Cambridge University Press
978-1-108-83713-2 — Understanding Intelligence
Ken Richardson
Excerpt
More Information

1 Testing, Testing

When people consider intelligence, they will first tend to think of IQ, and scores that distinguish people, one from another. They will also tend to think of those scores as describing something as much part of individuals' make-up as faces and fingerprints. Today, a psychologist who uses IQ tests and attempts to prove score differences are caused by genetic differences will be described as an 'expert' on intelligence. That indicates how influential IQ testing has become, and how much it has become part of society's general conceptual furniture.

And yet, it's never sat comfortably among us. Everyone – including experts – will agree that, whatever intelligence is, it's bound to be complex, enigmatic, and difficult to describe: probably the most intricate function ever evolved. To this day, psychologists argue over what it actually is. So how has measuring it become so easy and apparently so convincing? Most IQ tests take around half an hour, though they vary a lot. Some researchers claim to do it in a few minutes, or over the telephone, or online. How do they get away with that?

Well, for three reasons, I think. First, differences readily match common social observations, much as a model of the universe once matched everyone's experience of sun and stars going around the earth. In the case of intelligence, we sense it comes in grades, with people arranged on a kind of ladder. Second – and partly because of that – it's been easy for scientists to propose a *natural* ladder; that is, intelligence based on our biological make-up. That makes differences real and immutable. Third, IQ testing became influential because it’s been so useful as a socially practical tool. Psychologists have boasted about that for a long time.

Indeed, there’s a long history behind all of that. Proving that the intelligence we ‘see’ socially is perceived accurately, and is biologically inevitable, was the task of polymath Herbert Spencer. Writing in utilitarian Victorian Britain of the 1850s, one biographer described him as ‘the single most famous European intellectual’ of his time. Spencer did try to create a theory of intelligence, but didn’t get far. He was sure, however, that individual and group differences must originate in the physiology of the brain, and due more ‘to the completion of the cerebral organization, than to the individual experiences’. ‘From this law’, he went on, ‘must be deducible all the phenomena of unfolding intelligence, from its lowest to its highest grades’.

Spencer’s method was still theoretical. He read Darwin, and coined the term ‘survival of the fittest’. That idea was later to inspire a wider eugenics movement favouring policies like selective breeding. From common observation he told us that ‘the minds of the inferior human races, cannot respond to relations of even moderate complexity’. By the same token, helping the poor and weak in British society flew in the face of nature, he said – they should be allowed to perish. That’s also practical, in a way, if rather blunt. But sharper tools were on the horizon.

An Unnatural Measure

Gentleman scientist Sir Francis Galton followed much of Spencer’s drift. He was, in addition, very practically minded. In possession of a fortune (inherited), Galton was able to indulge many interests. Travels in Africa in the 1850s had already convinced him of the mental inferiority of its natives. He had read Charles Darwin’s theory, conversed with Spencer, and became convinced that there is something he called ‘natural ability’. It varies substantially among people, he argued, just like height and weight, and is distributed like the bell-shaped curve. Later he observed that members of the British establishment were often related to each other. That convinced him that differences in intelligence must lie in biological inheritance, which also implied that society could be improved through eugenics, or selective breeding programmes. However, he realised, that would need some measure ‘for the indications of superior strains or races, and in so favouring them that their progeny shall outnumber and gradually replace that of the old one'.

Galton was amazingly energetic and inventive. He believed that differences in natural ability, being innate, must lie in neurological efficiency, or the 'physiology of the mind' – functions obviously hidden and unidentified. He reasoned, however, that responses to simple sensorimotor tests could provide a window on those hidden differences. He even set up a special laboratory and got people to pay for the fun: reaction times, speed of hand motions, strength of grip, judgements of length, weight, and many others, which all provided the data he sought.

Of course, individuals varied in their scores. But how could he convince people that they were really differences in the unseen intelligence? He already had an idea: 'the sets of measures should be compared with an independent estimate of the man's powers', he said. The individuals' social status and reputation were what he had in mind. As he put it in *Hereditary Genius* (1883), 'my argument is to show that high reputation is a pretty accurate test of high ability'.

And that was it: a numerical surrogate of human worth, of 'strains and races', in a few quick tests of sensation, speed, and motion. If you're thinking about it, though, you might see some suspicious circularity in the logic and want to ask Sir Francis a few questions, like:

- How do you know that reputation is a good indication of natural ability (and not, for example, a consequence of social background)?

Possible answer: It's the only one I can employ (he actually said that) – or, more honestly, we don't really know.

- If you know in advance who is more or less intelligent, why do you need the test?

Possible answer: For mass testing. Also, because numbers look objective and scientific.

- If the tests are chosen to agree with what you already know, how can it possibly be more accurate?

Possible answer: It logically cannot be; but it looks as if it is.

These questions refer to the 'validity' of a test. Does it measure what it claims to measure? Or do we really know what score differences are differences in? Those questions have hung over intelligence testing like a dark cloud ever since. In this chapter, I hope to show you how IQ testers have dealt with it, and how that has entailed a very special understanding of intelligence. To get a better idea of validity, though, let's briefly compare the strategy with real tests of physiology, as in biomedical tests.

Physiological Testing

Like psychologists, real physiologists need to describe hidden causes of observed differences, especially in disease conditions. So they have long pored over whatever fluids, excretions, secretions, or expectorants they could coax from the insides to compare with outer symptoms.

Urinalysis was practised even in Ancient Greece, in the time of Hippocrates. In the Middle Ages, flasks and charts were carried by all respectable physicians, duly called 'pisse prophets', to assist diagnosis. Colour charts, used well into the nineteenth century, as well as notes on smell (and sometimes taste!) were the first attempts at standardised tests of physiology. Diabetic urine was noted for its 'exceeding sweetness'. It's thought that this is where the expression 'taking the piss' stems from.

Of course, the tests were rough and ready. But validity improved for a simple reason. Scientific research painstakingly revealed the true nature of the internal functions, including the many detailed steps in urine production, where they can go wrong, and how that is reflected in the chosen markers. So today we have a formidable array of valid tests of physiological functions. We can reliably rate differences in the measure on the 'outside' – even the colour of urine – to the unseen functional differences on the 'inside'. We can understand what variations in cholesterol measures mean. Likewise with blood pressure readings; why a white blood cell count is an index of levels of internal infection; and why a roadside breathalyser reading corresponds fairly accurately with level of alcohol consumption.

'We Classify'

It turned out that Galton's test didn't work, anyway. Differences between upper class and tradesmen, having experienced contrasting conditions in development, would have been unsurprising. But they turned out to be tiny. For example, reaction time to sound was 0.147 versus 0.152 seconds. For 'highest audible sound', Galton recorded 17,530 versus 17,257 vibrations per second. But what about intellectual differences? In the USA, doctoral student Clark Wissler tried to correlate results from Galton's tests with academic grades of university students. There was virtually no correlation between them. And the test scores did not even appear to correlate with each other. It is not possible, Wissler said, they could be valid measures of intelligence. The physiologists of the mind needed another approach.

It so happened that, around that time, the early 1900s, a psychologist in Paris was also devising psychological tests, though of a different kind and for a different purpose. Parisian schools were now admitting more children from poorer backgrounds, and some might struggle with unfamiliar demands. Alfred Binet was charged by the local school board to help identify those who might need help.

Like Galton, Binet devised series of quick questions and mental tasks. But he was looking for ones related to school learning rather than physiology. He got them, quite sensibly, from close observation of classroom activities, by devising short questions and tasks to reflect those activities, and then trying them out. Each item was deemed suitable, or not, according to two criteria: (1) whether the number of correct answers increased with age; and (2) whether a given child's performances matched teachers' judgements of his or her progress.

Binet and his colleague Henri Simon produced their first *Metrical Scale of Intelligence* in 1905. It contained 30 items, designed for children aged 3–12 years, arranged in order of difficulty. By 1911 the collection had expanded to 54 items. Here are some examples expected to be passable by two age groups:

Five-year olds:

- compare two weights;
- copy a square;

- repeat a sentence of 10 syllables;
- count four pennies;
- join the halves of a divided rectangle.

Ten-year olds:

- arrange five blocks in order of weight;
- copy drawings from memory;
- criticise an absurd statement;
- answer sentence-comprehension questions;
- use three given words into a sentence.

Average scores for each age group were calculated. The 'mental age' of individuals could then be worked out from how many items they could do. If a child achieved a score expected of six-year-olds, they would be said to have a mental age of six. A child passing them all would have a mental age of 12. Binet suggested that a deficit of two years or more between mental age and chronological age indicated that help was needed. Finally, in 1912, the German psychologist William Stern proposed the use of the ratio of mental age to chronological age to yield the now familiar intelligence quotient, or IQ:

$$\mathrm{IQ} = \frac{\text{mental age}}{\text{chronological age}} \times 100.$$

So IQ was born. But it is important to stress the narrow, practical, purpose of Binet's test: 'Psychologists do not measure . . . we classify', he said. However, it had an unintended, but hugely portentous, quality. By its nature it produced different scores for different social classes and 'races'. Galton's followers soon claimed Binet's to be the test of innate intelligence they had been looking for. Translations appeared in many parts of the world, especially in the USA. Binet himself condemned the perversion of his tests as 'brutal pessimism'.

Original Mental Endowment

In the USA, Henry H. Goddard translated the Binet–Simon test into English in 1908, and found it useful for assessing the 'feebleminded'. As a eugenicist, Goddard worried about the degeneration of the 'race' (and nation) by the mentally handicapped, and also by the waves of new immigrants from

Southern and Eastern Europe. He was commissioned to administer the test to arrivals at the immigration reception centre on Ellis Island. Test scores famously suggested that 87 per cent of the Russians, 83 per cent of the Jews, 80 per cent of the Hungarians, and 79 per cent of the Italians were feebleminded. Demands for immigration laws soon followed.

Lewis Terman, a professor at Stanford University, developed another translation of Binet's test in 1916. He enthused over the way it could help clear 'high-grade defectives' off the streets, curtail 'the production of feeblemindedness', and eliminate crime, pauperism, and industrial inefficiency. By using his IQ test, he said, we could 'preserve our state for a class of people worthy to possess it'. Binet's screen for a specific category thus became scores of the genetic worth of people in general. 'People do not fall into two well defined groups, the "feeble minded" and the "normal"', Terman said. 'Among those classed as normal vast individual differences exist in original mental endowment.' He, too, called for eugenic reproduction controls, which soon followed. Galton's programme had found its measure.

Mass Testing

Terman's test was applied to individuals, one at a time. During World War I, however, the US Army wanted to test recruits in large numbers from many different backgrounds. A group led by Robert Yerkes constructed two pencil-and-paper tests: one for those who could read and write English; the other for those who could not. These were quite ingenious, including tasks like tracing through a maze, completing a picture with a part missing, and comparing geometrical shapes. Up to 60 recruits could be tested at a time, taking 40–50 minutes. In *Army Mental Tests*, published in 1920, Clarence Yokum and Robert Yerkes claimed that the test was 'definitely known … to measure native intellectual ability'.

These group tests set the scene for mass IQ testing in populations generally, and for the spread of the ideology underlying it. Up to the early 1930s the IQ message became useful in the USA in the passing of compulsory sterilisation laws, immigration laws, and the banning of 'inter-racial' marriage. Hitler's ministers in Nazi Germany were impressed by America's IQ testing regimes, and its eugenics policies. They took the message away with even more deadly consequences, as we all know.

And in Britain

In Britain, the new intelligence tests were just as energetically promoted. Eugenics movements were popular, and, in 1911, a report to the *Board of Education* recommended their use for the identification of 'mental defectives'. By the late 1930s psychologists like Cyril Burt were urging their use in the British 11+ exam. 'It is possible at a very early age', they advised, 'to predict with accuracy the ultimate level of a child's intellectual power', and that 'different children . . . require (different) types of education'.

Since then, IQ testing has developed rapidly into the huge enterprise it is today. But constant controversy has hinged on the same burning question. Do IQ tests really measure what they claim to measure – even if we aren't sure what that is? It's important to see how IQ testers have dealt with the question, and how they have dodged it.

Validity Vacuum

In his book *IQ and Human Intelligence*, Nicholas Macintosh stated: 'If you are trying to devise a test that will measure a particular trait . . . it will surely help to have a psychological theory specifying the defining features of the trait, to ensure that the test maps on to them.' That means being clear about what differences on the 'outside' (the measure) really mean in terms of differences on the 'inside'. That's what we expect of modern physiological and biomedical tests, after all. Such mapping is called 'construct validity'. As mentioned earlier, simple definitions do not do that.

Lewis Terman used previously translated items from Binet's test, but added many more of similar types: memory span, vocabulary, word definition, general knowledge, and so on. Now called the Stanford–Binet, the test soon became the standard on both sides of the Atlantic. Like Binet's test items, they obviously tended to reflect school learning and a literary/numerical mindset, itself related to social background. But there is no systematic attempt at construct validity.

As regards the *Army Mental Tests*, mentioned above, Yokum and Yerkes duly emphasised (on page 2!) that the test 'should have a high degree of validity as a test of intelligence'. But little more was said about it. Instead, correlations of scores with those on Terman's test, and with teachers' ratings, as well as social class, are given.

David Wechsler devised the *Adult Intelligence Scale* in 1939, and then the *Wechsler Scale for Children* in 1949. Wechsler had worked on the Army tests and used similar items. Through a number of revisions, these tests have rivalled the Stanford–Binet in popularity. As for validity, as psychologist Nicholas Mackintosh noted, 'Wechsler had little evidence [of] the validity of his tests', except their correlation with Stanford–Binet scores and teachers' ratings. He quotes Wechsler to say, 'How do we know that our tests are "good" measure of intelligence? The honest answer we can reply is that our own experience has shown them to be so.' As apology, Wechsler mentions only that the same applies 'to every other intelligence test'.

These are the most popular, and almost the 'gold standard', tests. A huge number of other IQ-type tests have been constructed, and I will refer to some below. But they are almost all collections of items closely related to school learning and literacy and numeracy. Testers almost always imply that they're measuring learning *ability*, not simply learning. What that is has not been made clear. If you get the chance, have a look at the test manuals and search for any claims about test validity. What you will almost always get is reference to one or other of four substitutes for it. Let us look at these in turn.

Score Patterns

IQ test items are devised by individual psychologists introspectively, according to the cognitive processes they imagine will be invoked. They are then included or rejected in a test after exhaustive trial-and-error procedures known as item analysis. The aim is to end up with a pattern of average scores that match expectations of individual differences in intelligence. An item is included if, as seen in trials, it more or less contributes to such a pattern. Otherwise it is rejected. Although the pattern is thus 'built in', it is then taken to confirm test validity.

For example, test constructors have assumed that intelligence, as a ‘biological’ trait, should be distributed like physical traits, according to the bell-shaped curve (Figure 1.1). The pattern was achieved by Lewis Terman, after preliminary trials, by a simple device: rejecting items that proved to be either too easy or too difficult for most people. The practise has continued, and the results have been taken to confirm test validity, instead of being an artefact of test construction. In fact, many physiological variables, including many in the brain, are not distributed like that (see Chapter 6).

Test items are also selected such that progressively more of each age group respond correctly. That makes scores resemble those of a simple physical trait such as height or foot size. However, the school-related content of many items ensures that average test scores indeed increase with age – as desired – and then steadily decline thereafter (Figure 1.2). Even today, researchers worry about the declining average intelligence of adults, and are looking

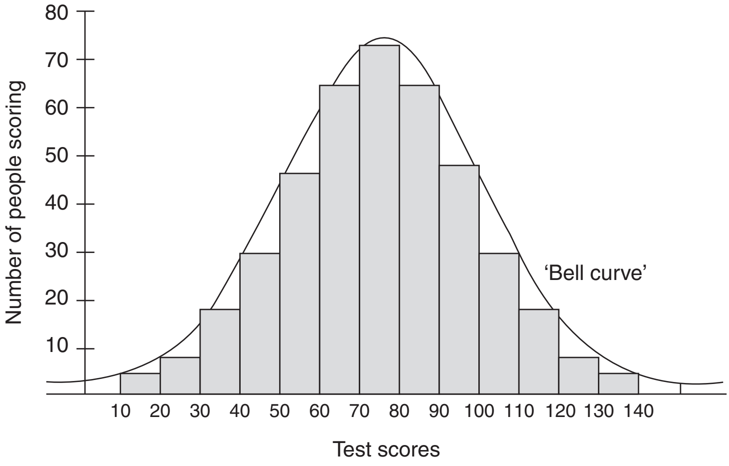

Figure 1.1 The famous bell-shaped curve ‘built in’ to the IQ test. Raw scores usually get converted statistically to IQs with a mean of 100.