1 A bombshell in a letter box

> The special function of scientific explanation is . . . to turn the unexpected, as far as possible, into the expected.
>
> (Stephen Toulmin, *Reason in ethics*, p. 88)

I am a teacher and rarely write for specialists alone. I have tried to avoid the dead-stick prose so beloved by journal editors. Anyone with a good education or a major in psychology should be able to read this book and the former is more important than the latter. It assumes that everyone is interested in intelligence and would like something exciting to provide a reason to learn more about it. Specialists will find that much has been omitted but will also, I hope, find something new in the argument and something worth pursuing in the research designs recommended.

A warning for everyone: there are problems that can simply be settled by evidence, for example, whether some swans are black. But there are deeper problems that pose paradoxes. Sometimes the evidence that would solve them lies in an inaccessible past. That means we have to retreat from the scientific level of explanation to the historical level where we demand only a plausibility that conforms to the known facts. I believe that my efforts to resolve the historical paradoxes we will discuss should be judged by whether someone has a more satisfactory resolution to offer. The reader should be wary throughout to distinguish the contentions I evidence from the contentions to which I lend only plausibility.

"The Flynn effect" is the name that has become attached to an exciting development, namely, that the twentieth century saw

massive IQ gains from one generation to another. To forestall a diagnosis of megalomania, the label was coined by Herrnstein and Murray, the authors of *The bell curve*, and not by myself. I have never done any studies of IQ trends over time in the sense of actually administering tests. Of those who had measured IQ gains here or there, Reed Tuddenham was the first to present convincing evidence using nationwide samples: he compared the mental test scores of US soldiers in World Wars I and II and found huge gains. Had I thought of attaching a name to the phenomenon, I would have offered his.

About 1981, it struck me that if IQ gains over time had occurred anywhere, they might have occurred everywhere and that a phenomenon of great significance was being overlooked. Therefore, I began a survey to see what data existed throughout the developed world. It was on a rather dull Saturday in November 1984 that I found a bombshell in my letter box.

It was data from the distinguished Dutch psychologist P. A. Vroon and some things were evident at a glance. Although Vroon had not developed the techniques to measure them, young Dutch males had made enormous gains in a single generation on an IQ test of forty items selected from Raven's Progressive Matrices. The sample was exhaustive. Raven's was supposed to be the exemplar of a culturally reduced test, one that should have shown no gains over time as culture evolved. These 18-years olds had reached the age at which performance on Raven's peaks. Therefore, their gains could not be dismissed as early maturation, that is, it was not just a matter that children today were about two years ahead of the children of yesterday. Current people would have a much higher IQ than the last generation even after both had reached maturity.

Over a period of twelve months, I was bombarded with data from another thirteen nations all of which showed huge gains. Today the total is almost thirty and includes data from developing nations as well. Our advantage over our ancestors is

relatively uniform at all ages from the cradle to the grave. Whether these gains will persist into the twenty-first century is problematic, at least for developed nations. But there is no doubt that they dominated the twentieth century and that their existence and size were quite unexpected. The very fact they occurred creates a crisis of confidence: how could such huge gains be intelligence gains? Either the children of today were far brighter than their parents or, at least in some circumstances, IQ tests were not good measures of intelligence. Paradoxes started to multiply. Now read on.

2 Beyond the Flynn effect

Yesterday upon the stair
I saw a man who wasn't there
He wasn't there again today
How I wish that man would go away
(Nursery rhyme)

I will try to make the problems posed by IQ gains go away, but do not really think that I can say the final word. I claim only that I can at last propose an interpretation that eliminates paradoxes. These paradoxes have been so intimidating as to freeze our thinking about the significance of IQ gains ever since we began to take them seriously (Flynn, 2006a).

Intelligence and the atom

Before I state the paradoxes, there are some concepts to convey. My fundamental line of argument will be that understanding intelligence is like understanding the atom: we have to know not only what holds its components together but also what splits them apart. What binds the components of intelligence together is the general intelligence factor or *g*; what acts as the atom smasher is the Flynn effect or massive IQ gains over time; the best IQ test to exemplify both of these is called the WISC (Wechsler Intelligence Scale for Children).

The WISC has ten subtests that measure various cognitive skills. For example, the Similarities subtest measures the ability to perceive what things have in common; the Vocabulary subtest

measures whether you have accumulated a large number of the words used in everyday life; Information measures your store of general (as distinct from specialized) information; Arithmetic measures your ability to solve everyday mathematical problems (how much change you should have if you bought certain items out of a five-dollar bill); and so forth (see Box 1).

Box 1

The WISC IQ test (The Wechsler Intelligence Scale for Children) has been administered since 1950 to children ages 6 to 16. The ten subtests given throughout most of that period are below (all items used to illustrate the subtests are fictitious but they fairly represent those used on the WISC). They are listed from the subtest with the lowest gains over time to the highest. Information has enjoyed a gain of only 2 IQ points while Similarities shows a gain of 24 points.

Information:	On what continent is Argentina?
Arithmetic:	If 4 toys cost 6 dollars, how much do 7 cost?
Vocabulary:	What does "debilitating" mean?
Comprehension:	Why are streets usually numbered in order?
Picture Completion:	Indicate the missing part from an incomplete picture.
Block Design:	Use blocks to replicate a two-color design.
Object Assembly:	Assemble puzzles depicting common objects.
Coding:	Using a key, match symbols with shapes or numbers.
Picture Arrangement:	Reorder a set of scrambled picture cards to tell a story.
Similarities:	In what way are "dogs" and "rabbits" alike?

There is a strong tendency for performance on these ten subtests to be inter-correlated. This means that people who are above average on one of them tend to excel on them all, that is, those who are good at seeing what concepts have in common and good at identifying the missing piece of a pattern tend to be the same people who accumulate large vocabularies, large funds of general information, and arithmetical skills. That is why we speak of a general intelligence factor or *g*. Naturally, there are other factors: some people are particularly good at the verbal portions of IQ tests, or the quantitative portions, or the items that require spatial visualization. I will largely ignore these subordinate factors because they pose no problem beyond that posed by the *g* factor.

There is nothing mysterious about the notion of *g*. In everyday life, all of us talk about general abilities that "lie behind" the fact that someone excels at a wide range of tasks or is superior in a wide range of traits. We talk about good people and mean that there are people who are above average not just in terms of kindness but also in terms of generosity and tolerance, so they have moral *g*. We have all said of someone that they have athletic ability and meant that they seem to excel at all sports not just at one, so they have athletic *g*. If someone is good at playing a wide variety of musical instruments, we tend to say that they are "musical," which is to say they have musical *g*. Similarly, if someone is good at a wide range of cognitively demanding tasks, we say that they have general intelligence or *g*(IQ).

A mathematical technique called factor analysis measures this tendency of performance on a wide variety of cognitive tasks to be inter-correlated and, technically, *g* is the quantified result. The *g* factor explains a surprising amount of individual differences in performance on the WISC subtests, but it is better at predicting performance on some rather than others. This is because good performers consistently open up a larger gap on the average person at some cognitive tasks than others. These tend to be the more cognitively complex tasks, which reinforces the claim of *g* to be a

measure of general intelligence. For example, a high-IQ person excels less on Digit Span forward, which is just remembering numbers in the order in which they were read out, and excels more on Digit Span backward, which is repeating numbers aloud in reverse of the order in which they were read out. The ten WISC subtests can be ranked in terms of their *g* loadings. That simply means you rank them from the subtest on which high-IQ people beat the average person the most down to the subtest on which they excel the least.

Once again, there is nothing mysterious about various traits or tasks having different *g* loadings. In the American South of my youth, people who were good tended to be farther above average in terms of kindness than tolerance, which is to say that kindness had a higher *g* loading than tolerance. Musical people tend to be farther above average on the piano than the drums. A talented cook is likely to exceed me more in making a soufflé than scrambled eggs because the former is more complex than the latter. Therefore, it is a better test of excellence in cooking.

The pervasiveness of the *g* factor creates certain expectations. If there is such a thing as general intelligence, and if it were to increase over time, we would expect gains on each of the ten WISC subtests to tally with their *g* loadings. With the exception of Coding, the *g* loadings are very similar on the various WISC subtests. But when we turn to IQ gains over time, we find something surprising: huge discrepancies between the magnitude of subtest gains and subtest *g* loadings. For example, Similarities and Information have much the same *g* loadings, yet the former shows twelve times the gains of the latter. Remember cooking. If cooking skills improved over time, it would be amazing if the *g* loadings were ignored, for example, if there was a huge gain in scrambling eggs but no gains in making soufflés.

Figure 1 presents a summary of IQ gains in America between 1947 and 2002. The WISC data are most complete for America but I could have chosen another developed nation such as France or Britain. Another test that will be important later is

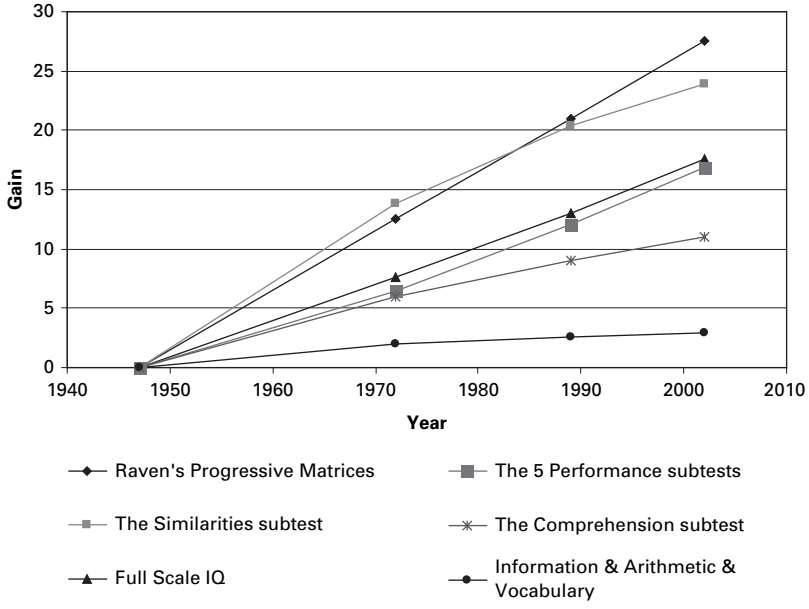

Figure 1 This figure shows WISC IQ gains starting in 1947–1948 and running through 2002. The test was updated three times, which means we get estimates of gains over three periods of 13 to 25 years. All gains are measured in IQ points (with SD set at 15). See Appendix I for Table 1 on which the figure is based. I have also included an "estimate" of American gains on Raven's. There are no reliable US data, but there is a huge literature showing that Raven's gains have proceeded at no less than 0.50 IQ points per year in every developed nation for which we have data. I will list these nations and give the years the data cover:

Belgium: 1958–1967 (Flynn, 1987, Table 2)
Norway: 1954–1980 (Flynn, 1987, Table 4)
The Netherlands: 1952–1982 (Flynn, 1987, Table 1)
Israel: 1971–1984 (Flynn, 1998b, Table 3)
Britain: 1942–1992 (Flynn, 1998a, Figure 3)
Argentina: 1964–1998 (Flynn & Rossi-Casé, under review)

Raven's Progressive Matrices, so I have given a minimal estimate of US gains on Raven's. As the caption to Figure 1 says, there are no good data on Raven's gains in America, so I have used a minimal estimate closely tied to US gains on Similarities. Data from Great

Britain show the two rising in tandem (Flynn, 1998a, Figure 3; Flynn, 2000b, Table 1).

Some trends to note in Figure 1. The various subtests show very different gains: Americans gained 24 points on Similarities between 1947 and 2002 (1.6 SDs), 4 points on Vocabulary, and only 2 points on Arithmetic and Information (for an average of 3 points on these three subtests collectively). The WISC gives not only subtest scores but also a summary judgment on our intelligence called Full Scale IQ. These gains are huge, amounting to about 18 points. The posited gains on Raven's come to fully 27.5 points. How can our recent ancestors have been so unintelligent compared to ourselves? Even worse, we will look at British data that suggest we have to extend these gains all the way back to 1900. So our distant ancestors must have been very stupid indeed. We are now in a position to state three paradoxes and I will throw in a fourth for good measure.

Stating the paradoxes

(1) The factor analysis paradox: how can intelligence be both one and many at the same time or how can IQ gains be so contemptuous of *g* loadings? How can people get more intelligent and have no larger vocabularies, no larger stores of general information, no greater ability to solve arithmetical problems?

(2) The intelligence paradox: if huge IQ gains are intelligence gains, why are we not stuck by the extraordinary subtlety of our children's conversation? Why do we not have to make allowances for the limitations of our parents? A difference of some 18 points in Full Scale IQ over two generations ought to be highly visible.

(3) The mental retardation (MR) paradox: if we project IQ gains back to 1900, the average IQ scored against current norms was somewhere between 50 and 70. If IQ gains are

in any sense real, we are driven to the absurd conclusion that a majority of our ancestors were mentally retarded. In passing, we are in a transitional period in which the term "mentally retarded" is being replaced by the term "mentally disabled" in the hope of finding words with a less negative connotation. I have retained the old term for clarity and because history has shown that negative connotations are simply passed on from one label to another.

(4) The identical twins paradox: there is no doubt that twins separated at birth, and raised apart, have very similar IQs, presumably because of their identical genes. Indeed a wide range of studies show that genes dominate individual differences in IQ and that environment is feeble. And yet, IQ gains are so great as to signal the existence of environmental factors of enormous potency. How can environment be both so feeble and so potent?

We will address each of these paradoxes in turn but it may help to signal the solutions in shorthand:

(1) The WISC subtests measure a variety of cognitive skills that are functionally independent and responsive to changes in social priorities over time. The inter-correlations that engender *g* are binding only when comparing individuals within a static social context.

(2) Asking whether IQ gains are intelligence gains is the wrong question because it implies all or nothing cognitive progress. The twentieth century saw some cognitive skills make great gains, while others were in the doldrums. To assess cognitive trends, we must dissect "intelligence" into solving mathematical problems, interpreting the great works of literature, finding on-the-spot solutions, assimilating the scientific worldview, critical acumen, and wisdom.

(3) Our ancestors in 1900 were not mentally retarded. Their intelligence was anchored in everyday reality. We differ