10.9	8.3
		Simpson (1970)	82.5	88.8	11.2	8.6
		Quereshi (1970)	114.2	111.0	13.3	9.0
WISC, 1947.5	SB72, 1971.5	Brooks (1977)	96.4	87.2	19.1	16.0
WPPSI, 1964.5	SB72, 1971.5	Sewell (1977)	95.8	91.5	12.3	13.8
WPSSI, 1964.5	WISC-R, 1972	Rasbury (1977)	119.3	114.5	9.3	10.8
		Wechsler (1974)	99.8[c]	96.8[c]	–[c]	–[c]

Notes: [a] Studies are identified by only the name of the first author, to economize space in the table; see References for full citations.

[b] The subscripts on the means and standard deviations refer to the "age" of the IQ test in the comparison. $\overline{X}_1$ and S_1 refer to the mean and standard deviation of the earlier test, $\overline{X}_2$ and S_2 refer to the later test. Following the logic assumed in Flynn (1984), a mean decline in sample means from the earlier to the later test implies increasing population mean IQ over time. The logic developed in this paper suggests that changes in the sample means can also result from changes in the population standard deviations as well.

[c] Means were inferred and standard deviations were not available.

In his second article, Flynn (1987) used a different methodology than that in Flynn (1984). This second approach involved direct comparisons of IQ patterns across different ages of cross-sectional respondents. For example, among the "strongest" data used by Flynn (1987) came from Belgium, where all 18-year old men took a battery of tests upon military induction. Between 1958 and 1967, cross-sectional means increased between 2.65 and 4.50 IQ units on the several tests. Because these data are based on (essentially) a whole population, the selection problem from the 73 studies in Flynn (1984) did not exist.

Because many of these studies rested on direct observation of means, the problem of potentially misplaced inference that occurred in Flynn (1984) is not of any concern. Further, the finding of Flynn (1987) of direct mean changes over time in cross-sectional samples suggests that Flynn's (1984) interpretation of his results as deriving from mean changes is likely to be correct (assuming the causal processes underlying these patterns are the same in the US as in the other countries studied). However, I emphasize that mean changes and variance changes can be occurring simultaneously, and the distributional nature of these changes lies exactly at the heart of our understanding of what the Flynn Effect is and what it means.

The re-analysis of Flynn's (1984) data and Flynn's (1987) research account for changes in the overall distributions. However, whether the mean changes are systematic ones across the whole distribution, or whether they are caused by decreasing variance in the lower half of the distribution and/or increasing variance in the upper half is not resolved by this investigation. Since some empirical evidence (reviewed above) suggests contraction in the lower half of the distribution, the issue of whether the Flynn Effect is a fundamental shift in the mean of the overall distribution, or in the variance from a subset of the distribution is still open to future research. In fact, other moments of the IQ distribution besides the mean and variance — those related to skewness and kurtosis — are of theoretical interest, and are substantively interpretable. We will only begin to understand the Flynn Effect when we interpret it at the distribution level, with particular focus on the raw scores themselves. Only at this level can the issues of select samples, differential contraction or expansion of parts of the distribution, and the substantive interpretations attached to these patterns begin to clear up. Micceri (1989) studied many empirical distributions found in nature, and few of those followed the traditional normal form. He found, for example, many more "lumpy" distributions than standard statistical procedures (or researchers who account for them) typically expect. Following Micceri, we need to understand in much more detail the nature of the whole distributions of IQ scores to unravel the puzzles underlying the Flynn Effect.

Ten Proposals for Future Research

Excellent research often results in more questions than answers. Flynn's work has identified an important and fascinating empirical pattern that begs for additional research in response to the questions it has raised. In this final section, I will present 10 research projects that would inform our understanding of the nature, meaning and causes of the Flynn Effect. Some of these lines of research will be defined by specifying the research questions that should be addressed (and data, designs and analyses will be suggested that might help answer the research question). In other cases, direct research designs or data collection efforts will be suggested. These research efforts will be organized around efforts to understand the *nature,* the *meaning* and the *causes* of the Flynn Effect.

The actual *nature* of the empirical pattern underlying the Flynn Effect is still far from being well defined. Empirical analysis directed toward answering several additional questions about its nature would be helpful.

1. Does the Flynn Effect show up in major demographic subgroups, or only in aggregate patterns?

Flynn has documented its cross-cultural occurrence, but there are many gender, race and regional differences in intellectual functions. A great deal of research has been devoted to understanding such subgroup differences, and some of the knowledge gained from that research could be applied to understanding the Flynn Effect if we knew how it behaves within those subgroups.

2. Is the Flynn Effect a period or cohort phenomenon, and how does it apply to within-individual patterns of intellectual development (i.e., are there any age effects)?

A whole series of articles from the 1960s to the early 1980s developed longitudinal, cross-sequential and cross-sectional designs that can be used to resolve this important question (e.g., Schaie, 1965, 1976; Baltes, 1968; Adam, 1978; Costa & McCrae, 1982). If the Flynn Effect has some cohort component, then we need to know to what cohorts the effect applies, and why. If the Flynn Effect is purely a period effect, we need to know what is happening in the US and other economically developed countries to create such systematic IQ increases (with a particular interest in continuing to promote its elevation and extending it to other domains besides abstract problem solving skills). If there is any aging implication to the Flynn Effect, we need to know at what part of the lifespan it applies, whether there are other compensatory processes that additively or interactively compete, and what are the specific mechanisms of its contribution to aging.

3. If the Flynn Effect does have a strong period interpretation, how far back does it go? Is it continuing? Why has the pace been so consistent during the period that Flynn has studied?

Presumably, IQ — or abstract problem solving skills, as Flynn (1987) suggested underlie the "massive IQ gains" — has not been systematically increasing forever, and cannot keep doing so in perpetuity. Is there evidence that can be found to define the beginning of the Flynn Effect, to project the end, or to suggest changes in its pace? Different paradigms may be necessary to answer these questions than those used in Flynn (1987) and, especially, Flynn (1984). But the statement of the effect can be used to make specific predictions about IQs (and other outcomes) that have been measured outside the intervals that Flynn has studied. The constancy of its pace is one of the intriguing (and most surprising) pieces of the "Flynn puzzle." Further efforts to understand this constant pace would help our understanding of the effect.

4. Does the Flynn Effect occur only in economically developed countries, or also in LDCs?

It will be much harder to obtain data to identify a Flynn Effect in Lesser-Developed Countries, but there certainly exists IQ information in some LDCs.

Besides many additional questions about the nature of the Flynn Effect (some of which can be answered with the help of available — or still-to-be-collected — empirical data and standard designs), there are also a number of questions about the *meaning* of the Flynn Effect.

5. What does it mean to say "massive gains in IQ?"

This question harkens back to a number of issues raised earlier in this critique, and these issues define the requisite "next steps" before any additional meaningful progress can be made in understanding the Flynn Effect. Question #5 implies several sub-questions: Do the "massive gains" refer to the overall distribution of IQ, to sub-parts of the distribution, to family units themselves, and/or to each individual within that distribution? A similar and critical question is whether the gains only show up in IQ distributions, or whether they can also be detected in raw score distributions themselves.

Another way to address the "meaning" question is through a series of thought-experiments: If a US individual from 1950 ages normally to 1970, would we expect that exact individual to arrive with an IQ around seven points higher? If all individuals below (say) the median IQ in 1950 age normally to 1970, would they arrive with a mean around seven points higher? What about all 10-year-olds? All high school students? All US Hispanics? All US males? Those above the median IQ? What if different years were used than 1950 and 1970? What if the aging occurred over a 20-year period, but for all individuals from age 20 to 40? What if the aging occurred over a 20-year period to all high school students born between 1930 and 1950? What if all of these questions were posed at the level of the raw score, rather than in relation to IQ scores normed against the population? Do the increases in IQ actually represent a higher probability of answering IQ questions correctly? If so, to what kinds of questions does the increase probably apply? The repeated reference to "massive gains" falls short of answering these types of specific questions, and until they are addressed with appropriate designs, we will not really understand what the term "Flynn Effect" means.

Early in this critique, I suggested that we cannot properly search for causes until we better understand the nature and meaning of the Flynn Effect. The search has already begun, however, even if prematurely. In fact, extensive research to help us understand what social and biological factors influence intellectual performance has been conducted for many years, and work on the Flynn Effect falls into this much larger domain. For example, many studies of the effect of nutrition on IQ have already been performed, and these may have a great deal to say about the role of nutrition in understanding the Flynn Effect. Research on the effect of schooling, or Head Start, or nutrition, or dropping out, or having a certain number of siblings can all be brought to bear on understanding the Flynn Effect. In return, research on the Flynn Effect can ultimately help us understand the general underlying causes of individual differences in intelligence and the various measures of it.

It is not premature to speculate about what kind of data and designs will be necessary to help identify *causes* of the Flynn Effect.

6. Research on existing data like that in Flynn (1984) could be done in which common items are used to link across time (or, alternatively, in which whole batteries are compared across time).

Flynn (1984) used studies based on a common sample and different IQ forms as his basic design. Flynn (1987) used studies based on common IQ forms (for the most part). Even different IQ forms contain common questions, which could be used for equating purposes. This might, for example, allow research on IQ scores from earlier in the 20th century than Flynn (1984) was able to define. Extensive recent work published in the psychometric literature on test equating and linking procedures could be applied fruitfully to studying the Flynn Effect.

Arthur Jensen (personal communication, 1998) suggested a broader version of this idea. If a whole battery of tests (or subtests from a common IQ form) have been administered to two samples at disparate points in time, the raw score difference on these batteries could be investigated to elucidate the patterns behind and the causes of the Flynn Effect. Further, factor structures of the batteries could be compared across time, and subtest patterns could be investigated across time to identify specific domains that contribute (or that do not) to the Flynn Effect. Identifying data that would permit either item-level comparisons across time, or comparisons of intellectual subscales across time, would provide the basis for valuable and fascinating research projects.

7. Research explicitly using the information contained in select samples could help us understand the nature, meaning and causes of the Flynn Effect.

As discussed in the second criticism, Flynn (1984) erred in failing to account for the selection processes involved in the studies he reviewed. Going back through these studies and accounting for the select nature of those samples would be very helpful research effort. Alternatively, identifying highly select samples (from either the upper or lower tail or both) and studying their IQ patterns would help us identify where our causal models should be applied. It seems axiomatic that many of the processes driving individuals into the upper tail of the intelligence distribution are very different from those driving them into the lower tail.

8. A re-analysis of the Flynn (1984) data as a formal meta-analysis would be a helpful contribution.

Flynn's (1984) Table 2 is the basis for a formal meta-analysis. When he conducted that research, however, the methodology of meta-analysis was early in its development. A less biased estimate of the effect size estimated by Flynn would be one product of such an effort. An even more useful contribution would be to code additional information from the studies reviewed in Flynn (1984), and correlate that information with the size of the effects. Through such an effort, the role of the period, the role of selection, the role of differential effects in different parts, and other problems raised above concerning the IQ distributions could all be formally evaluated.

9. Presumably IQ forms change systematically over time. Research into the specific nature of the questions that causes differences between two IQ forms would help us understand the causes of the Flynn Effect.

One reason samples who take two different forms of an IQ test (as in Flynn, 1984) do better on the earlier form may be that some of the questions from the earlier form become more a part of the general cultural knowledge than they were for the norming sample for the older test. A content analysis of these question changes with the Flynn Effect specifically in mind would be a useful study. The idea of a cultural collective memory developed by Mahlberg (1997) provides a useful theoretical basis for this type of analysis.

10. At some point, when the nature and meaning of the Flynn Effect is better specified and understood, some of the social innovations that might have influenced intelligence should be evaluated using appropriate quasi-experimental design procedures.

Many evaluation studies have been performed of educational and social innovations, and the methods for those are well-developed and sophisticated (e.g., Cook & Campbell, 1979). Creative designs and analyses to identify potential causes of the Flynn Effect can be derived from such methods.

CONCLUSION

I have provided motivation and suggested questions and methods to address the nature, meaning and causes of the Flynn Effect. I have assumed throughout that careful scrutiny of the logic and constructive criticism of the methods on which it is built is prerequisite to acceptance of the Flynn Effect as a well-defined phenomenon in search of an explanation. Even with a healthy dose of skepticism, the effect rises above purely methodological interpretation, and appears to have substantive import. But its nature, meaning and causes are still far from being well understood.

Acknowledgements: The author would like to acknowledge discussions with Amalia Bjornsdottir and Robert Terry that improved this article. Presentations to and comments from the Bi-Weekly Quantsem group from the Oklahoma University Psychology Department also contributed to the ideas developed in this article. Finally, the reviewers — Arthur Jensen, Lee A. Thompson and Keith Widaman — and the editor — Douglas Detterman — made many suggestions that substantially improved both the ideas and the presentation.

References

Adam, J. (1978). Sequential strategies and the separation of age, cohort, and time-of-measurement contribution to developmental data. *Psychological Bulletin, 85*, 1309–1316.

Azar, B. (June, 1996). People are becoming smarter — why? *American Psychological Association Monitor, 20.*

Baltes, P. B. (1968). Longitudinal and cross-sectional sequences in the study of age and generation effects. *Psychological Bulletin, 11*, 145–171.

Belmont, L., & Marolla, F. A. (1983). Birth order, family size, and intelligence. *Science, 182*, 1096–1101.

Berbaum, M. L., & Moreland, R. L. (1980). Intellectual development within the family: A new application of the confluence model. *Developmental Psychology, 16*, 506–515.

Blake, J. (1981). Family size and the quality of children. *Demography, 18*, 421–442.

Brand, C. (1996). G, genes, and pedagogy: A reply to seven (lamentable) chapters. In D. Detterman (Ed.), *Current topics in human intelligence, vol. 5: The environment*. Norwood, NJ: Ablex.

Breland, H. M. (1974). Birth order, family configuration, and verbal achievement. *Child Development, 45*, 1101–1109.

Brooks, C. R. (1977). WISC, WISC-R, S-B L and M, WRAT: Relationships and trends among children ages six to ten referred for psychological evaluation. *Psychology in the Schools, 14*, 30–33.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis for field settings*. Boston: Houghton Mifflin.

Costa, P. T., & McCrea, R. R. (1982). An approach to the attribution of aging, period, and cohort effects. *Psychological Bulletin, 92*, 238–250.

Detterman, D. K. (1996). *Current topics in intelligence, vol. 5: The environment*. Norwood, NJ: Ablex.

Ernst, C., & Angst, J. (1983). *Birth order: Its influence on personality.* New York: Springer-Verlag.

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

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

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

Galbraith, R. C. (1982). Sibling spacing and intellectual development: A closer look at the confluence models. *Developmental Psychology, 18*, 151–173.

Hannon, J. E., & Kicklighter, R. (1970). WAIS vs. WISC in adolescents. *Journal of Consulting and Clinical Psychology, 35*, 179–182.

Herrnstein, R. J, & Murray, C. (1994). *The bell curve*. New York: The Free Press.

Horgan, J. (November, 1995). Get smart, take a test: A long-term rise in IQ scores baffles intelligence experts. *Scientific American*, 12–13.

Horn, J. L., & Donaldson, G. (1976). On the myth of intellectual decline in adulthood. *American Psychologist, 31*, 701–719.

Jensen, A. R. (1991). Speed of elementary cognitive processes: A chronometric anchor for psychometric tests of g. *Psychological Test Bulletin, 4*, 59–70.

Jensen, A. R. (1996). Secular trends in IQ: Additional hypotheses. In D. Detterman (Ed.), *Current topics in human intelligence, vol. 5: The environment*. Norwood, NJ: Ablex.

Knopf, I. J., Murfett, B. J., & Milstein, V. (1954). Relationships between the Wechsler–Bellevue Form I and the WISC. *Journal of Clinical Psychology, 10*, 261–263.

Loehlin, J. C. (1996). Environment and intelligence: A comment. In D. Detterman (Ed.), *Current topics in human intelligence, vol. 5: The environment*. Norwood, NJ: Ablex.

Lynn, R. (1990). Differential rates of secular increase of five major primary abilities. *Social Biology, 37*, 137–141.

Lynn, R., & Hampson, S. (1986). Intellectual abilities of Japanese children: An assessment of 2 1/2–8 1/2-year-olds derived from the McCarthy Scales of Children's Abilities. *Intelligence, 10*, 41–58.

Mahlberg, A. (1997). The rise in IQ scores. *American Psychologist, 52*, 71.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*, 156–167.

Neisser, U. (1997). Rising scores on intelligence tests. *American Scientist, 85*, 440–447.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory.* New York: McGraw-Hill.

Outhit, M. C. (1933). A study of the resemblance of parents and children in general intelligence. *Archives of Psychology, 149*, 1–60.

Page, E. B., & Grandon, G. (1979). Family configuration and mental ability: Two theories contrasted with U.S. data. *American Educational Research Journal, 16*, 257–272.

Price, J. R., & Thorne, G. D. (1955). A statistical comparison of the WISC and Wechsler–Bellevue, Form I. *Journal of Consulting Psychology, 19*, 479–482.

Quereshi, M. Y. (1968). The comparability of WAIS and WISC subtest scores and IQ estimates. *Journal of Psychology, 68*, 73–82.

Quereshi, M. Y., & Miller, J. M. (1970). The comparability of the WAIS, WISC, and WB II. *Journal of Educational Measurement, 7*, 105–111.

Rasbury, W., McCoy, J. G., & Perry, N. W. (1977). Relations of scores on WPPSI and WISC-R at a one-year interval. *Perceptual and Motor Skills, 44*, 695–698.

Retherford, R. D., & Sewell, W. H. (1991). Birth order and intelligence: Further tests of the confluence model. *American Sociological Review, 56*, 141–158.

Rodgers, J. L. (1984). Confluence effects: Not here, not now! *Developmental Psychology, 20*, 321–331.

Schaie, K. W. (1965). A general model for the study of developmental problems. *Psychological Bulletin, 64*, 92–107.

Schaie, K. W. (1976). Quasi-experimental research designs in the psychology of aging. In J. Birren, & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (pp. 39–58). New York: Reinhold-VanNostrand.

Schaie, K. W. (1994). The course of adult intellectual development. *American Psychologist, 49*, 304–313.

Sewell, T .E. (1977). A comparison of the WPPSI and Stanford–Binet intelligence scale 1972 among lower SES black children. *Psychology in the Schools, 14*, 158–161.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B, 13*, 238–241.

Simpson, R. L. (1970). Study of the comparability of WISC and the WAIS. *Journal of Consulting and Clinical Psychology, 34*, 156–158.

Steelman, L. C., & Mercy, J. A. (1980). Unconfounding the confluence model: A test of sibship size and birth-ofer effects on intelligence. *American Sociological Review, 45*, 571–582.

Stelzl, I., Merz, F., Ehlers, T., & Remer, H. (1995). The effect of schooling on the development of fluid and crystallized intelligence: A quasi-experimental study. *Intelligence, 21*, 279–296.

Teasdale, T. W., & Owen, D. R. (1987). National secular trends in intelligence: A twenty-year cross-sectional study. *Nature, 325*, 119–121.

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

Triggs, F. O., & Cartee, J. K. (1953). Pre-school performance on the Stanford–Binet and the Wechsler intelligence scale for children. *Journal of Clinical Psychology, 9*, 27–29.

Velandia, W., Grandon, G. M., & Page, E. B. (1978). Family size, birth order, and intelligence in a large South American sample. *American Educational Research Journal, 15*, 399–416.

Wechsler, D. (1974). WISC-R manual. New York: The Psychological Corporation.

Zajonc, R. B. (1976). Family configuration and intelligence. *Science, 192*, 227–236.

Zajonc, R. B., & Markus, G. B. (1975). Birth order and intellectual development. *Psychological Review, 82*, 74–88.