# Chapter 1. The role of the measurement theory in modern science

## 1.1. Historical background of testing

Tests and measurement have very long history. The early origins of psychological testing can be back traced in ancient times. Times, when the Chinese emperors had their official examined to determine their mental tests. China was the first country to use testing for the selection of talents [Jin, 2001; Qui, 2003]. Earlier than 500 BCE, Confucius had argued that people were different from each other. In his words, their nature might be similar, but behaviors are far apart, and he differentiated between the superior and intelligent and the inferior and dim. Mencius (372–289 BCE) believed that these differences were measurable. He advised: assess, to tell light from heavy; evaluate, to know long from short. Xunzi (310–238 BCE) built upon this theory and advocated the idea that we should "measure a candidate's ability to determine his position".

**Fig. 1.1.** Imperial examinations in China.
Source: https://pl.pinterest.com/pin/229402174753484934/

In the Xia Dynasty (c. 2070–1600 BCE), the tradition of selecting officers by competition placed heavy emphasis on physical strength and skills, but by the time of the Zhou Dynasty (1046–256 BCE) the content of the tests had changed. The emperor assessed candidates not only based on their shooting skills but also in terms of their courteous conduct and good manners. From then on, the criteria used for the selection of talent grew to include the Six Skills, including arithmetic, writing, music, archery, horsemanship, and skills in the performance of rituals and ceremonies; the Six Conducts such as: filial piety, friendship, harmony, love, responsibility, and compassion; and the Six Virtues: insight, kindness, judgment, courage, loyalty, and concord. During the Warring States period (475–271 BCE), oral exams became more prominent.

In the Qin Dynasty, from 221 BCE, the main test syllabus primarily consisted of the ability to recite historical and legal texts, calligraphy, and the ability to write official letters and reports. The Sui (581–618 CE) and Tang Dynasties (618–907 CE) saw the introduction of the imperial examinations, a nationwide

testing system that became the main method of selecting imperial officials. Formal procedures required, then as they do now, that candidates' names should be concealed, independent assessments by two or more assessors should be made, and conditions of examination should be finally standardised.

The general framework of assessment set down then, including a syllabus of material that should be learned and rules governing an efficient and fair "examination" of candidates' knowledge, has not changed for 3,000 years. While similar but less sophisticated frameworks may have existed in other ancient civilisations, it was models based on the Chinese system that were to become the template for the modern examination system. The British East India Company, active in Shanghai, introduced the Chinese system to its occupied territories in Bengal in the early 19th century. Once the company was abolished in 1858, the system was adopted by the British for the Indian Civil Service. It subsequently became the template for civil service examinations in England, France, the USA, and much of the rest of the world.

The Chinese examination system also influenced neighboring countries, such as Japan, Korea and Vietnam. The Chinese examination system was introduced to Europe in the reports of European missionaries and diplomats, and encouraged France, Germany, and the British East India Company to use a similar method to select prospective employees. Following the initial success in that company, the British government adopted a similar testing system for screening civil servants in 1855. Modeled after these previous adaptations, the United States established its own testing program for certain government jobs after 1883. Chinese tests served as model for developing exams and tests later introduced in the USA and in Europe and measurements procedures for examining tests results. Now, we present the history of measurement and tests development that took place in the 18th century and continues constantly till nowadays.

Nevil Maskelyne (1732–1811) was the fifth British Astronomer Royal. He held the office from 1765 to 1811. He was the first person to contribute to the measurement with his achievement by the scientifically measure the mass of the planet Earth. Nevil Maskelyne's 1774 experiment on the Scottish mountain Schiehallion set out to derive the mean density of the Earth, from astronomical observations of the deflection of the vertical and calculation of the mountain's relative gravitational attraction. Using Maskelyne's results and lithological survey results, John Playfair estimated mean Earth specific gravity to be 4.56–4.87, while Charles Hutton argued in 1821 that the Earth was "very near five times the density of water; but not higher".


AN
ACCOUNT
OF
OBSERVATIONS
MADE ON THE
MOUNTAIN SCHEHALLIEN
FOR FINDING ITS
ATTRACTION.
LONDON:

**Fig. 1.2.** An Account of Observations made on the Mountain Schehallien for finding its attraction. Read at the Royal Society, July 6, 1775.
Source: https://www.nigelphillips.com/product/maskelyne-nevil/

Hutton challenged future workers to identify any areas in which his analysis could be improved. The geometry of the 1774 experiment has therefore been

recomputed within a digital elevation model extending 120 km from the mountain. Three contributions to the deflection of the vertical have been included: topography, and local and regional subsurface density variations. Local subsurface densities have been modelled using geological maps, cross-sections and laboratory measurements. Regional subsurface effects have been included from analysis of the Bouguer gravity anomaly. The outcome of the new modelling is to credit Maskelyne for his accurate astronomical observations, as together with the new density structure model, they yield a mean Earth density of $5480 \pm 250$ kg/m$^3$, in agreement with the modern value of 5515 kg/m$^3$.

**Fig. 1.3.** Statue of Adolphe Quetelet in the gardens of the Palais des Académies in Brussels.
Source: https://www.brusselsremembers.com/memorials/adolphe-quetelet-at-palais-des-academies

Adolph Quetelet (1796–1874) was a Belgian statistician, and famously envisioned *l'homme moyen* – an image of the average man developed through the measurement of human features with the deviation plotted around the mean. He started with human physical features, like the chests of Scottish Highland regiment soldiers, and moved on to moral and intellectual qualities including suicide, crime, madness, and even poetic ability. For Quetelet, the average body presented an ideal beauty; the normal, conceived of average, emerged as an ideal type to be desired. It was Quetelet who formulated the BMI, initially through the measurement of typical weights among French and Scottish conscripts. Instead of labelling the peak of the bell-curve as merely normal, he labelled it "ideal", with those deviating either "overweight" or "underweight" instead of heavier than average or lighter than average.

Thus, while informed by statistics, Quetelet was still working within the medical context of the normal; that is, he envisioned the normal (i.e., typical) as the ideal or something desirable.

**Fig. 1.4.** First Wundt Laboratory in Leipzig in Germany.
Source: https://bibliolore.org/2017/06/12/rhythm-and-experimental-psychology/

The next stage of the development of measurement methods is very strongly connected to experimental psychology. With advances in science, much had been learned about the physical world, including about the physical stimulation of the sense receptors, which convert that stimulation into nerve impulses, and about the brain structures where those impulses terminate.

There was never much doubt about the existence of consciousness; the problem was in determining what we were conscious of and what caused that consciousness.

Ernst Heinrich Weber (1795–1878), a contemporary of Johannes Müller, born in Wittenberg the son of a theology profesor, was the third of 13 children. Weber obtained his doctorate from the University of Leipzig in 1815 and taught there until his retirement in 1871. Weber was a physiologist who was interested in the senses of touch and kinesthesis (muscle sense). Most of the research on sense perception before Weber had been confined to vision and audition. Weber's research consisted largely in exploring skin and muscle sensations. Weber was among the first to demonstrate that the sense of touch is not one but several senses. For example, what is ordinarily called the sense of touch includes the senses of pressure, temperature, and pain. Weber also provided convincing evidence that there is a muscle sense. It was in regard to the muscle sense that Weber performed his work on just noticeable differences, which we consider shortly.

Gustav Theodor Fechner (1801–1887) was a brilliant, complex, and unusual individual. At the age of 16, Fechner began his studies in medicine at the University of Leipzig (where Weber was studying) and obtained his medical degree in 1822 at the age of 21. Upon receiving his medical degree, Fechner's interest shifted from biological science to physics and mathematics. At this time, he made a meager living by translating into German certain French handbooks of physics and chemistry, by tutoring, and by lecturing occasionally.

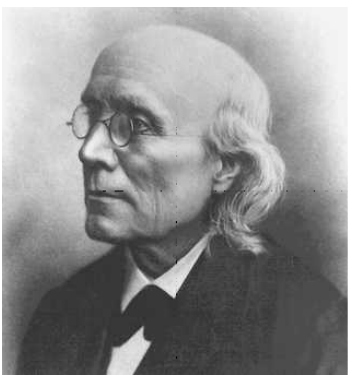


**Fig. 1.5.** Gustav Fechner.
Source: Archive of the History of American Psychology, The Center of the History of Psychology the University of Akron

Fechner was interested in the properties of electric currents and in 1831 published a significant article on the topic, which established his reputation as a physicist. In 1834, when he was 33 years old, Fechner was appointed profesor of physics at Leipzig. Soon his interests began to turn to the problems of sensation, and by 1840 he had published articles on color vision and afterimages. He accepted Spinoza's double-aspect view of mind and body, and therefore believed that consciousness is as prevalent in the universe as is matter. Because he believed that consciousness cannot be separated from physical things, his position represents panpsychism; that is, all things that are physical are also conscious. In his lifetime, Fechner wrote 183 articles and 81 books and edited many others [Bringmann, Bringmann, & Balance, 1992]. He was eulogised by his friend and colleague Wilhelm Wundt and their works were fundamental to experimental psychology. From Fechner's philosophical interest in the relationship between the mind and the body sprang his interest in psychophysics. He wanted desperately to solve the mind–body problem in a way that would satisfy the materialistic scientists of his day. Fechner's mystical philosophy taught him that the physical and mental were simply two aspects of the same fundamental reality. Thus, as we have seen, he accepted the double aspectism that Spinoza had postulated. But to say that there is a demonstra-

ble relationship between the mind and the body is one thing; proving it is another matter. According to Fechner, the solution to the problem occurred to him the morning of October 22, 1850, as he was lying in bed [Adler, 1996]. His insight was that a systematic relationship between bodily and mental experience could be demonstrated if a person were asked to report changes in sensations as a physical stimulus was systematically varied. Fechner speculated that for mental sensations to change arithmetically, the physical stimulus would have to change geometrically. In testing these ideas, Fechner created the fundamentals of psychophysics.

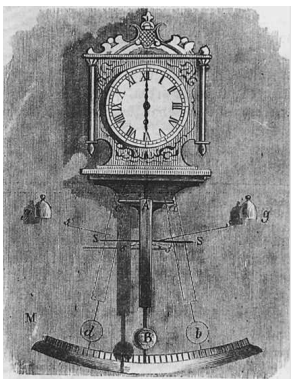


**Fig. 1.6.** Wundt's "thought meter".
Source: Wundt [1862b, p. 264].

**Fig. 1.7.** Wilhelm Maximilian Wundt.
Source: Archives of the History of American Psychology, The Center for the History of Psychology, The University of Akron.

Wilhelm Wundt (1832–1920), one of the most infuencial psychologist, philosopher, physician that had an impact on the development of psychometrics, known as a founder of modern psychology, opened first Institute for Experimental Psychology in 1879 at the University of Leipzig in Germany. Wundt is often treated as the father of psychology.

Wundt, called the father of experimental psychology, held that attention was an inner activity that caused ideas to be present to differing degrees in consciousness. He distinguished between perception, which was the entry into the field of attention, and apperception, which was responsible for entry into the inner focus. He assumed that the focus of attention could narrow or widen. This view that has also enjoyed popularity in recent years.

Wundt disagreed with individuals like Galileo, Comte, and Kant who claimed that psychology could never be a science. He also disagreed with Herbart, who said that psychology could be a mathematical science but not an experimental one. He strongly believed that psychology had, in fact, become an experimental science. As we have seen, however, in his comprehensive view of psychology, experimentation played only a limited role. He proved that experimentation could be used to study the basic processes of the mind but could not be used to study the higher mental events.

The clock was arranged so that the pendulum ($B$) swung along a calibrated scale ($M$). The apparatus was arranged so that a bell ($g$) was struck by the metal pole ($s$) at the extremes of the pendulum's swing $(d, b)$, Wundt discovered that if he looked at the scale as the bell sounded, it was never in position $d$ or $b$ but some distance away from either. Thus, determining the exact position of the pendulum as the bell sounded was impossible. Readings were always about 1/10 of a second off. Wundt concluded that one could either attend to the position of the pendulum or to the bell, but not both at the same time.

For the latter, only various forms of naturalistic observation could be used. We will see how Wundt proposed to study the higher mental thought processes when we discuss his Völkerpsychologie. Still, the role of experimental psychology was vital to Wundt. Learning about the simpler conscious processes was fundamental for understanding those that are more complex: "Let us remember the rule, valid for psychology as well as for any other science, that we cannot understand the complex phenomena, before we have become familiar with the simple ones which presuppose the former" [Wundt, 1912/1973, p. 151]. According to Wundt, psychology's goal was to understand both simple and complex conscious phenomena. For the former, experimentation could be used; for the latter, it could not. Wundt believed that all sciences are based on experience and that scientific psychology is no exception.

Wundt was the founder of both experimental psychology as a separate discipline and the school of voluntarism. One of Wundt's goals was to discover the elements of thought using experimental introspection. A second goal was to discover how these elements combine to form complex mental experiences. Wundt found that there are two types of basic mental experiences: sensations and feelings. Wundt distinguished among sensations, which are basic mental elements; perceptions, which are mental experiences given meaning by past experience; and apperceptions, which are mental experiences that are the focus of attention. Because humans can focus their attention on whatever they wish, Wundt's theory was referred to as voluntarism. Wundt believed that reaction time could suplement introspection as a means of studying the mind. Following techniques developed by Donders, Wundt presented tasks of increasing complexity to his subjects and noted that more complex tasks resulted in longer reaction times. Wundt believed that the time required to perform a complex mental operation could be determined by subtracting the times it took to perform the simpler operations of which the complex act consists. Wundt eventually gave up his reaction-time studies because he found reaction time to be an unreliable measure.

Now we move to another scientist who played an important role in psychometrics and measurement theory. One of them is Charles Darwin, a giant figure of that age, whose theory of evolution had considerable implications for how differences both between and within species would be understood, was among those who held this Eurocentric approach, something that perturbed the development of evolutionary science in a way that would increasingly be recognised as racist. In *The Descent of Man*, first published in 1871, Darwin argued that the intellectual and moral faculties had been gradually perfected through natural selection, stating as evidence that "at the present day, civilised nations are everywhere supplanting barbarous nations".

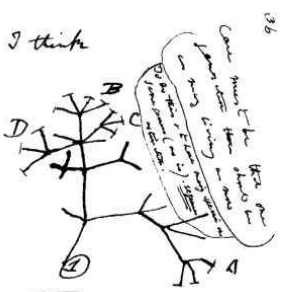



**Fig. 1.8.** Charles Darwin's 1837 sketch.
Source: https://commons.wikimedia.org/wiki/File:Darwin_tree.png

Darwin's view that natural selection in humans was an ongoing process, with "the savage" and "the lower races" being evolutionary inferior to "the civilised nations", had considerable influence.

But it was not a view that was shared by every scientist of that time. Others chose to differ, including Alfred Wallace, Darwin's copresenter of papers to the seminal 1858 meeting of the Linnean Society of London that introduced the idea of natural selection through survival of the fittest [Wallace, 1979]. Wallace took issue with his former colleague, believing Darwin's arguments for differences between the races to be fundamentally flawed.

His observations in Southeast Asia, South America, and elsewhere convinced him that so-called "primitive" people all exhibited a high moral sense. He also drew attention to the ability of children from these groups to learn advanced mathematics, having taught five-year-olds in Borneo how to solve simultaneous equations, and pointed out that no evolutionary pressure could have ever been exerted on their ancestors in this direction by the natural environment. In Wallace's view, the evolutionary factors that had led to the development of intelligence and morality in humanity had happened in the distant evolutionary past and were shared by all humans.

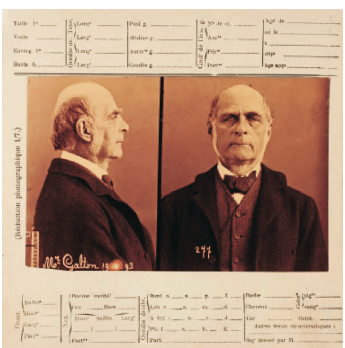


**Fig. 1.9.** Antropometric record card of Francis Galton.
Source: https://www.researchgate.net/figure/Anthropometric-record-card-of-Francis-Galton-including-profile-and-full-face-photos_fig63_325272698

Sir Francis Galton (1822–1911) was an English mathematician, psychometrist, inventor, meteorologist, geneticist, and eugenicist. Galton made numerous contributions to biometrics and statistics, including introducing the concepts of correlation and of regression towards the mean. He coined the term eugenics and is considered to be the father of the eugenics movement. The phrase "nature versus nurture" is also credited to Galton. Galton was born in Birmingham on February 16, 1822. He came from a prominent, intellectually-driven family; for example, Charles Darwin, his cousin, first described the theory of evolution. Galton originally studied medicine, but he was convinced by Darwin to study mathematics in Cambridge to enhance his medical education [Gillham, 2001].

Galton tried to complete an advanced program at Cambridge, but became overwhelmed and finished with a regular degree in mathematics. He did not return to medicine, but instead choosing to travel and explore. The evolution of the human intellect was also of particular interest to Francis Galton, who published *Hereditary Genius: An Inquiry into Its Laws and Consequences* [1896]. He had first carried out a study of the genealogy of famous families in 1865, based on a compendium by Sir Thomas Phillipps entitled *The Million of Facts* [Galton, 1865]. Galton argued that genius, genetic in origin, was to be found in these families. But when he spoke of genius, he was considering much more than mere

intellect. He believed that these people were superior in many other respects, be it the ability to appreciate music or art, performance in sport, or even simply physical appearance. Galton's desire to measure individual differences among humans inspired him to create what he called an "anthropometric laboratory" at London's International Health Exhibition in 1884. Here, in about one year, Galton measured 9337 humans in just about every way he could imagine. For example, he measured head size, arm span, standing height, sitting height, length of the middle finger, weight, strength of hand squeeze (measured by a dynamometer), breathing capacity, visual acuity, auditory acuity, reaction time to visual and auditory stimuli, the highest detectable auditory tone, speed of blow (the time it takes for a person to punch a pad). Some of these measures were included because Galton believed sensory acuity to be related to intelligence, and for that reason, Galton's anthropometric laboratory can be viewed as an effort to measure intelligence, or even the beginning of the mental testing movement in psychology. In 1888 Galton set up a similar laboratory in the science galleries of the South Kensington Museum, and it operated for several years [Hergenhahn, Henley 2014].

When Darwin published his work entitled *The Origin of Species*, Galton changed his work of direction. Due to Darwin's thoughts, Galton was fascinated by the idea of evolutionary progress. He wrote *The Origin of Species* which gave him a purpose in life: using the concepts described by his cousin to improve the human race.

Galton was not the first person to address controlled breeding in humans in order to create a better species – the idea dates back at least two thousand years ago, in ancient Greece – but he did help to generate new interest on the subject [Galton, 1998]. Galton was convinced that social and mental traits, like talent and intelligence, were inherited [Galton, 1865; Galton, 1869]. Galton published his thoughts in the popular *Macmillan's Magazine* [Galton, 1865] and conducted extensive research to try to establish that personality, work ethic, and other traits were hereditary, and could be traced through family lineages. In 1869, he released a compilation of the data he had collected and published it under the title *Hereditary Genius* [Galton, 1869]. In this work, he showed that success seemed to run in families, and that the more closely related a person was to a high-achiever, the more likely they were to become one themselves [Galton, 1869]. He argued that this proved that intelligence, accomplishment, and various other features were inherited. He thought that a person's environment had very little to do with the development of such characteristics. He used Brass instruments


**Fig. 1.10.** Francis Galton's Anthropometric Laboratory at the International Health Exhibition, London, 1884.
Source: Psychology Pictures/Archives of Dutch Psychology.

to measure sensory threshold and reaction time to visual and auditory stimulus. Galton tested around 17,000 people and demonstrated that objective tests could provide meaningful scores.

Galton as first introduced the word eugenics in 1883 and described it as a brief word to express the science of improving stock, which is by no means confined to questions of judicious mating, but which, especially in the case of man, takes cognisance of all influences that tend in however remote a degree to give to the more suitable races or strains of blood a better chance of prevailing speedily over the less suitable than they otherwise would have had [Galton, 1883, p. 24–25]. The term eugenics literally means well-born, but it was eventually given various definitions relating to actions and ideas aimed at improving the inheritable qualities of the human race [Galton, 1998]. By the early 20$^{th}$ century, the word eugenics was being used in public and academic spheres. Galton intended for eugenics to become a sort of religion, and he believed that eugenics could lead to a perfect, happy and successful human race [Galton, 1869; Kevles, 1985]. Originally, he imagined that species improvement could be achieved through the elite marrying and having large numbers of children. However, in Galton's later work, he also focused on the least desirable – explaining who these undesirables were, creating classification systems for them, and suggesting how they should be treated [Mackenzie, 1976]. Galton did not seem to promote cruelty or extreme measures in the name of eugenics, but he did propose that the so-called unfit be segregated.

Overall, Galton was a significant early contributor to a movement that would grow and develop long after his death. He encouraged an in-depth exploration of the inherited differences between classes of people and promoted the belief that certain groups were fundamentally and genetically superior to others. While he also discovered important concepts in statistics and psychology, many of these were simply mechanisms to further his eugenic ideas [Gillham, 2001; Kevles, 1985]. He believed that the key to a Utopian society was a eugenic religion, and Galton dedicated his life to eugenics.

Galton originated and named the eugenics movement, in service of which he created the basic idea of the intelligence test. Galton believed that intelligence was a matter of neurological efficiency. Therefore, he theorised that it could be tested by measuring reaction time and sensory acuity. In the middle of 1880s he established an anthropometric laboratory at London's South Kensington Museum, cleverly enticing visitors to spend three pence apiece to enjoy the novelty of undergoing a variety of psychophysical tests. Galton‘s interest in the nature of genius led him to examine developments in Europe in the newly emerging field of osychophisics. He cooperated with James McKeen Cattell who worked in Wilhelm Wundt‘s laboratory in Leipzig.

At about the same time, an American doctoral student, James McKeen Cattell (1860–1944), Wundt's assistant in his laboratory in Leipzig was conducting a series of reaction time experiments in Germany. When he became aware

of Galton's anthropometric laboratory, he began to correspond with him, and designed a series of 50 psychophysical tests based on Galton's earlier work. He established his own anthropometric laboratory at Cambridge University, and became a successful international advocate of the psychophysical approach to mental testing.

With this first laboratory, the field of psychometrics could differentiate from psychophysics and the major differences can be grouped as the following:

1) while psychophysics aimed to discover general sensory-perception laws (i.e. psychophysical functions), psychometrics was (is) concerned with studying differences between individuals;
2) the goal of psychophysics is to explore the fundamental relations of dependency between a physical stimulus and its psychological response, but the goal of psychometrics is to measure what we call latent variables, such as intelligence, attitudes, beliefs and personality;
3) the methods in psychophysics are based on experimental design where the same subject is observed over repeated conditions in a controlled experiment, but the majority of studies in psychometrics are observational when the measurement occurs without trying to affect the participants [Jones, Thissen, 2007].

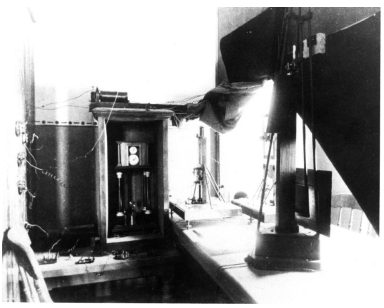


**Fig. 1.11.** Cattell's laboratory showing the Hipp Chronoscope and gravity chronometer.
Source: https://www.sas.upenn.edu/psych/history/cattelltext.htm

In fact, neither Galton's antropometric tests nor Cattell's early psychometric tests were successful. Venn [1889] and Wissler [1901] proved that there are no differences among students in one class using their approach to correlation.

The work of these two men was of pivotal importance to the field of experimental psychology. Psychophysical testing had great popular appeal, and was being enthusiastically embraced by researchers from many different countries. In the late 1880s Cattell, recently arrived in Cambridge from Wundt's psychophysics laboratory in Germany, introduced Galton to many of Wundt's psychological testing instruments. Hence – mental testing called psychometrics as a science was born.

A few years before Galton's death, Charles Spearman [1904] published his landmark paper that introduced the factor analysis model. As a result of his empirical studies, building in part on the early experiments of Galton, he concluded that: "all branches of intellectual activity have in common one fundamental function (or group of functions), whereas the remaining or specific elements of the activity seem in every case to be wholly different from that in all the others (p. 284)". One of these fundamental functions introduced by Spearman was what he termed general intelligence or $g$, which he regarded as a universal factor underlying mental attributes. It was not long before this claim was contested, most

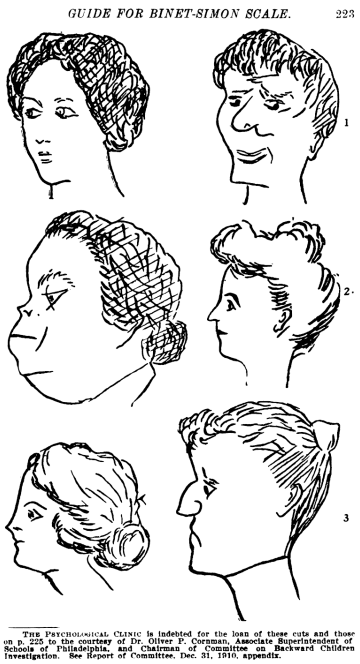
notably by Burt [1909] and then by many others, with Spearman defending his position over the next quarter century.

Karl Pearson was very important scientist and pioneer in the founding of the school of biometrics, which was a competing theory to describe evolution and population inheritance at the turn of the 20th century. His series of eighteen papers entitled *Mathematical Contributions to the Theory of Evolution* established him as the founder of the biometrical school for inheritance. In fact, Pearson devoted much time during 1893 to 1904 to developing statistical techniques for biometry. These techniques, which are widely used today for statistical analysis, include the chi-square test, standard deviation, correlation coefficient or regression model. Pearson's Law of Ancestral Heredity stated that germ plasm consisted of heritable elements inherited from the parents as well as from more distant ancestors, the proportion of which varied for different traits. Karl Pearson was a follower of Galton's achievements, and although the two differed in some respects, Pearson used a substantial amount of Francis Galton's statistical concepts in his formulation of the biometrical school for inheritance, such as the law of regression. While Galton proposed a discontinuous theory of evolution, in which species would have to change via large jumps rather than small changes that built up over time, Pearson pointed out flaws in Galton's argument and actually used Galton's ideas to further a continuous theory of evolution, whereas the Mendelians favored a discontinuous theory of evolution.

Another milestone in the development of the measurement theory was due to Binet's testing of the intelligence. The main impetus to provide an intelligence test for educational selection took place in France in 1904, when the minister of public instruction in Paris appointed a committee to find a method that could identify children with learning difficulties. It was urged that "children who failed to respond to normal schooling be examined before dismissal and, if considered educable, be assigned to special classes" [Binet and Simon, 1916]. Drawing from item types already developed, the psychologist Alfred Binet and his colleague Théodore Simon put together a standard set of 30 scales that were quick and easy to administer. These were found to be very successful at differentiating between children who were seen as bright and children who were seen as dull (by teachers), and between children in institutions for special educational needs and children in mainstream schools.

Furthermore, the scores of each child's scales could be compared with those of other children of the same or similar age, thus freeing the assessment from teacher bias. The results of Binet's testing program not only provided guidance on the education of children at an individual level but also influenced educational policy. The first version of the Binet–Simon Scale was published in 1905, and an updated version followed in 1908 when the concept of "mental age" was introduced–this being the age for which a child's score was most typical, regardless of their chronological age. In 1911, further amendments were made to improve the ability of the test to differentiate between education and educability. Scales of reading,

writing, and knowledge that had been incidentally acquired were eliminated. The English-language derivative of the Binet–Simon test, the Stanford–Binet, is still in widespread use today as one of the primary assessment methods for the identification of learning difficulties in children.


GUIDE FOR BINET-SIMON SCALE. 223

THE PSYCHOLOGICAL CLINIC is indebted for the loan of these cuts and those on p. 225 to the courtesy of Dr. Oliver P. Corman, Associate Superintendent of Schools of Philadelphia, and Chairman of Committee on Backward Children Investigation. See Report of Committee, Dec. 31, 1910, appendix.

**Fig. 1.12.** Reproduction of an item from the 1908 Binet-Simon intelligence scale. Source: https://en.wikipedia.org/wiki/Stanford%E2%80%93Binet_Intelligence_Scales

Binet's tests emphasised what he called the higher mental processes that he believed underpinned the capacity to learn: the execution of simple commands, coordination, recognition, verbal knowledge, definitions, picture recognition, suggestibility, and the completion of sentences. In their book *The Development of Intelligence in Children*, first published in 1916, Binet and Simon, using the language of the time, stated their belief that good judgment was the key to intelligence: It seems to us that in intelligence there is a fundamental faculty, the alteration or the lack of which, is of the utmost importance for practical life. This faculty is judgment, otherwise called good sense, practical sense, initiative, the faculty of adapting one's self to circumstances. To judge well, to comprehend well, to reason well, these are the essential activities of intelligence. A person may be a moron or an imbecile if he is lacking in judgment; but with good judgment he can never be either. Indeed, the rest of the intellectual faculties seem of little importance in comparison with judgment [Binet and Simon, 1916].

With the active support of graduate students and colleagues of L.L. Thurstone at The University of Chicago, the Psychometric Society was founded in 1935. The Society sponsored the journal *Psychometrika*, of which Volume 1, Number 1 appeared in March, 1936. At nearly the same time, the first edition of Guilford's (1936) *Psychometric Methods* was published. From one perspective, these events may serve to temporally locate the beginning of a formal sub-discipline of psychometrics. The founding of the Psychometric Society led to the publication in *Science* of Thurstone's presidential address to the meeting of the Society in September of 1936 that presented a strong plea to recognise a mathematical underpinning for psychological research [Thurstone, 1937].

In work published in the middle of the 1920s, L.L. Thurstone provided the conceptual cornerstone and foundation upon which much of IRT has been built. In *A Method of Scaling Psychological and Educational Tests*, Thurstone [1925] proposed an analytic procedure to be applied to test items that can be graded right or wrong, and for which separate norms are to be constructed for successive age-

or grade-groups. Thurstone used the terms "mental age", "achievement", and "intelligence" interchangeably. Thurstone's inspiration was seen in Cyril Burt's [1922] data collected using his translation into English of the Binet intelligence test questions. Burt's [1922] book contained a table of the percents of British children who responded correctly to each Binet item. Thurstone [1925] graphed the percentage correct as a function of age for eleven of the questions in Burt's [1922] table (upper part of Fig. 13).

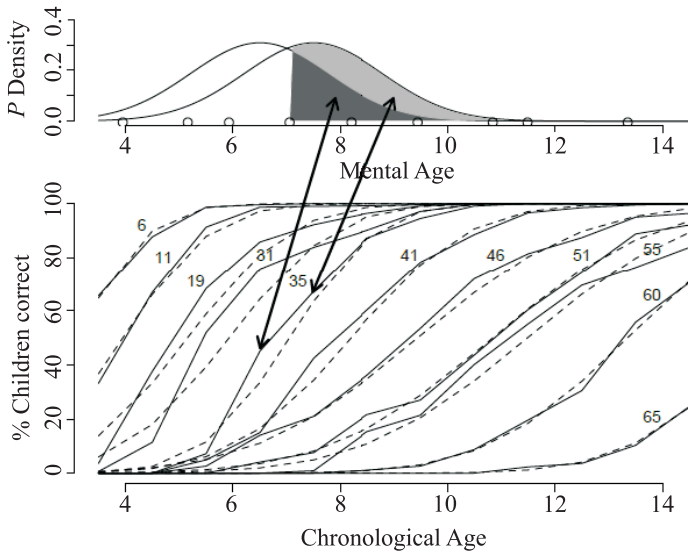



**Fig. 1.13.** Upper plot: Two normal curves representing the distribution of mental age for 6- and 7-year old children (modeled after Thurstone's [1925]. Lower plot: The observed percentage correct (solid lines) for eleven of Binet items in Burt's [1922, p. 132–133] data.
Source: Thiessen, Steinberg [2020, p. 25].

Following Thurstone's example, the points for each age are located at the midpoint of each year (4.5, 5.5, 6.5, …) because the data were grouped by age in years in the original tabulation. Thurstone was struck by the resemblance between those empirical curves and the cumulative normal called also the normal ogive. In 1925 it would not have been easy to add fitted probit curves to the graphic, but those have been included as the dashed lines.

Late 40s and 50s of the 20$^{th}$ century were considered the golden period of psychological assessment, particularly in the USA. The statistical methods of factor analysis were widely applied in test construction and validity studies. By the 1950s the major forms of psychological tests were designed mostly for the assessment of behavioral differences. In this period, new tests were designed primarily of the self-report inventory and behavior scale type, which addresses other domains apart from intelligence and personality, such as attitudes, achievement, temperament, or aggression. The period between 1960 and 1990 focused more on the assessment

of cognitive, memory, and related neuropsychological functions. In the 1970s computers were introduced in test administration as a mean to evaluate the reliability of behavior measures. Factor and confirmatory analysis were applied to test construction and in particular to the study of construct validity.

The chronological age scale proposed by Binet allowed for direct measurement. Thurstone's innovation was to put Binet's items on an intelligence scale that cannot be measured directly. He did so because he recognised that intelligence is an example of what is now known as a latent variable; that is, an unmeasured hypothetical construct. The scale of this variable was defined by postulating a normal distribution for each age group and inferring the scale values of the items from the response data. Thurstone also showed how to check this distributional assumption. The assumption of a normal cumulative distribution function as a response function was not new but borrowed from the earlier work on psychophysics by Fechner, who used them to describe how psychological sensations vary with the strength of experimentally manipulated physical stimuli. But Thurstone's idea to separate intelligence from age and define it as a latent variable with a scale defined by such response functions was entirely new. The idea of response functions on a latent variable was picked up again by authors like Ferguson, Lawley, and Mosier in the 1940s (and led to much confusion between the use of the normal ogive as a definition of a population distribution and a response function on a latent variable). But we had to wait until achievements by Lord [1952] and Rasch [1960] until the developments really began. From a statistical perspective, later contributions by Birnbaum [1968] were very important. He proposed replacing the normal ogive by the logistic function, introduced additional item parameters to account for guessing on items (which is typical of most educational measurements), derived maximum-likelihood estimators for the model, and showed how to assemble tests from a bank of calibrated items to meet optimal statistical specifications for their application. Intelligence was not the only concept of interest measured by scientists that time. Psychological testing and factor analysis were also applied to the measurement of attitude, preferences or other psychological constructs.

Guttman published a number of papers during the 1950s that defined and refined important concepts in the field of measurement such ad reliability, scaling and factor analysis. In 1947 Thurstone published a first full-length text on factor analysis that helped to make a statistical tool widely accessible to researchers working in the area of measurement. Therefore, throught the ideas were certainly relevant, their application was largely forced to wait until the 1970s when computing power caught up to the the data analyses. It was noted that the crucial work from 1950s and 1960s led to the development of item response theory (IRT), which serves as the basis for all modern large-scale testing programs. Those ideas are that items are "located" on the same scale as the "ability" variable [Thurstone, 1925], the "ability" variable is latent (or unobserved) [Lazarsfeld, 1950; Lord, 1952], and the unobserved variable accounts for the observed interrelationships among the

item responses [Lazarsfeld, 1950]. These ideas saw some use in theoretical work concerning the structure of psychological tests, by Lord [1952, 1953], Solomon [1956, 1961], Sitgreaves [1961a, 1961b, 1961c], and others. However, there was still no practical way to estimate the parameters (the item locations and discriminations) from observed item response data.

The next two decades of the 20$^{th}$ century offered much works on item response theory models for test items with other response formats than simple dichotomous scores as well as on newer procedures for parameter estimating and model evaluation. Especially, the development of Bayesian procedures for parameter estimation and model validation. Due to the availability and the development of computer software this method became more powerful and popular in the 1980s. The first programs to exploit IRT to score test takers in real time and deliver computerised adaptive tests were launched in the 1990s. Nowadays, item response theory models are no longer the main instruments only in the educational testing industry, but are becoming increasingly popular in psychological testing, survey research, medicine, economy and other scientific fields.

In the 2000s the IRT field was promoted by a new wave of researchers who not only expanded the technical aspects of the framework (estimation, model identification, and goodness of fit), but also advanced its computational aspects. The extensive study of IRT during the past 50 years was manifested in a rise in the number of software packages designed for analysing item response data from surveys or tests. Various IRT commercial software was also created including BILOG, MULTILOG, WINSTEPS, IRTPRO, MPLUS, STATA and HLM. More importantly, a number of IRT packages developed in the open source **R** (`www.r-project.org`) software to estimate various IRT models also appeared and gained recognition. These included the packages ltm for unidimensional IRT [Rizopoulos, 2006], `eRm` for extended Rasch models [Mair & Hatzinger, 2007], `mlirt` for multilevel and Bayesian estimation of some IRT models [Fox, 2007], `gpcm` [Johnson, 2007] for a Bayesian estimation of the generalised partial credit model, `MCMCpack` for Bayesian IRT [Martin, Quinn, Park, 2011], `irtGUI` for IRT analysis with a user-friendly Graphic User Interface, and `mirt` for multidimensional IRT [Chalmers, 2012]. De Boeck, Wilson, Acton [2008] made use of the general statistics package `lme4` and incorporated Rasch models under the generalised linear mixed model framework. Such a wide range of packages dedicated to different datasets allows to apply item response models in almost all scientific fields. We may also assume that IRT methodology will be incorporated and overlap with other frameworks such as structural equation modeling and factor analysis. There also may be a new development in models and estimation methods, as well as computer software that allows model application by non-statisticians and a large group of researchers that are willing to apply IRT modeling in their practical research.