Part I History, Theory and Utility

1 The History of Psychometrics

Craig Knight

He had the personality of kipper; on an off day.

Joan Collins

Think about the people you know for a moment. Have you ever wondered how Chris manages to maintain a sense of equilibrium under even the most testing circumstances, or why Sam is more irritating than a starched collar? Why are some people like balm to a wound, while others look to start a fight in an empty room? And wouldn't it be useful if you could predict people's behaviour patterns *before* an event rather than ruefully mopping up afterwards?

Humans have been speculating on and assessing their own variables since Cain weighed up Abel, often with the success of somebody nailing fog to a wall. If it's hard to judge those we claim to know best, just how can you assess the personality of a good accountant, manager or leader? Of course Tibetan Buddhists re-select the same leader on an eternal basis. The rest of us have to make a more or less educated assessment of the candidates available.

It is this assessment that is central to psychometrics. If we accept the definition of psychometrics as 'the science of measuring mental capacities and processes' (en.oxforddictionaries.com, 2016) then the *quality* of that science becomes the predictor of its success.

As we will see, psychometrics is a flawed discipline. Its advocates can be vociferous and wrong. Vaunted predictive capabilities go unchecked and snake oil oozes from the cracks of many psychometric creations. No matter how persuasive the personality advocate and how beguiling the evidence, we do well to remember that nobody ever equates to a yellow circle, a traffic light or a bear. Only decent instruments – probably in the hands of trained assessors – can link skills, propensities and personalities to jobs, proclivities and outcomes.

Well-researched psychometrics can test for the qualities required in a boardroom or back office or bakery. So while these tools – like all tools – arrive in various shades of imperfection, their lack during times of recruitment and appraisal can be costly. This chapter will explore the origins and development of psychometrics, its uses and abuses. It will close by reading the runes of future developments.

GREAT MEN AND THEIR HUMOUR

From when time was in its cradle people have believed that personality traits can be divined. The gift of leadership was particularly prized. Leaders were said to have natural charisma and

ability which others instinctively lacked. Even as infants leaders waved their rattles like sceptres (Haney, Sirbasku & McCann, 2011). Thus followers innately looked to trail behind, while women were 'fitted to be at home as is their nature' (Buss & Schmitt, 2011). Scientifically illiterate though these ideas may be (Haslam, 2004), moot them in the Red Lion and witness the levels of assent amongst the crowd. The idea of a born leader remains powerfully salient. With due deference to Meir, Thatcher and Merkel, as Carlyle had it (1841, p. 47), 'The history of the world is but the biography of great men'.

However, even a cursory look at different leaders' personalities reveals considerable variety within the camps. Alexander the Great's propensity for megalomania would have sat poorly with Nelson's service ethic; Kublai Khan's extravagance is unlikely to have appealed to Karl Marx, while Mahatma Gandhi's peaceful resistance would probably leave Emperor Hadrian somewhat perplexed. Discussion over the cornflakes would have been tense. And the same differences of approach are found amongst carpenters, midwives and tennis players. So how does any instrument assess for role, aptitude and skill?

PERSONALITY AND THE FOUR HUMOURS

Many of the chapters of this book will explore how various instruments gauge aspects of personality. Even between the most widely respected psychometric tools the number of perceived personality traits varies widely and runs from five to 32. However, originally there were just four.

It is a matter of conjecture whether a belief in the need for bodily balance was developed by the Indian Ayurveda system of medicine or by the Ancient Greeks. What is certain is that the concept of four distinct bodily fluids – hydraulically interdependent and all influencing human nature – survived from Hippocrates through Galen and the Roman Empire, right through to the Renaissance. Indeed we retain much of the terminology today. To be *sanguine*, *choleric*, *phlegmatic* or *melancholy* is to echo a system of personality assessment that resonates through the centuries (Figure 1.1).

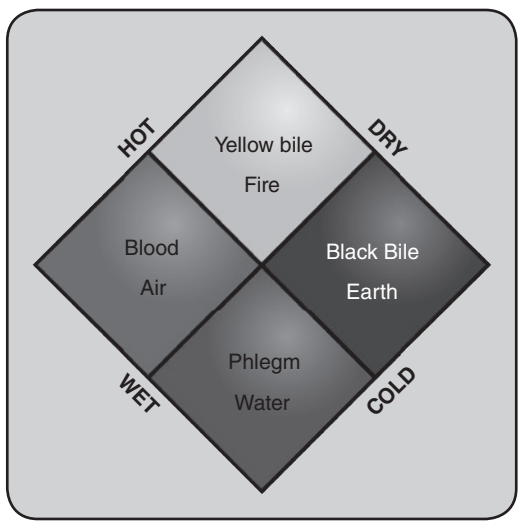

Figure 1.1 *The four humours*

A surplus or deficiency of any one of four elemental bodily fluids – or *humours* – was thought to directly affect one's feelings and health. All four humours may originate from just one bodily fluid: blood. In the open air blood sedimentation shows a dark, thick clot at its base (black bile), and erythrocytic cells (or red blood) sit on top below a layer of white blood cells which could easily have been labelled as phlegm. Phlegm was not the expectorated gloop we know today. Finally a top pool of yellow liquid (yellow bile) completes the basic substances which were thought to comprise the corporeal human.

An excess of yellow bile was expressed through overt aggression, an issue said to be associated with an agitated liver. Even now we will call somebody who is peevish and disagreeable 'liverish' or 'bilious', while alternative medicine often insists that anger remains a symptom of a disturbed liver (Singh & Ernst, 2008).

Meanwhile those said to have an excess of what the Greeks called *melaina kholé*, or black bile, were said to be suffering from 'melancholy' or depression. An excess of phlegm was thought to be behind a stolid, fixedly unemotional approach to one's affairs, and gave rise to the modern phlegmatic personality.

In contrast to the other three humours an excess of blood carried clear personality benefits. People who are *sanguine* (from the Latin *sanguis,* 'blood') have always been cheerful, optimistic and confident.

Each individual had their own humoral composition, which they shared to a greater or lesser degree with others. This mix of humours precipitated personality in a view that held good from Hippocrates to Harvey via Ancient Rome and Persia. Indeed, this holistic approach is still used in *personality type* analysis today, where psychometricians are keen to label individuals with marks of similarity (Pittenger, 1993).

Thus, while it is considered pseudo-scientific to tell somebody that they possess a mostly phlegmatic personality (Childs, 2009), you are very likely to hear that you have the temperament of a team worker, or of an introvert, or that you have a blue/green personality. You may even be assigned a group of incongruous-sounding letters such as ENTJ from the globally dominant Myers–Briggs Personality Type Indicator. Amongst other attributes ENTJs are 'born leaders' (personalitypage.com, 2015). And we see the ancient terminology being recycled in the twenty-first century, even when it is known to be psychologically flawed. So are some modern interpretations any less pseudo-scientific than their rather longer-lasting forebears (Sipps, Alexander & Friedt, 1985)?

THE BEGINNINGS OF MODERN PSYCHOMETRICS

The history of psychometrics intertwines with that of psychology. Its modern incarnations have two main progenitors. The first of these concentrates on the measurement of individual differences; the second looks at psychophysical measurements of similarity.

Charles Darwin's (1809–82) *The Origin of Species* (Darwin, 1859) explained why individual members of the animal kingdom differ. It explored how specific characteristics show themselves to be more successful and adaptive to their environment than others. It is these adaptive traits that survive and are passed on to successive generations.

Sir Francis Galton (1822–1911) was a Victorian polymath whose panoply of accomplishments encompassed sociology, psychology and anthropology. He was also related to Charles Darwin and was influenced by his half-cousin's work. Consequently Galton wondered about various adaptive traits in human beings. Not content with merely studying the differences, however, Galton wanted to *measure* them.

In his book *Hereditary Genius* (1869), Galton described how people's characteristics make them more or less *fit* for society and for positions within it. Galton – often called 'the father of psychometrics' – was drawn to measuring intelligence, as was Alfred Binet (1857–1911) in France (Hogan & Cannon, 2007). This work was later taken up by James McKeen Cattell (1860–1944), who coined the term *mental test*.

As Darwin, Galton, Binet and Cattell developed their measures of fitness and intelligence, Johann Herbart – a German philosopher and psychologist – was also working to scientifically unlock 'the mysteries of human consciousness' (Wolman, 1968). Herbart was responsible for creating mathematical models of the mind in his field of psychophysics. Psychophysics influenced Wilhelm Wundt, who was often credited with founding the science of psychology itself (Carpenter, 2005). Thus Herbart, via Wundt, and Galton, via Cattell, have strong claims to be the pioneers of modern psychological testing.

THE TWENTIETH CENTURY

The twentieth century saw psychometrics become increasingly reliable, valid and robust. Louis Thurstone, founder and first president, in 1936, of the Psychometric Society, developed the *law of comparative judgement*, a theoretical approach to measurement that owed much to psychophysical theory. Working with statistician Charles Spearman, Thurstone helped to refine the application and theory of factor analysis, a statistical method that explores variability and error without which psychometrics would be greatly diminished and considerably less accurate (Michell, 1997).

Working at the same time, Hungarian psychiatrist, Leopold Szondi was in something of a revolt against this forensic but narrow statistical treatment of people's psyche. He did not believe that the make-up of something as complex, changeable and irrational as a human being could be captured by a series of focused numbers, no matter how thorough the statistics that underlay them (Szondi, Ulrich &Webb, 1959).

In developing his own, eponymous test, Szondi instead tried to capture as much of the essence of the spirit of humankind as possible by widening the assessments that were made. The test's goal was to explore the innermost recesses of our repressed impulses. Constructs were elicited by assessing the levels of sympathy or aversion engendered by showing clients specific photographs of psychopaths. The client was expected to point to the person she or he would least like to meet on a dark night and explain why (Szondi et al., 1959).

Szondi held that the characteristics of – and emotions in – others that bother us are those that most disturbed us early in our lives. That is why we repress these factors in ourselves. His test is said to address fundamental drives which classify the entire human system but in a more qualitative manner than instruments offered by his psychometrician contemporaries.

In this gestalt approach Szondi is closer in spirit to Hermann Rorschach, the Swiss Freudian psychiatrist and psychoanalyst. Rorschach developed perhaps the most famous psychological instrument the world has seen. The Rorschach inkblot test assesses clients' perceptions of a series of patterned smudges, some of which are shown in Figure 1.2 (Wood, Nezworski,

Figure 1.2 *Ink blots from the Rorschach test. © Zmeel Photography/iStockphoto*

Lilienfeld & Garb, 2003). There were ten original inkblots, which Rorschach presented on separate white cards, each approximately 7 × 10 inches in size, each with near-perfect bilateral symmetry. First, the client interpreted the shapes in a *free association* phase. 'Oh, that one looks like a prehistoric moth …', and so forth. Then the cards were presented in a fixed order, and held, rotated and pored over by the client, who was quizzed at each stage. Responses were tabulated.

Rorschach wanted his test to act as a series of pegs upon which aspects of human personality could be hung. The interpretation of Rorschach is both complex and contested. Rorschach interpreters are effectively on probation for up to four year before being considered sufficiently competent to handle the test alone. Nevertheless some critics consider the interpretation of odd blobs nothing more than pseudo-science (Wood et al., 2003). Even so, the Rorschach test, like Freud, the man who inspired Rorschach himself, may be flawed and a little past its peak, but it continues to be very influential – one of the tests most used by members of the Society for Personality Assessment (Gacano & Reid,1994).

MEASUREMENT, CONTROVERSY AND THEORETICAL DEVELOPMENT

The split between the preferred types of psychometric assessment grew. At the same time, the importance of accurate psychometric measurement became ever more key and contentious. Even the definition of measurement itself caused argument.

In 1946 Stanley Smith Stevens defined measurement as 'the assignment of numerals to objects or events according to some rule' (Michell, 1997). At first glance this definition benefits from a certain vagueness, useful to some social scientists but slightly and importantly different from the definition used by physical science, where measurement is 'the estimation or discovery of the ratio of some magnitude of a quantitative attribute to a unit of the same attribute' (Michell, 1997).

An opposite view quickly formed. This was that as physicists and psychologists were both scientists there should be no convoluted semantic differences between how they measure their inputs, throughputs and outputs (Hardman, 2009).

While picking up the niceties of measurement can be a little like eating consommé with a fork, the different theories themselves are happily salient. *Classical Test Theory* grew from the combination of three mathematical developments and the genius of Charles Spearman in the early twentieth century (Novick, 1966). First, there was the realisation that there are errors when people are measured. If, before an assessment, you have slept like a contented elephant and eaten a hearty breakfast, you are likely to feel and perform differently than had you rolled in from an all-night party, unwashed, unrested and unfed. Second, it is not always possible to predict where the error will occur (you might perform brilliantly when hung over) and, third, some aspects of your performance are usefully correlated while others are not. You may, for example, be happier in the morning than in the afternoon, so your happiness and the 24-hour clock are correlated *and* linked. However the freshness of the morning milk also correlates with your moods, but the correlation is incidental and unlinked.

By harnessing the mathematics to the psychometrics, Classical Test Theory was able to improve the predictive power of psychological testing. It used people's performances to feed back into the reliability and validity of the instruments. It made useful estimates as to how psychometric performances would translate into real-world successes (Novick, 1966).

However, a major flaw in Classical Test Theory is that the characteristics of the test taker and the characteristics of the test itself are impossible to separate. Each can only be interpreted in the context of the other. Furthermore, the standard error of measurement (which is the difference between what you would score on a test in ideal conditions – your true score – and the score you did achieve in the conditions prevailing at the time of the test) is assumed to be the same for everybody, regardless of mood swings or innate personality stability.

During the 1950s and 1960s three men working, independently but serendipitously, on parallel research led to the development of *Item Response Theory*. Danish mathematician Georg Rasch, American psychometrician Frederic Lord and Austrian sociologist Paul Lazarsfeld developed a framework for evaluating how well psychological assessments work, and how valid specific items within these assessments may be.

Item Response Theory is also known as *latent trait modelling*. This is because IRT models the relationship between concealed, or latent, traits within a test taker and the responses that a test taker makes to test items. Thus somebody's sociability can be assessed by asking questions such as 'Do you enjoy meeting people?' and 'Do you take the initiative in making new friends?' (Cook & Cripps, 2005).

Traits, constructs or attributes therefore do not need to be directly observed, but can be inferred from the responses given. Item Response Theory is argued to be an improvement over Classical Test Theory (Uebersax, 1999). IRT is said to provide a basis for obtaining an estimate of comparisons between related but different groups with varying levels of ability.

For example, a chemistry graduate's knowledge of her or his subject can be examined via a university test. The result can then be reliably compared to the test result of a senior school pupil sitting a similar but easier examination. By contrast, Classical Test Theory relies on comparisons with a norm group (a norm group is a collective representation of a relevant group, such as 'graduates', 'taxi drivers' or 'senior managers'), so that while there are comparisons within groups, there is no *relative* comparison between groups.

Item Response Theory is especially popular in education. It is used in designing, comparing and balancing examinations across disciplines and age groups. It is, perhaps, at its best in computerised adaptive testing where questions change with and mould to the test taker's ability level (Lord, 1980).

While Classical Test and Item Response Theories compete for psychologists' and statisticians' attentions, *Generalisability* – or *G – Theory* is now staking its claim. In the 1960s another Swiss psychologist, Jean Cardinet, began to explore the specificity and generalisability of data (Cardinet, 1975). G Theory looks at the reliability of measures under specific conditions.

In practice, generalisability allows researchers to explore what would happen if aspects of a psychometric investigation were altered. For example, an opinion poll company could now discover whether assessments of voting intention varied much depending on whether 10, 100, 1,000 or 1,000,000 politically active adults were interviewed. Implications for time and money are plain.

These advancements may not be as clear-cut as they first appear. Classical Test Theory still tends to dominate psychometrics. Most instruments remain norm-based, with comparisons between norms fraught with unreliability. Similarly the most popular statistical packages still present and prepare data in ways, and to standards, that Charles Spearman would recognise. So what of the instruments themselves?

TOOLS FOR THE JOB

The first modern psychometric instruments measured intelligence. Probably the best-known of its type was the Binet–Simon IQ test. At the end of the nineteenth century the French Government introduced universal education. Significantly underperforming children were categorised as *sick* and removed to asylums for their own welfare (Nicolas, Andrieu, Croizet, Sanitioso & Burman, 2013). In 1899, working with Théodore Simon, a psychologist and psychometrician, Alfred Binet looked to develop a way of identifying 'slow' rather than sick children, so that they could be placed in special education programmes instead of being separated from society (Avanzini, 1999).

By testing a wide range of children across many measures, Binet and Simon developed a baseline of intelligence. Their original goal was to find one, clear indicator of intelligence, of general mental excellence. In this, they failed. Instead children were compared within categories and age groups (Fancher & Rutherford, 2012). Binet and Simon were able to set common levels of achievements, and from here developed benchmarks for high and low achievers. They produced a portable, generalisable test that is still in use in modified form today. This categorisation of intelligence within the Binet–Simon test (which became the Stanford–Binet test in 1916) may be seen in Table 1.1 in both its present and its original classification (Bain & Allin, 2005).

Table 1.1 *Stanford–Binet IQ classification 2015 versus final Binet–Simon IQ classification 1916*

Stanford–Binet (2015) IQ range	IQ classification	Binet–Simon (1916) IQ range	IQ classification
145–160	Very gifted or highly advanced		
130–144	Gifted or very advanced	Above 140	Near genius or genius
120–129	Superior	120–140	Very superior intelligence
110–119	High average	110–120	Superior intelligence
90–109	Average	90–110	Normal, or average, intelligence
80–89	Low average	80–90	Dullness, rarely classifiable as feeble-mindedness
70–79	Borderline impaired or delayed	70–80	Border-line deficiency, sometimes classifiable as dullness, often as feeble-mindedness
55–69	Mildly impaired or delayed	Below 70	Definite feeble-mindedness

There is a distinct problem in comparing IQ scores across even a few decades. By convention, the average IQ score is set to 100. When tests are revised – as good practice demands – a new cohort of people take the test. Because of the passage of time, new test takers tend to be from younger generations than their predecessors. In almost every case new average scores are significantly above 100 and means have to be revised (Flynn, 2009). The average rate of increase seems to be about three IQ points per decade in the Western world (Marks, 2010). In other words, an average person sitting the 1916 Binet-Simon IQ test in the year 2016 would register a true score of 130.

So, given the apparent growth in intelligence, average members from the Binet– Simon IQ test in 1916 would be 'borderline impaired' compared to today's average cohort (see the scales and labels shown in Table 1.1). Meanwhile, Napoleon Bonaparte and Benjamin Franklin must have had IQs that would barely challenge a modern Border collie. It is not difficult to see that there may be a slight flaw in this logic!

Better nutrition is a popular explanation for rising intelligence scores. However, given that what constitutes good nutrition is debatable – and that today's diet is probably too fat- and sugar-rich compared to that of earlier generations – this explanation does not fit the evidence well (Marks, 2010). Similarly, longer school careers may account for some but far from all the variation (Flynn, 2009).

Perhaps the answer lies with modern humans' familiarity with everyday quiz type documents. This means that as a whole we are more *au fait* with conundrums in general and with dealing with abstract notions in particular (Mackintosh, 2011). Mackintosh's argument also helps to account for large cultural variations in test performance which are largely ironed out with time and exposure (Dickens & Flynn, 2002).

PSYCHOMETRICS, WAR AND A PEACETIME DIVIDEND

As with invention so with psychometrics; war is a great driver of innovation. The recruitment of men for the Great War was so inept as to have been a potentially significant contributory factor to desertion, shell shock and slaughter (Mazower, 1999). It was the Americans who tested their soldiers for IQ in 1917 using Army Alpha and Beta tests (for literate and illiterate soldiers respectively). Plotting IQ and literacy against suitability may have been coarse, but it was an improvement on what the British, French and – before the war at least – the Germans had done. It is argued that it was the Americans and their new strategies that helped bring the First World War to a speedier conclusion than would otherwise have been the case (Mead 2000).

Putting, as far as possible, the right soldiers into challenging conditions was a maxim employed by the Duke of Wellington, forgotten by the technologically reliant Victorians and then rediscovered during the Second World War, which saw greatly increased psychological intervention (Reid, 1997). The American Army and Navy General Classification Tests (AGCT and NGCT respectively) replaced the Army Alpha and Beta tests used on the Western Front over 20 years earlier. As a consequence aptitude replaced literacy as a deployment device. The new psychometrics used a broad spectrum of applied psychology to drill into individuals' personalities and to test mental acuity. Data from these psychological tools were used to improve instruction and training. Today the US armed forces use the multiple-choice Armed Services Vocational Aptitude Battery (ASVAB), originally introduced in 1968. The ASVAB features written and computerised tasks, which test talent across a range of ten disciplines including mechanical acuity, electronics, mathematics and English (Hogan & Cannon, 2007).

Over the past century, the British armed forces have moved forward at different speeds. The army lagged behind both the senior and junior services during the world wars, when new recruits were subjected to the most basic training. The Navy, however, insisted on finding skilled seamen, while the Royal Air Force (initially the Royal Flying Corps) has always used challenging recruitment regimes which include physical, mental and intelligence checks (Ballantyne & Povah, 2004).

When National Service was abolished in the UK in 1960, military recruitment standards were tightened and made more rigorous. Today the Army Recruiting and Training Directorate

(ARTD) has an estimated annual budget of approximately £700m with which to enlist candidates and train all ranks. The British Army Recruit Battery (BARB) test is a psychometric array to match the American model. It is designed to steer recruits towards their most suitable roles and predict 'trainability'. Candidates who pass BARB are interviewed, after which those who pass the interview face a two-day assessment centre at an Army Development and Selection Centre (ADSC) (Reid, 1997). And what is an assessment centre? For that we need to turn the other side of the trenches and look at the defeated forces of Kaiser Wilhelm II.

Assessment centres date back to the aftermath of World War I when Germany, now a republic, decided to select its serving officers differently. In 1914 the German officer class had been drawn from the ranks of the nobility, which by the early 1920s had become a more restricted and less trusted elite (Ballantyne & Povah, 2004). The army was only allowed to enlist men for twelve years of service and was restricted in size. It therefore needed to pick the most able, promising candidates. Step forward German psychologist Dr Wolfgang Simoneit, who developed novel methods of identifying the best candidates.

Under Simoneit's aegis the army looked for four major attributes: leadership, adaptability across different situations, mental perspicacity and the ability to work as part of a team (Morgan, 1955). Simoneit developed a range of techniques, the application of which allowed him to judge how well candidates displayed selected competencies.

Everything an applicant said or did was – in theory at last – linked to a competency area. So aspirant officers were exposed to stresses they might face in the field, such as imagined provision shortages, poor communications, even the deaths of key command personnel supposedly killed in action. Would candidates be able to extract the facts from a reticent colleague during a role play? Could they present their ideas cogently to a forceful senior officer, and how did they interact with their peers?

Simoneit had a team of fellow assessors who would compare their opinions of the knowledge, skills, abilities and attitudes displayed. How did men's personalities alter under stress? Did they show more, or less, intelligence when the pressure was on? Were these people trustworthy, capable and inspirational (Morgan, 1955)?

MOVING FORWARD, MIND THE SNAKE OIL

From its ancient origins, psychometrics has been fired in the ovens of war and now proliferates in business. At the top end of the recruitment market, AT&T was one of the first commercial organisations to take up the idea of military-style recruitment in the 1950s. Today the big four accountancy firms make extensive use of psychological analysis within their businesses: 80 per cent of managerial recruitments across the Fortune 500 and FTSE 100 companies depend on psychometric tools, while 68 per cent of *all* employers in Western Europe and the USA now use some form of psychometric assessment as part of their recruitment and development processes (Sponton & Wright, 2009). This growth seems likely to continue. Given increasingly esoteric job roles, the need to uncover the latent talents will call for ever more forensic examination of characteristics, preferences and ability (Cook & Cripps, 2005).

But beware.

Personality and ability assessment depend upon instruments being both reliable and valid. Your wristwatch may be reliable but it is not a valid tool with which to measure your personality. Meanwhile a questionnaire drawn up on the back of an envelope after a discussion with colleagues is likely to be neither reliable nor valid. Validity means that an instrument is consistent and actually measures that which it is supposed to measure.

Many excellent organisations can provide such tools. Companies such as SHL (formerly Saville and Holdsworth Limited), Psytech and ECCOS offer tools of significant worth. Meanwhile bodies like the American and British Psychological Societies maintain the standards required to apply and interpret psychometric data to a high level.

However, the market remains unregulated. Dear reader, you yourself could jot down 40 questions about work, based on your own experience, and publish your list tomorrow, calling it, for the sake of argument, the *Crawchester Psychometric Inventory (or CraPI)*. You could develop your own scoring index, hire a glossy publisher and produce a good-looking, saleable but entirely spurious instrument that lacks market research, science and statistical support. It would be unreliable and invalid. And it would be one of many such tools on the market.

An invalid psychometric instrument is business poison, whereas a good psychometric is a litmus test. Valid psychometric tools used together can significantly improve recruitment, assessment and development. For this we owe a good many thanks to those Ancient Greeks.

REFERENCES

Avanzini, G. (1999). *Alfred Binet*. Paris: Presses Universitaires de France.

Bain, S. K. & Allin, J. D. (2005). Book review: Stanford–Binet intelligence scales, fifth edition. *Journal of Psychoeducational Assessment*, *23*, 87–95.

Ballantyne, I. & Povah, N. (2004). *Assessment and development centres* (2nd edn). Aldershot: Gower.

Buss, D. M. & Schmitt, D. P. (2011). Evolutionary psychology and feminism. *Sex Roles*, 64(9–10), 768–87.

Cardinet, J. (1975). International Test Commission: Application of the Liège recommendations for the period 1971–4. *International Review of Applied Psychology*, 2(1), 11–16.

Carlyle, T. (1841). *Heroes, hero-worship, and the heroic in history*. London: James Fraser.

Carpenter, S. K. (2005). Some neglected contributions of Wilhelm Wundt to the psychology of memory. *Psychological Reports*, 97, 63–73.

Childs, G. (2009). *Understand your temperament*. New York: Rudolf Steiner.

Cook, M. & Cripps, B. D. (2005). *Psychological assessment in the workplace: A manager's guide*. Chichester: John Wiley.

Darwin, C. (1859). *The origin of species*. London: John Murray.

Dickens, W. T. & Flynn, J. R. (2002). The IQ paradox: Still resolved. *Psychological Review*, 109, 764–71.

Fancher, R. E. & Rutherford, A. (2012). *Pioneers of psychology*. New York: W. W. Norton & Company.

Flynn, J. R. (2009). *What is intelligence? Beyond the Flynn effect*. Cambridge: Cambridge University Press.

Gacano, C. B. & Reid, M. J. (1994). *The Rorschach assessment of aggressive and psychopathic personalities*. Hillsdale, NJ: Lawrence Erlbaum.
Galton, F. (1869). *Hereditary genius*. London: Macmillan.
Haney, B., Sirbasku, J. & McCann, D. (2011). *Leadership charisma*. Waco, TX: S & H Publishing.
Hardman, D. (2009). *Judgment and decision making: Psychological perspectives*. Hoboken, NJ: Wiley-Blackwell.
Haslam, S. A. (2004). *Psychology in organizations: The social identity approach* (2nd edn). Thousand Oaks, CA: Sage Publications.
Hogan, T. P. & Cannon, B. (2007). *Psychological testing: A practical introduction*. Hoboken, NJ: John Wiley & Sons.
Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Mahwah, NJ: Erlbaum.
Mackintosh, N. J. (2011). *IQ and human intelligence*. Oxford: Oxford University Press.
Marks, D. F. (2010). IQ variations across time, race, and nationality: An artefact of differences in literacy skills. *Psychological Reports*, 106, 643–64.
Mazower, M. (1999). *Dark continent: Europe's twentieth century*. London: Penguin.
Mead, G. (2000). The doughboys: America and the First World War. London: Penguin.
Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, 88, 355–83.
Morgan, W. J. (1955). *Spies and saboteurs*. London: Victor Gollancz.
Nicolas, S., Andrieu, B., Croizet, J.-C., Sanitioso, R. B. & Burman, J. T. (2013). Sick? Or slow? On the origins of intelligence as a psychological object. *Intelligence*, 41(5), 699–711.
Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, 3, 1–18.
Personalitypage.com (2015). *Portrait of an ENTJ*. https://www.personalitypage.com/ENTJ.html. Accessed 13 August 2015.
Pittenger, D. J. (1993). Measuring the MBTI … and coming up short. *Journal of Career Planning and Employment*, 54, 48–52.
Reid, B. H. (1997). *Military power: Land warfare in theory and practice*. London: Frank Cass.
Singh, S. & Ernst, E. (2008). *Trick or treatment: Alternative treatment on trial*. New York: W.W. Norton.
Sipps, G. J., Alexander, R. A. & Friedt, L. (1985). Item analysis of the Myers–Briggs Type Indicator. *Educational and Psychological Measurement*, 45, 789–96.
Sponton, J. & Wright, S. (2009). *Management assessment centres pocketbook*. Alresford: Management Pocketbooks.
Szondi, L., Moser, U. & Webb, M. W. (1959). *The Szondi Test in diagnosis, prognosis, and treatment*. Philadelphia, PA: J. B. Lippincott.
Uebersax, J. S. (1999). Probit latent class analysis: Conditional independence and conditional dependence models. *Applied Psychological Measurementt*, 23(4), 283–97.
Wolman, B. B. (1968). *Historical roots of contemporary psychology*. New York: Harper & Row.
Wood, J. M., Nezworski, M. T., Lilienfeld, S. O. & Garb, H. N. (2003). The Rorschach Inkblot Test, fortune tellers, and cold reading. *Skeptical Inquirer*, 27(4), 23–8.