Chapter 1
Early Roots of Psychometrics Before Francis Galton

Willem J. Heiser

Abstract Although one of the flagships of psychometrics, factor analysis, could not have been invented without Francis Galton's (1822–1911) groundbreaking concept of correlation, some other psychometric concepts had been explored already before his time. Christian Thomasius (1655−1728) pioneered personality assessment using numerical rating scales and introduced a first notion of psychometric reliability. It was Christian Wolff (1679−1754) who coined the term "psychometria" and who identified the basic difficulty of finding a suitable unit for measurement of psychological variables. Halfway the nineteenth century, Gustav Fechner (1801–1887) not only founded psychophysics but also introduced before Galton the statistical approach to the analysis of psychological data—which is so typical for psychometrics in general. He also developed some pathbreaking experimental designs for data collection, as well as the notions of a psychological scale and the psychometric function.

1.1 Introduction

When the first laboratory worldwide for both research and teaching of experimental psychology was founded in Leipzig (1879) by Wilhelm Wundt (1832–1920), it immediately attracted many students, not only from Germany and neighboring countries in Europe but also from the United States. The first of them, James McKeen Cattell (1860–1944), while discussing the topic of his dissertation with Wundt, experienced a bit of a culture shock:

The author is grateful to Denny Borsboom, Larry Hubert, Jacqueline Meulman, Stephen Stigler, and Robert Tijssen for their useful comments on an earlier draft of this chapter.

W. J. Heiser (✉)
Leiden University, Leiden, The Netherlands
e-mail: heiser@fsw.leidenuniv.nl

L. A. van der Ark et al. (eds.), *Essays on Contemporary Psychometrics*,
Methodology of Educational Measurement and Assessment,
https://doi.org/10.1007/978-3-031-10370-4_1

> As a large part of the work of the laboratory was then on reaction-time experiments, it was not surprising that such a subject fell to my lot, and it was fortunate, for I had already in America begun experimental work on the time of sensori-motor processes. Wundt, however, was mainly interested in experiment for the aid it gave to introspection, and the subject assigned to me was to react as soon as I saw a light and in a second series to react as soon as I recognized its color, with a view to analyzing the factors of apperception. This I could not do, and in my second interview with Wundt I presented an outline of the work I wanted to undertake, which was the objective measurement of the time of reactions with special reference to individual differences. Wundt said that it was 'ganz Amerikanisch'; that only psychologists could be the subjects in psychological experiments. (Cattell, 1921, p. 156)

Despite Wundt's negative reaction, Cattell was allowed to start his project as he conceived it, with his own apparatus and in his own room, and Wundt prepared him graciously for his doctorate examination. His dissertation work in Leipzig was published as Cattell (1886).

Meanwhile in England, Francis Galton (1822–1911) had founded an anthropometric laboratory, at the occasion of the International Health Exhibition in London (1884–1885), where he measured and recorded "the chief physical characteristics of man," including "keenness of sight, colour sense and hearing" (Galton, 1885). In total, he was able to measure 9337 ordinary persons on 17 variables. Attracted by Galton's interest in empirically studying individual differences between people, Cattell went to London soon after leaving Leipzig and joined Galton in his research projects. That joint effort resulted in the paper *Mental tests and measurement*, of which Cattell wrote the main part, opening with the programmatic statement:

> Psychology cannot attain the certainty and exactness of the physical sciences, unless it rests on a foundation of experiment and measurement. A step in this direction could be made by applying a series of mental tests and measurements to a large number of individuals. The results would be of considerable scientific value in discovering the constancy of mental processes, their interdependence, and their variation under different circumstances. (Cattell & Galton, 1890, p. 373)

Cattell continues to describe a long series of 60 tasks concerning sight, hearing, taste and smell, touch and temperature, sense of effort and movement, mental time, and memory. These were the type of tasks used by Fechner, Wundt, and Helmholz, the pioneers of experimental psychology. However, Cattell broke with their habit of using only a few psychologists as subjects and with the high priority these German pioneers gave to finding general laws. Wundt's low regard of individual differences continued to dominate experimental psychology for a long time. Only quite recently, we have seen several attempts to bring together classical tasks used by experimental psychologists with a serious look at individual differences. For example, Schmiedek et al. (2007) considered individual differences in reaction time and their relations to working memory capacity (WMC) and intelligence, Wilhelm et al. (2010) studied individual differences in face recognition, and Wilhelm et al. (2013) used confirmatory factor analysis to obtain a broader perspective of WMC as an individual difference construct. After more than a century of delay, these researchers are effectuating the program Cattell, Galton, and others originally had in mind.

Returning to the above quote from Cattell and Galton (1890), key terms are *constancy*, *interdependence*, and *variation*, which shows the influence of Francis Galton's statistical ideas. It is also remarkable that the type of tasks listed was far removed from what one might suppose mental testing is primarily about: ability, personality, or character. The paper has an appendix with comments by Galton, where he pays attention to exactly this aspect:

> One of the most important aspects of measurement is hardly if at all alluded to here and should be emphasized. It is to obtain a general knowledge of the capacities of a man [. . .] In order to ascertain the best points for the purpose, the sets of measures should be compared with an independent estimate of the men's powers [. . .]. The sort of estimate I have in view and which I would suggest [. . .] is something of this kind,—'mobile, eager, energetic; well-shaped; successful at games requiring good eye and hand; sensitive; good at music and drawing. (Cattell & Galton, 1890, p. 380)

We also see a quest here for establishing the *validity* of a mental test in being able to identify important characteristics of a gentleman. However, it must be noted that these were only plans; Cattell and Galton never actually collected and analyzed these type of personality data! By contrast, there were certainly earlier attempts of mental testing of personality, and one of them (in the seventeenth century) that attempted to assess reliability will be described in Sect. 1.2 of this paper, which discusses Christian Thomasius.

A much earlier example is ability testing in ancient Greece, which represented the different facets of the ideal Greek citizen. These tests were primarily of a vocational nature, but also included athletic abilities (Doyle, 1974). Even still earlier, in ancient China:

> The great Chinese philosopher and educator Confucius (551–479 B.C.) first classified people into three categories on the basis of intelligence: (1) people of 'great wisdom'; (2) people of 'average intelligence'; (3) people of 'little intelligence'. Confucius also made personality assessments of his students. (Zhang, 1988, p. 101)

Moreover, it is well-known that ancient China had a Civil Service Examination system, even though there is uncertainty about exactly how old it is (Bowman, 1989). In any event, in these earlier examples of mental testing, we do not find any notion of reliability or some form of advanced statistical analysis of the test results, which are basic elements of psychometrics. In the Dutch literature, Kouwer (1963) has taken perhaps the broadest possible historical view on the development of systems to characterize personality, but again without paying attention to actual measurement or quantification. However, for educational testing we do know when serious quantification started. Stigler (1992) and Mellenbergh (2011, p. 18) have identified that the first psychometric papers on the analysis of examination scores were published by Edgeworth (1888, 1890), in which he discussed the scaling of exams by using the normal distribution, correction for examiners bias, reliability of a single examiner, and more.

It is often stated that the modern form of intelligence testing started with the psychologist Alfred Binet and psychiatrist Theodore Simon in the period 1895–1910 (Boake, 2002), which was also the period that modern personality research started. As a particularly interesting example of the latter, Heymans and Wiersma (1906) collected large-scale questionnaire data on personality characteristics—such as introversion-extraversion and emotionality—in 437 families, including 3 generations of each family, summarized in 90 4-way contingency tables (*cf.* Heiser, 2008). But it should be noted, as argued by Mülberger (2017), that the emergence of mental testing in this period was more widespread and gradual than just the Binet-Simon breakthrough (as is also evident from Spearman's (1904) extensive summary of previous correlational studies of mental test data).

There is no doubt that Galton's major contributions to psychometrics have been, as pointed out by Drenth and Sijtsma (1990, pp. 4–5), his keen interest in individual differences, the need to work with standardized research designs, and his conceptualization of regression and correlation. How Galton developed the concept of correlation has been nicely described by Stigler (2010), while Walker (1929, pp. 92–102) explained why earlier writers in the nineteenth century hovered on the verge of discovery of correlation, but did not actually uncover it.

However, to regard Galton as "the founding father of psychometrics" (e.g., Furr & Bacharach, 2008, p. 9) is perhaps one step too far, for there are earlier roots of psychometrics, at least if we take a broader view of the field and do not restrict it to mental testing. Such a broader view was sketched by Jones and Thissen (2007), and the present paper tries to add three historical lines to their paper. Apart from the already mentioned early attempt to assess reliability, we will also discuss how the name and perspective of a discipline of psychometrics was conceived by Christian Wolff in the eighteenth century. The third and most important early root is the groundwork given by Gustav Fechner's psychophysics[1].

1.2 An Early Notion of Reliability: Thomasius' Numerical Rating System of Personality

Rating has been a method of assessing the degree of some natural characteristic by a human observer since ancient times (e.g., temperature, *cf.* Wright, 2016). However, according to McReynolds and Ludwig (1984, 1987) and Ramul (1963, p. 657), it was the German Enlightenment philosopher and jurist Christian Thomasius (1655–1728) who devised and applied the first quantitative rating scales for personality attributes of individuals. His purpose was to characterize each individual on 4 scales with 12 categories ranging from 5 to 60 in steps of 5 so that a *personality profile*

[1] As this chapter is part of a Festschrift in honor of Klaas Sijtsma, someone with a keen interest in the evolution of psychometrics (e.g., Van der Heijden and Sijtsma, 1996; Sijtsma and Junker, 2006; Sijtsma, 2016), who defends a position of psychological measurement between physics and statistics (Sijtsma, 2012), it is my hope and expectation that he welcomes these new trace lines in our history.

could be formed, and he conceived a concept of *interrater reliability* by asking several observers to rate the same person.

Thomasius' rating scales followed from his overall theory of personality, which was announced in a first programmatic publication entitled *New Discovery of a Solid Science, Most Necessary for the Community for Discerning the Secrets of the Heart of Other Men from Daily Conversation, Even Against Their Will* [English translation of Thomasius (1692a) by McReynolds and Ludwig (1984)]. The motivation for formulating this empirical approach using practical field work, interviews, and informal discussions with the common citizen was to arrive at the kind of knowledge a politician needs for effective policy making. Thomasius was convinced that a science of policy should not be legalistic, let alone philosophical. He started an autonomous discipline addressing what makes people tick (Barnard, 1971).

Further details about his rating scale system were provided by a second publication in the same year, entitled *Further Elucidation by Different Examples of the Recent Proposal for a New Science for Discerning the Nature of Other Men's Minds*. This English translation of the German title of Thomasius (1692b) is again by McReynolds and Ludwig (1984), and they also provided a translated version of the five basic postulates of his personality theory:

> I. There are four major inclinations from which all other inclinations spring. These are:
>
> 1. Rational love [*Vernünftige Liebe*]
> 2. Sensuousness [*Wollust*]
> 3. Ambition [*Ehrgeiz*]
> 4. Acquisitiveness [*Geldgeiz*]
>
> II. All human beings are characterized by these inclinations and all possess some part of each of them.
> III. At all times one of the four inclinations is dominant in a person.
> IV. The difference among persons in human inclinations must be recognized not only from the dominant inclination but also from the proportion of the other three.
> V. One can appropriately assign 60 points to the strongest inclination and 5 points to the weakest (or at times more) and then judge the remaining two in accordance with the difference between the 60 points and the value of the lowest inclination. (Thomasius, 1692b, p. 239)

The four inclinations indicated in postulate I are the basis of the four rating scales that are to be used to rate any individual on the basis of conversations with the rater (a trained observer). According to McReynolds and Ludwig (1984), "All kinds of data went into the rating determinations—educational, occupational, and familial information about the subject; reports of his daily habits; interpersonal styles; behaviors that the individual found pleasurable; and so on." They also comment that Thomasius' description of *Rational love* comes close to what we would now call *Altruism*, that *Sensuousness* is concerned with *Hedonic tone* (seeking pleasure and avoiding pain), that *Ambition* must be understood as *Social ambition*, and that *Acquisitiveness* not only relates to a *Passion for money* but also to *Stinginess* and *Envy*.

Regarding the numerical rating categories on the four attributes, it is noteworthy that they are seen as proportions (postulate II) and kept within the range 5–60 (postulate V). It is plausible that this particular choice of values for the rating categories was inspired by the usual scale markers in time measurement (60 min in an hour, 12 months in a year). Furthermore, only the dominant attribute gets the maximal score of 60 (postulate III), and the other attributes need to be seen in proportion to the dominant one. Due to the aim to compare patterns of attribute proportions between individuals (postulate IV), the whole approach seems to fit into what two and a half centuries later has been called *Q-methodology* (Stephenson, 1936, 1953; Cattell, 1952)—a small, but basic part of psychometrics.

Thomasius included a section in *Further Elucidation* called "About the Test of Certainty of This Science," beginning as follows: "Just as in mathematics, where there is no better way to check to see if one has calculated correctly than to repeat the process two or three times in order to find out if the sum is the same, I have thought that in the discovery of other truths, regardless of what the discipline it may be, this method might be the best way of checking" (quoted in McReynolds and Ludwig (1984)). He then gives an example of a single individual who was rated by himself and by two students who had been trained well in the method of scoring. It turned out that the three patterns were very close, a sign of considerable *interrater reliability*. Even without recourse to a numerical reliability coefficient, the expression of the patterns in quantitative terms obviously facilitated comparisons enormously. Note that this form of interrater reliability is different from the more usual form at present, in which for each attribute separately the ratings of different raters across individuals would be compared.

What was the impact of Thomasius' quantitative methodology? The short answer is: by the end of the eighteenth century, he was not taken seriously anymore. According to Barnard (1971), "To his German contemporaries and near-contemporaries Thomasius was something of an idol. Nineteenth-century intellectuals—Hegelians in particular—generally dismissed him as an unsystematic, facile eclectic, and only the present century has witnessed a moderate revival of interest in him, though scarcely beyond the confines of Germany." However, there does seem to be a renewed interest in the Anglo-Saxon world for his "desacralization of philosophy" (Hunter, 2000). Moreover, it is the irony of history that after Thomasius was forgotten, *Q*-methodology is thriving presently in political science and communication science (*cf.* Brown, 1991, 1993; McKeown & Thomas, 2013).

1.3 Qualities of the Soul Can Be Measured: Wolff's Proposal of Psychometria

Christian Wolff (1679–1754) was "arguably the most eminent German philosopher between Leibniz and Kant, and an important figure in the development of thought about the state and its tasks as well as about the national economy" (Drechsler, 1997). But he was also vitally important for the *Sciences of the Soul* (Vidal, 2011) and in particular laid the groundwork for their methodology, which he coined *Psychometria* (Ramul, 1960; Feuerhahn, 2004)).

In the beginning of his career, he had chosen to specialize in mathematics, obtaining his doctorate in 1703 at the University of Leipzig, where he was soon invited to become a staff member of the first scholarly journal in Germany, the *Acta Eruditorium Lipsiensium*. Apart from mathematics, he soon expanded into other areas within the *Faculty of Arts*, then including all fields of learning except *Divinity*, *Law*, and *Medicine* (Drechsler, 1997).

Due to the Great Northern War between an alliance of Denmark-Norway, Saxony, and Russia against the Swedish empire, Wolff decided to leave Leipzig in 1706, and he accepted an offer of the University of Halle, where he became Professor of Mathematics, upon recommendation of no less than Gottfried Wilhelm Leibniz (1646–1716). Drechsler relates:

> Wolff greatly enjoyed teaching [. . .] and also began lecturing in what we would today call Philosophy. [. . .] He was also by then a prolific and celebrated author, and was thus unanimously elected as Fellow of the Royal Society in London. [. . .] Embarrassed by the fact that a Prussian subject had thus been honored abroad but not at home, the Berlin Academy subsequently made him a member as well. [. . .]
>
> In 1723, however, Wolff had to flee from Halle in one of the most celebrated dramas in the academy in the eighteenth century. The incident which caused the drama was his farewell address as *Prorector* in 1721. [. . .] In it, Wolff described the Chinese philosophy and ethics, namely Confucianism, as rather admirable and really as largely in agreement with his own moral principles. Indeed, his lecture submitted *proof that one could find moral truths through the powers of reason of natural Man without the help of divine revelation*. [. . .] If one follows Wolff's argument, there remains little place for Christian mission; to the contrary, it seems that one could actually learn a few things from the Chinese. (Drechsler, 1997, p. 112)

It became a scandal of immense proportions, where opinion leaders adhering to a strict form of Protestantism forced the King of Prussia to accuse him of gross impiety and to order him to leave the city of Halle and all other Prussian lands immediately. Fortunately, Wolff could use his influential network to escape and obtained the *Papin's chair of Mathematics and Physics*, as well as the *chair of Philosophy* at the University of Marburg, just a week later.

It was in Marburg that Wolff published major works about philosophy and psychology, including *psychometria*. Concerning philosophy, "He reemphasized Leibniz's conviction that mathematics has a role in philosophy. As he wrote in his *Discursus praeliminaris de philosophia in genere* (Wolff, 1963, original work published in 1728), philosophy must use mathematical knowledge. For in philosophy we wish to have complete certitude [. . .] [and] in many cases, complete certitude depends on mathematical knowledge and demonstrations" (Leary, 1980, p. 155).

Concerning psychology, Wolff's point of departure was Leibniz' doctrine that *Intensity*, *Continuity*, *Variation*, and *Covariation* apply equally well in the material as in the mental realm. What now follows is a summary of Leary (1980, pp. 154–155). With respect to *Intensity*, the concept of *force* in physics corresponds to the concept of *clarity of ideas* in psychology. The principle of *Continuity* states that all differences in nature are different in degree rather than in kind, whether we consider *physical motion* or *mental consciousness*. *Variation* refers to the principle that every material object and every mental idea undergoes continuous change in the degree of its intensity. Material objects change in *momentum*, and mental concepts change in *amount of consciousness*. *Covariation* refers to the principle that change in one part of the system leads to a (reverse) change in some other part of it. For the material world, it implies that an *increasing force* in one body corresponds to a *decreasing force* in another, while for the mental world it implies that an *increase of clarity* in one idea corresponds to a *decrease of clarity* in another.

In his important work *Psychologia Empirica, methodo scientifica pertractata* Wolff (1962a, original work published in 1732), explained that

> The art of discovery (*ars inveniendi*), which involves deducing unknown truths from already known ones, can proceed either *a priori* or *a posteriori*. In the latter case, which is the only one of interest to empirical psychology, findings are based on observation or experimentation (*ex experimentis*). Both are forms of 'experience' (*experientia*), that is, of knowledge acquired by paying attention to our perceptions. Observation involves no voluntary alteration of nature, experimentation (*experimentum*), by contrast, requires it. Watching the sky cloud over is an observation, whereas pumping air from a pneumatic machine is experimentation. The *ars observandi* used by physicists, doctors, and above all astronomers, is the proper method of empirical psychology. *Ars experimendi*, on the other hand, is used only by physicists—even if, Wolff suggested, it could be applied to the whole of philosophy and even to natural theology. (Vidal, 2011, pp. 128–129; footnotes and references to the source omitted here)

In the same work, Wolff also formulated his mathematical law about the magnitudes of pleasure and displeasure: "Pleasure is proportional to the perfections of which we are conscious, as well as to the certainty of our judgments concerning these perfections." In a footnote he added: "These theorems belong to psychometry, which conveys a mathematical knowledge of the human mind and continues to remain a desideratum. It should teach us how to measure the magnitudes of perfection and imperfection and also the certainty of a judgment, and insofar determine [both measures]" (Ramul, 1960, p. 256). As recently noted by Mei (2021, p. 91), Wolff's psychometria is "a form of methodological mediation that implies the ability to measure the *effects* of the soul rather than its *substance*. In other words, psychometria allows us to take into scientific consideration the possibility of a first form of the *naturalization* or *mathematization* of the mind."

More specifically, in his *Philosophia prima, sive Ontologia, methodo scientifica pertractata, qua omnis cognitionis humanae principia continentur,* Wolff (1962b, original work published in 1736) gave a number of examples of how psychometrics could proceed to measure the qualities of the soul. What now follows is a summary of Mei (2021, pp. 91–96), who also gives source references. As a preliminary, Wolff states that each quality is measurable (and calls it a "common prejudice" that not all qualities are measurable). For instance, *density of fluids* is a quality and can be measured with an aerometer, *temperature* can be measured with a thermometer, and the *gravity of air* can be measured with a barometer. Moreover, qualities have a degree, and therefore we have the possibility to establish the *size of the degree*, which he regards as an *imaginary notion* (recall Leibniz's principle of the continuity of nature). He mentions *degree of speed* and the notion of *substance* as other examples of imaginary notions. Wolff mentions the following three possibilities for psychometric measurement.

1. Measuring *duration* and *clearness* of psychic phenomena.

Thoughts are not immediate and some time is required to allow human thought to proceed. The term *time* refers to continuous processes and *duration* to the simultaneous existence of several successive things. Time can be represented through the imaginary notion of a straight line consisting of a continuous series of points, so that there is an analogy between time and number. Furthermore, perceptions can be partial or composite. If ideas belonging to a visible object and its corresponding word become clearer over time, it is because the movement of the material ideas is faster. A composite perception consists of several partial perceptions, and if the partial perceptions become clearer, then the corresponding composite ones are perceived distinctly. The greater the number of the particular, clear perceptions, the greater the degree of the distinctiveness of the subsequent composite perception. So it is duration and/or the number of required perceptions that allows measurement of psychic phenomena.

2. Measuring the *intensity* of psychic phenomena: Memory and the imagination.

According to Wolff, if something is distinctly perceived, it is also easier to retain in memory. Therefore, the quality of memory admits different degrees that may vary from individual to individual. We can identify this degree of quality by looking at the time spent holding on to an idea in the mind or to the number of acts by which the reproduced ideas are delivered to memory or with which they are held in memory. Therefore, people with a “great” memory can reproduce the ideas of many things, like those who can remember the whole Bible and can quote each part of it in the right order or those with a “long” memory who can remember a long series of things or events. Imagination also has different degrees, to the extent that it reproduces the ideas of many things, while memory recognizes ideas reproduced. There are individual differences in the quality of the soul, due to a possible diversity of nerve fibers. Body and soul are closely connected and interdependent, which implies that psychometria enables measurement of the *effects* of the soul, instead of measuring the soul as *substance*.

3. *Measuring degrees of attention and individual differences.*

A major pillar of Wolff’s psychometria is that degrees are the “quantities of qualities.” Also, every time we talk about degrees, we do not refer to objects, individuals, or activities, but to *relations* between them. For example, we say that this line is three or five times as thick as another one. Likewise, for intellectual qualities we can say that one person’s ability is larger than someone else’s. Degrees of attention can be greater or smaller depending on (*a*) how much the sense organs are involved in perception (which can be measured by their arousal), (*b*) how long mental content is preserved or extinguished, (*c*) how many different things a person can pay attention to simultaneously, (*d*) the selectivity with which a person pays attention to some objects but not others, and finally (*e*) whether someone typically pays attention to actual objects or to imaginary objects. These examples show that levels of attention and individual qualities can be measured by counting relational data.

What happened to Wolff's program of empirical psychology and psychometria? First and foremost, he did not collect any data himself to see how his methodological ideas would work, and neither did his contemporaries. But there were several authors in the eighteenth century who also dealt theoretically with the question of mental measurement (for an overview, see Ramul, 1960 and Vidal, 2011). One of them was Gottlieb Friedrich Hagen (1710–1769), a philosophy teacher at the Bayreuth gymnasium, who was a follower of Wolff and had a position as Adjunct Professor in the faculty of philosophy at Halle in the period 1731–1737. Like Wolff, he wanted his work to have the universal applicability of mathematics while also being socially useful. As Vidal notes:

> He imagined psychological experiments [. . .] that would alter the soul, for example, by scaring people and then observing their reactions. Such experiments could contribute significantly to self-knowledge [. . .]. Hagen also conceived a *dynametria* to measure the faculties (*dunamis*) of the soul, again within the framework of a sort of quantitative casuistry. He argued that, like the mechanical faculties of the body, the representative faculties of the soul are finite in number; since they vary considerably from individual to individual, they may be compared quantitatively. (Vidal, 2011, p. 130)

Ramul (1960) concluded his pioneering essay by noting that although measurement of mental phenomena attracted the attention of several eighteenth-century scholars, much of what the individual authors had to say were their personal "ideas" with little continuity in their development, except for some of Wolff's students. No one carried out any actual measurements. By contrast, he says, "by far the larger part of the psychological measurements known to us from that time [. . .] have been carried out not by psychologists (or philosophers) but by naturalists [who studied such elementary phenomena as visual acuity, the size of the blind spot, and the duration of visual afterimages]. And thus the program of [. . .] psychometry remained wholly on paper in the eighteenth century" (Ramul, 1960, p. 264). Although Wolff's psychometria did influence Immanuel Kant (1724–1804) and Johann Friedrich Herbart (1776–1841) on a conceptual level (Leary, 1980; Sturm, 2006), the definite start of psychometrics had to wait until the second half of the nineteenth century.

1.4 Birth of Experimental Design and Psychological Scaling Methodology: Fechner's Psychophysical Paradigm

Psychophysics is the brainchild of Gustav Theodor Fechner (1801–1889), physicist and philosopher with important contributions to psychology, psychometrics, and statistics. Born in Gross Särchen (a small village in the German region of Saxony) as the son of a clergyman, he started studying medicine in 1817 at the University of

Leipzig and earned a baccalaureate in 1822. But he had a lot of other interests: “at about the same time he began writing a series of sometimes mystical philosophical pieces on the identity of mind and matter, a practice that was to last throughout the rest of his life” (Stigler, 1986, p. 242). He did not finish medicine, however, and got more interested in physics. To earn some money, he began translating the textbook of Jean-Baptiste Biot (1774–1862) on experimental physics from French into German and started lecturing in 1824. Then he published a paper on the galvanic battery (Fechner, 1831) that was inspired by the pathbreaking experimental work of Georg Simon Ohm (1787–1854) on the laws of electricity published in 1827. It made his reputation as a physicist, and he was elected as extraordinary professor of physics at Leipzig in 1831, and in 1834 he was promoted to full professor of physics at the same university.

Stigler (1986) has emphasized the lasting influence of Ohm’s work on the young Fechner, because his 1831 paper already:

> bore the hallmark of Fechner’s later work. Even though it made no use of probability in its analysis, it was an extensive, painstakingly-detailed account of a series of multifactor experiments. Everything that could be varied was varied; everything that could be measured was measured; everything that could be recorded was recorded. And in all this mass of detail (the record of the experimental results alone covers about 200 pages) he did not lose sight of overall objectives. (Stigler, 1986, p. 243)

Several biographies of Fechner have associated his turn to psychophysics to a period of illness and personal crisis in his early forties after he had ruined his eyesight by doing experiments in subjective color perception, looking often at the sun through colored glass. He recovered when he entered his garden not wearing the mask that covered his eyes for many years. Overwhelmed by how beautiful everything looked, especially the flowers, it seemed to him “like a glimpse beyond the boundary of human experience” (the last quote is from Fechner’s autobiographical notes as cited in Murray, 2021, pp. 76–79). But Stigler is not impressed:

> As appealing as such stories are as devices for raising the origin of scientific ideas to the level of heroic myth, they do not seem to be essential to un understanding of Fechner’s intellectual development. The urge to experiment, the interest in physics and both mind and body, and an ambition to influence human thought—all the essential ingredients were already in place in the 1820s. The Fechner who by 1855 had begun the extensive experimentation that led to his *Elemente der Psychophysik* was essentially the same Fechner who had devoted two full years to the study of electrical current in 1829–1831. (Stigler, 1986, p. 243)

For additional intellectual influences of earlier scientists on Fechner, these go back “nearly a hundred years to the measurement of sensitivity and of the discriminatory capacity of the senses as accomplished by physiologists and other natural philosophers” (Boring, 1961). At this point, we cannot elaborate on that story because of our focus on psychometrics, but we have to give a brief introduction to Fechner’s Law and how to check it experimentally.

In the next summary, we use Fechner’s notation, as given in the short excerpt from the 1860 *Elemente der Psychophysik* reproduced in Miller (1964, ch. 4). It is well-known that *Weber’s law* states that the sensation difference between two stimuli remains constant when the relative stimulus difference, or the increase in one stimulus, remains constant. Let the stimulus which is increased be called β and the small increase $\mathrm{d}\beta$, where the letter d is to be considered simply as a sign that $\mathrm{d}\beta$ is a small increment of β. The relative stimulus increase therefore is $\mathrm{d}\beta/\beta$. Choosing d so that two sensations are "just noticeably different" (*jnd*), Fechner took the *jnd* as the unit of sensation, which could be counted to form *magnitudes of sensation*. Let the sensation that is dependent upon the stimulus be called γ and the small increment of sensation be $\mathrm{d}\gamma$. Now, Weber's law is usually stated as $\mathrm{d}\beta/\beta =$ constant. By invoking the assumption that the change in sensation $\mathrm{d}\gamma$ is equal for all *jnd*s, Fechner could transform Weber's law into

$$\mathrm{d}\gamma = \kappa \frac{\mathrm{d}\beta}{\beta}, \qquad \text{(fundamental formula)}$$

where κ is a constant dependent on the units for γ and β. Fechner's next step was to consider the fundamental formula as a differential equation and integrate it. The result is

$$\gamma = \kappa \log \frac{\beta}{b}. \qquad \text{(measurement formula)}$$

Here b is the threshold value of the stimulus β, a value at which the stimulus is no longer detectable, called the *stimulus limen L* or *RL* (from the German *Reiz Limen*), corresponding to $\gamma = 0$. The scale of γ is then the number of *jnd*s that a sensation is above zero. Finally, Fechner made one more step by regarding b as the unit for the measurement scale of the stimulus β, so that the measurement formula simplifies to what he called the *metric formula*:

$$\gamma = \kappa \log \beta, \qquad \text{(metric formula)}$$

the form usually found in the textbooks, where the metric formula is usually called *Fechner's law*. Fechner himself preferred to keep using the name *Weber's law*, out of respect for his physiology professor in medical school, Ernst Heinrich Weber (1795–1878) himself (for more on Weber's importance as a pioneer of quantitative psychology, especially his experiments on the sensitivity of the touch sense, see Murray, 2021, ch. 3). But according to Stigler (1986, p. 243), such emphasis on Weber might be misleading in the light of the fact that Fechner was so well acquainted with the early work of Ohm who developed a similar logarithmic relationship between the loss of force in a current and the length of a wire.

Fechner's law generated substantial objections and a lot of discussion among the psychologists of his time and later (for brief overviews, see Boring, 1961; Stevens, 1961; Zudini, 2011, pp. 82–84). Also, it might be noted by the reader, as did George Miller, that

> Fechner's law relating subjective sensation to objective stimulation is exactly the same as D. Bernoulli's law relating subjective utility to objective money. But Fechner's law was

> immediately strengthened by his proposals for psychometric methods of measurement, whereas methods for measuring the subjective magnitudes that Bernoulli was talking about were not developed until the middle of the twentieth century. A theory is good, but a theory plus measurements is a great deal better. (Miller, 1964, p. 99)

As a matter of fact, Fechner did know that Daniel Bernoulli (1700–1782) came up with the concept of diminishing marginal utility and suggested a logarithmic function for it. In particular, in *Elemente der Psychophysik*, he quoted Bernoulli's treatise *Specimen theoriae novae de mensura sortis* published in 1738, where Bernouilli writes: "Certainly the value must not be estimated from the price of the thing, but from the advantage acquired therefrom. The price is estimated by the thing itself; the advantage, by the state of the persons involved. Thus, without doubt, the gain of 1000 ducats is far more important for poor persons than for rich persons, although the amount is the same for both. [. . .] Thus, it is indeed exceedingly probable that any small advantage adds to the ultimate good in reciprocal proportion to their status of the people involved" (quoted in Fechner, 1860, p. 197). But immediately after this quote, Fechner remarks: "He bases his differential formula [. . .] and his logarithmic formula [. . .] on these considerations. We later base the same on Weber's law in a more general way" (Fechner, 1860, pp. 197–198). A theory is good, but a theory motivated by verified empirical regularities is even better! He also points out the role of Laplace, who developed Bernoulli's idea further in his *Théorie analytique des probabilités* (1812), and to Poisson, who mentioned and accepted it in his *Recherche sur la probabilité des jugements en matière criminelle et en matière civile, précédés des règles générales du calcul de probabilités* (1837).

In checking the logarithmic law for sense data, values of the physical stimulus R (from German *Reiz)* were taken as its strength β in a measure valid for the chosen domain (weight, touch, brightness, pitch, and so on) and given in terms of b as the unit of measurement (*cf.* the measurement formula). Let us elaborate what Fechner's proposals were for finding scale values γ of the psychological response S (from German *Sensation*). [Confusingly, the initials of the German terms are the reverse of the English terms Stimulus and Response!] There are three basic methods and two additional ones, which are regularly described in the classical texts of Guilford (1936), Brown and Thomson (1940), and (partly) Bock and Jones (1968). They all refer to Titchener (1905) as the basic source. Here is a brief description, where we follow the distinction suggested by Brown and Thomson (1940) to distinguish between names for *methods of experimenting* in order to collect data (*experimental design* in modern terms) and *processes of calculation* after the data have been collected (*analysis methods*). The basic psychophysical methods are as follows.

1. *Method of reproduction or adjustment*. This experimental design is one of the oldest and most fundamental of psychophysical methods. According to Titchener (1905, p. 160), it is "a free gift to psychophysics from the exact sciences of physics and astronomy." Fechner introduced it in *Elemente der Psychophysik* with tactual and visual measurements. In his own words, from the 1882 revision of the *Elemente*, English translation by Guilford (1936, p. 25):

> A certain distance, e.g., between compass points or parallel threads, is presented. This I call the normal distance. I am to make another distance, the error distance, as nearly equal to this as it can be made by eye. First of all, starting from an error distance that is too large or too small, I adjust it roughly, in an irresponsible sort of way, to apparent equality with the normal. Then I consider whether or not it really corresponds to sensible equality, and I shift the boundary of the error distance, thread or compass point, to and fro—until I seem, with a definitive adjustment, to have touched equality as closely as I may. (Fechner, 1882, p. 105)

In this case, the stimulus was an interval. More generally, the task for the subject is to *adjust* or *reproduce* a variable stimulus V, so that it appears subjectively equal to a given standard or *comparison stimulus* C. Anyway, the task is repeated a large number of times, so that we get a distribution of numerical adjustments. Method of analysis for this design is called the *method of average error*, meaning that we take the arithmetic mean of the observed scale values of the reproduced Vs. This choice was driven by the time-honored decomposition $Observation = Truth + Error$, used by astronomers in the 1820s, who had no doubt in their mind that they were "after something real, definite, objective, something with an independent reality outside of their observations, a genuinely Platonic reality inherited from the then-unshakable edifice of Newtonian theory" (Stigler (1992, pp. 61–62). It is also the basis of *classical test theory*, pioneered by Spearman (1910), and of *signal detection theory* (Link, 1994).

2. *Method of limits or method of minimal changes*. Primary use of this experimental design is the determination of sensory thresholds. For the stimulus limen RL, the experimenter decreases a variable stimulus V in small steps until it is no longer detected. For the *difference limen DL*, we have a pair of stimuli, V (a *variable* stimulus) and C (a *constant* or *standard* stimulus). V is first made equal to or slightly smaller than C and then decreased in small steps until the observer calls it *just noticeably smaller* than C. If there are N repetitions of the procedure, the simplest analysis method used is to calculate again the mean of the midpoints between C and the last V. According to Guilford (1936, p. 115), the original *method of just noticeable differences*, which was already used by Weber in 1829 to measure *jnd*s in passive pressure and lifted weights, presupposed that a human observer can recognize a *jnd* when he sees one. Weber would follow the procedure described above and was ready when the observer reported that he perceived a *jnd*. Fechner recommended a change in the method that was an improvement and has been permanently adopted. The change is to also start from positions of *extreme inequality*; now, the sequence results in the new notion of a *just not noticeable difference* (a *jnnd*), which is usually slightly smaller than the *jnd*. One then takes the average of the *jnd* and the *jnnd* as the true limen. The occurrence of different limiting values for these two starting positions suggests the presence of a *perceptual hysteresis* effect. As a matter of fact, Hock and Schöner (2010) have recently considered several possible mechanisms for such effects, detectable by a *modified* method of limits. There were already more variations in experimental design earlier, for which Urban (1907) is a good source.

3. *The constant method (or method of constant stimuli) and the method of right and wrong cases*. These methods can be used for determination of stimulus limens (*RL*s), differential limens (*DL*s), and equal sense distances, as well as the determination of other psychological scale values outside the strict realm of psychophysics. It is regarded as the most satisfactory of all Fechnerian scaling methods. The experimenter selects in a pilot study a limited number of stimuli, usually four to seven, that are going to be *constant* during the experiment. Let us call them $C_j, j = 1, \ldots, n_c$. Next, an additional stimulus T is selected as the *target*, somewhere on the physical continuum depending on the specific purpose of the experiment. For example, if the target is a stimulus limen *RL*, T is typically defined as the physical stimulus that has a probability equal to 0.5 of producing a response, which corresponds to a scaled value of $\gamma = 0$ on the psychological scale. Each constant stimulus C_j is then paired with the target T, and these pairs are presented either simultaneously or successively to the observer in prearranged or random order. The observer has to tell which of the two is "greater than" or "above" the other one, or the reverse. The presentation of each pair is repeated a large number of times, say n_j times, and the observations can be summarized in n_c relative frequencies p_j. For a differential limen *DL*, the C_j and T_j are in fact *pairs of stimuli*, and this case is called the *method of constant stimulus differences*, in which we have *pairs of pairs*, which are compared in terms of the magnitude of their sense differences.

The major analysis method developed by Fechner was called the *method of right and wrong cases*. This ingenious method finds the scale value of the target as the point on the physical scale that is the *median* of the discrete distribution of the comparison stimuli. It is by definition the location for which half of the judgments "C_j greater than T " are right (those to the right of the median), while the other half of the judgments "C_j greater than T " are wrong (those to the left of the median). Now, how do you find the median of a discrete distribution? Several simple methods were used to determine the median, such as linear interpolation, but Fechner came up with a new, pathbreaking procedure. Since each C_j has a relative frequency p_j of "greater than" judgments up to that point, these observed relative frequencies will tend to be *monotonically increasing*, within the interval 0.0–1.0. Fechner proposed to fit a cumulative distribution function to the data, in particular the *normal ogive*. Given that choice, it is easy to find the median as the inflection point of the curve (the scale value β corresponding to a probability of 0.5 on the y-axis), where the curve's positive acceleration changes into negative acceleration. Note that to use the inflection point of the normal ogive to find the scale value of a particular stimulus is essentially the same as the definition of the item difficulty parameter in *item response theory* (IRT), as used by Lord (1952).

Fechner used the original parameterization of the cumulative normal curve that Gauss had used in his first publication on least squares in 1809 (see Stigler, 1986, pp. 140–143). Gauss expressed the argument of the exponent as $-h^2\Delta^2$, with Δ the usual error term and h a *precision* parameter, which indicates the steepness of the normal ogive, or the *sensitivity* of the observer. Thus, the relation of h with the standard deviation that is now commonly used is $h = 1/\sigma\sqrt{2}$. Fechner had a justification for the hypothesis of the normal ogive (which has become known as the *phi-gamma function*, and hence the *phi-gamma hypothesis*). It was suggested to him by his Leipzig University colleague August Ferdinand Möbius (1790–1868). Asked to judge whether one stimulus is “greater than” another one, the observer would form a mental estimate of each stimulus, making a normally distributed error, and reports the difference between the two mental estimates (Stigler, 1986, p. 247). In this way, Fechner could measure not only (possibly different) limens for each individual observer but also individual differences in their sensitivity or precision, associated with smaller standard deviations in their mental estimates. The normal ogive or phi-gamma function in the context of psychophysics has been called the *psychometric function* by Urban (1910), in analogy with the *biometric function*, which models a binary outcome (e.g., dying) as a function of some predictor (e.g., age).

Fechner’s method for fitting psychometric functions was simple *Gaussian least squares*, which in the early nineteenth century had become a standard analysis method for astronomers and geometers, but for psychologists it was an important innovation (Fechner, 1859). Nevertheless, Müller (1878, 1879) argued that proportions near 0.5 should be weighted more than proportions deviating from 0.5 in either direction, because the standard errors of proportions are a function of the mean. A further justification for using *weighted least squares* with weights $n_j\, /p_j(1 - p_j)$ was provided by Urban (1908, 1910). Hence they were called *Müller-Urban weights*—and still mentioned as a term in the current APA Dictionary of Psychology. This weighted procedure is known as *probit analysis*.

There are two more classical experimental designs that aim at finding scale values for psychophysical or psychological stimuli to which Fechner contributed only partly. They were proposed with special interest in scaling *supraliminal* stimuli (relatively far apart), in which case a psychological S-scale of sensations cannot be formed by counting *jnd*s.

4. *The method of equal appearing intervals (or method of equal sense distances).* In its original form, the *method of equal sense distances* required the observer to *bisect* a given distance on a specific psychological continuum. For example, “given two sound intensities, R_1 and R_3, the latter being of greater intensity than the former, O [the observer] had the problem of finding a stimulus R_2 such that the interval $R_1 - R_2$ equaled $R_2 - R_3$” (Guilford, 1936, p. 143). This task is the simplest one for obtaining equal sense distances, but it can already be used for testing Weber’s law in cases where the stimuli are supraliminal. The reasoning is that if we define the small but supraliminal increments $\Delta R_1 = R_2 - R_1$ and $\Delta R_2 = R_3 - R_2$, then it is easy to show that Weber’s law implies that R_2 must be the geometric mean of R_1 and R_3. This prediction can be tested on observations obtained in a bisection experiment (Guilford, 1936, p. 144). For a classic application to tonal intervals, the reader is referred to Pratt (1928).

In the more general *method of equal appearing intervals*, the observer is asked to sort a relatively large set of n stimuli into a relatively small set of m *piles*, or *classes*, separated by *equal sense distances*. The stimuli sorted within a pile should have high psychological *similarity* (e.g., they should sound about equally intense). The observations can be collected in a frequency table of n rows (stimuli in the order of their scale value, if known) by m columns (classes labeled with consecutive integers), and the pattern that one would expect is that of a discrete bivariate normal distribution with negative correlation: i.e., high frequencies in the upper left corner of the table, extending along the diagonal to the lower right corner, tapering off toward the upper-right and the lower-left corners. We can now define the sensation scale S by allocating equal intervals between the classes, with the consecutive integers as scale values. That allows us to use the *method of right and wrong cases* to calculate S-values for the stimuli. For each row, we first transform the frequencies into cumulative frequencies and next fit the psychometric function to smooth them, with the equally spaced S-values on the x-axis. Then the median can be found as the inflection point of the curve and defines the scale value of the row stimulus. Fechner's law can be checked by plotting these against the physical R-values. This analysis method was suggested by Thurstone (1929), as a correction on the approach taken earlier in the notorious *Sanford weight experiment* on lifted weights (Sanford, 1898; also see Titchener, 1905, pp. 82–85; Murray, 2021, pp. 86–89), in which the average physical scale value in each pile was considered as the adjusted R-value and plotted against the equally spaced class intervals on the S-scale to check Fechner's law. Thurstone (1929) illustrated his corrected procedure with an example of 96 cards filled with irregularly spaced dots, where the stimulus magnitude was the number of dots on the card, and they were sorted in 10 piles. It showed convincingly that Fechner's law could be verified to hold for supraliminal stimuli.

5. *The method of choice and the method of paired comparisons*. Only the first steps in the development of these methods will be briefly described. Fechner was the first to study systematically the aesthetic properties of the so-called golden section in his treatise *Zur experimentellen Ästhetik* (Fechner, 1871). Among other methods, he proposed the *method of choice* (die *Wahlmethode*), in which an observer must choose one stimulus among k alternatives (see Guilford, 1936, pp. 222–225, and Green, 1996, for more detailed accounts of this development). For $k = 2$, the observer has to choose one stimulus out of a pair, and if this basic element is repeated for more pairs of m stimuli (not necessarily all of them), we arrive at the *method of paired comparisons*. An early application of this method was in the construction of a handwriting scale by Thorndike (1910), but a good method to analyze paired comparisons was still to be desired at that time.

With its methodological innovations, experimental designs and analysis methods, Fechner's work prepared the ground not only for experimental psychology but also for psychometrics, statistics (Sheynin, 2004), and even probability theory: the prominent applied mathematician Von Mises (1912) referred to Fechner's posthumously published work *Kollektivmasslehre* (1897) as one of the inspirations that later brought him to introduce *randomness* as a basic concept in the theory of probability (Von Plato, 1994, pp. 182–183). Several improvements in the design of the constant method were introduced in a large-scale weight-lifting experiment by Peirce & Jastrow (1885). They wanted to measure *jnd*s as precisely as possible, because of their skepticism about the existence of difference limens. The most important improvement was to determine the order of presentation of the pairs of weights by randomization, using two packs of playing cards (Stigler, 1978). They found that the sensitivity of the subjects was far below Fechner's threshold and concluded that there was no evidence for a difference limen (Stigler, 1992). In connection with this experiment, and similar ones in early experimental and educational psychology, Dehue (1997) has defended the claim that randomized designs were introduced by psychologists before Ronald Fisher introduced them in his classic handbook *The Design of Experiments* (1935). This claim was challenged by Hall (2007), who placed Fisher's rationale for the promotion of randomization in the tradition of agricultural field experiments starting in the middle of the nineteenth century. But it would lead us outside the scope of this chapter to pursue this priority issue further.

The impact of Fechner's psychophysics on Wundt and his doctoral students has been very large, as most of their experiments involved his methodology, except for Wundt's notion that psychologists should be the observer or subject (and not an arbitrary person). For only trained psychologists could use *introspection* to report their *apperception*, which is an unconscious process that interprets raw sense data in relation to past experiences. Prominent among Wundt's students were psychologists from the United States (in total 33 of them), including James McKeen Cattell (1860–1944) and Edward B. Titchener (1867–1927), who already made their appearance in this paper. Especially Cattell at Columbia University was a strong advocate for the statistical turn in American psychology in the period 1890–1915, part of a more general rise of statistical methodology in anthropology, sociology, and economics (cf. Camic & Xie, 1994). One of Cattell's students, Edward L. Thorndike (1874–1949), became the founder of modern educational psychology and educational testing at Teachers College of Columbia University and was one of the founding fathers of the Psychometric Society. From England, Charles Spearman (1863–1945) was also a Wundt PhD student, but he was more influenced by Francis Galton and the rise of mental testing in the last decade of the nineteenth century. Galton himself has praised Fechner in a letter from 1875, for having laid "the foundation of a new science [. . .] [in which a] mass of work by Arago, Herschel and various astronomers fall in as a part of the wide generalizations of Fechner, and much criticism and recognition of him will be found in Helmholtz" (Sheynin, 2004).

The formulation of the concept of the *psychometric function* was for sure Fechner's greatest contribution to psychometrics. It was recognized in standard textbooks, not only in the specialized ones already mentioned but also in more general popular texts about statistics. For instance, Truman Kelley's textbook *Statistical Method* (Kelley, 1924) has a section on psychophysical methods, in particular the method of right and wrong cases. But the psychometric function was also a topic for further clarification and research (e.g., Boring, 1917; Thomson, 1919; Urban, 1933). By far the most clarifications, generalizations, and extensions came from Louis Leon Thurstone (1887–1955) during the start of his career.

In our discussion of the method of equal appearing intervals, the psychometric function was used to find S-values of the stimuli as the median of a discrete distribution defined on equally spaced intervals on the sensation scale. It should be noted that nowhere in Thurstone's (1929) procedure he needed to use R-values (physical magnitudes) to derive the S-values. More generally, Thurstone (1927a) had already sketched a new framework for psychophysics, in which he introduced the concept of normally distributed *discriminal processes* with which an organism differentiates stimuli by calculating *discriminal differences*, which leads to the normal ogive psychometric function. It reminds us of the already mentioned suggestion made by Möbius to Fechner for justifying the phi-gamma hypothesis. In this suggestion, Möbius assumed that *mental estimates* of each stimulus were made, with a normally distributed error, and supposed that the differences between two mental estimates are reported by the observer. Indeed, Thurstone (1927a) fully developed the same idea and showed that Weber's law and Fechner's law are independent of each other and also that equally often noticed differences are not necessarily equal on the psychological continuum.

Within the same general framework, Thurstone (1927b) formulated his famous *Law of comparative judgment*. We already briefly met the experimental design to which this law applies in our discussion of the 5th Fechnerian method, i.e., the method of choice and paired comparisons. In a paired comparison design, m pairs of stimuli out of a set $\{S_1, \ldots, S_n\}$ are formed, where m can be the total number of possible pairs ½ $n(n - 1)$, or some subset of it. The subject is asked to compare the stimuli in a pair on any psychological attribute of interest and make a choice which one *dominates* the other. The dominance judgment may be *personal preference*, for example, when the stimuli are odors and the subject has to indicate which one smells more pleasant. Or the judgment may be an expression of *social or moral values*, for example, when the stimuli are crimes or offenses and the subject has to indicate which one is more serious. Upon replication of these judgments for one subject or across a number of different subjects, relative frequencies can be determined for each pair, and the Law of Comparative Judgment forms the basis for a statistical analysis that finds scale values for the stimuli on a psychological continuum, or *Thurstone scale*.

The assumption that each stimulus generates a normally distributed discriminal process in the mind of the subject(s) leads to a model in which the probability of making a choice of S_i over S_j is equal to the normal ogive of the difference in the means μ_i and μ_j, divided by the discriminal dispersions σ_{ij}, which are related to the standard deviations and the correlation between the two processes. The most often encountered case in which one assumes that $\sigma_{ij} = \sigma$ is called *Case V*. It is in fact equivalent to a simple probit model in terms of the differences in the means and without interaction terms. An authoritative treatment of Thurstonian scaling and some of its extensions is Bock and Jones (1968). An overview including modeling the discriminal dispersions σ_{ij} by multidimensional scaling was given by Heiser and De Leeuw (1981), while Takane (1989) and Maydeu-Olivares (2001) gave formulations of Thurstonian models in terms of the analysis of covariance. Böckenholt (2004) proposed solutions for the arbitrary location of the origin in a Thurstone scale.

Out of the general model, a whole series of other psychological scaling methods emerged, such as the *method of successive intervals* (Saffir, 1937), *method of graded dichotomies* (Attneave, 1949), and the *law of categorical judgment* (Torgerson. 1958). In these methods, there are pairs of stimuli by judgment categories (or category boundaries), instead of pairs of stimuli by stimuli. If we consider pairs of ability test items by persons and collect dominance responses for them, we obtain an *Item Response Theory* (IRT) model with a normal ogive *item characteristic curve* (a psychometric function that gives the probability of a correct response for a person with a score somewhere on the *S*-scale of *ability*, where the inflection point of the curve is called the *item difficulty* parameter). A theoretical development of such an IRT model formed the basis of the new test theory by Lord (1952). In fact, Thurstone (1925) had already formulated the basic idea and had illustrated it with a set of data with Binet-type questions, collected by Cyril Burt on 3000 London school children. A more detailed treatment and comparison with current IRT methods, including the reasons for switching from the cumulative normal to the logistic function, can be found in Bock (1997). For an enthusiastic review of Thurstone's general scaling framework, see Lumsden (1980), who concludes his paper by saying: "During the 1920s Thurstone stole fire from the gods. (As a punishment they chained him to factor analysis.)" (Lumsden, 1980, p. 7). Indeed, Thurstone's work in the beginning of his career expanded the scope of psychophysical scaling enormously by allowing the inclusion of non-physical stimuli, which made it the early root of two main branches of psychometrics, the scaling branch and the IRT branch.

1.5 Conclusions and Discussion

We have seen that Christian Thomasius had already more or less done in 1692 with his rating system of personality what James McKeen Cattell and Francis Galton had in mind in 1890 with their outline of mental testing—except of course that Thomasius was not in the position to calculate the interrater correlations between the 12-point rating scales that he and his two students had collected on one particular individual. Anyway, as pointed out by Jones and Thissen (2007, p. 5), the proposal to use sensory reactions and motor skills as a way of assessing mental ability was invalidated by a study of a graduate student at Columbia, who found that, for each of Cattell's proposed tasks, the correlations with class grades were essentially zero (Wissler, 1901)—which effectively ended this approach to mental testing.

Unfortunately, we must also conclude that Thomasius, with his proposal to study different personality profiles of political leaders, was simply too far ahead of his time, because political psychology started to have an interest in this topic only somewhere in the 1970s (Simonton, 2014). Ironically, Galton and Cattell became best known for their interest in quantitative studies of another subpopulation of the human race: men of science.

For Francis Galton, that project started already with the publication in 1869 of his notorious book *Hereditary Genius*, in which he tried to demonstrate the genetic transmission of intelligence, drawing on data culled from biographies and biographical dictionaries of scientists, eminent military leaders, philosophers, lawyers, and artists (Galton, 1869). The French botanist Alphonse de Candolle, who had read *Hereditary Genius*, responded acutely with the publication of De Candolle (1873), a book in which he offered an elaborate statistical study of the lives of outstanding scientists (members of the Academies of Science from Paris, Berlin, and London, including their foreign members). As noted by Ruth Schwartz Cowan: "He found that a very high proportion had come from countries or cities that possessed a moderate climate, a democratic government, a tolerant religious establishment, and an important trade centre. He concluded that Galton was wrong and that environmental factors did indeed play a crucial role in the production of outstanding men" (Cowan, 1970, p. *ix*).

Galton's immediate response was to produce a similar study called *English Men of Science: Their Nature and Nurture* (Galton, 1874), in which he aimed to show—not surprisingly—the dominance of nature over nurture. This time his data were autobiographical replies to a long questionnaire, sent out to 180 eminent scientists (fellows of the Royal Society, and the like), of whom 100 were selected for statistical treatment. One conclusion was that a strong and innate taste for science is a prevailing characteristic among scientific men and another that they had fewer children than their parents (Godin, 2007, p. 696). It served Galton well in pursuing his political program of eugenics.

Cattell followed Galton with several projects of measuring eminent scientists. In 1895, he acquired the weekly journal *Science*, established in 1883, which had run into financial difficulties. He used it as a vehicle for reporting the results of his statistical studies on science, based on an extending directory of researchers, called *American Men of Science*. He started with 4131 entries in 1906 (Cattell, 1906a, b), which accumulated to 34,000 entries in 1944. The directory included their background characteristics, fields of study, estimates of scientific merit, measures of productivity, mobility, and so on. In addition, as documented in Webster (1985), he also developed a system of academic quality rankings on the level of institutions instead of individuals (Cattell, 1910).

Because of these projects in the measurement of science, Galton and Cattell are now regarded as *pioneers of scientometrics* (Godin, 2007), as is De Candolle (Szabó, 1985). Furthermore, Godin (2006) has described how and why systematic counting of publications, citations, and acknowledgements (the output side of science, a branch of scientometrics known as *bibliometrics*) originated with several other psychologists, following in Cattell's footsteps.

What can be said about the lasting influence of Christian Wolff? He was certainly correct in thinking that the duration and clearness of thoughts could be empirically studied, as well as intensity, memory, attention, and individual differences. He, too, was ahead of his time, for these topics had to wait more than a hundred years before they became incorporated in the empirical research programs of people like Wundt, Müller, Helmholz, Ebbinghaus, and their students in the second half of the nineteenth century. In evaluating Wolff's impact, Vidal (2011, p. 111) remarks: "When the Aristotelian frameworks disintegrated, by the 1720s at the latest, psychology became the science of the human mind. In university circles, it was Christian Wolff who gave this shift its most systematic form. Hegel mentions in his *Lectures on the History of Philosophy* that Wolff gave the discipline a systematic structure which had served as a standard 'down to the present day', that is, until the 1820s."

Empirical psychology had to wait until Fechner laid out the psychophysical paradigm for measuring sensation. Wundt and contemporaries incorporated Fechner's pioneering work on experimental design and measurement of sensation and at the same time started to criticize him and to come up with alternatives (Murray, 2021, Ch. 5–8; Zudini, 2011). A much discussed criticism was that mental phenomena would *in principle* not be accessible for quantification, called the *quantity objection*. Michell (2006) phrased the denial of this criticism as the *psychometricians' fallacy*. For nuanced discussions of the quantity objection, see Hornstein (1988) and Sturm (2006). We have already noted the strong influence of psychophysics in the early twentieth century on experimental psychology, psychometrics, and educational psychology in the USA, including the upcoming testing movement in both Europe and the USA. However, after World War II its influence was waning, partly because of the upsurge of Stevens's "new psychophysics" in experimental psychology, based on magnitude estimation (Bolanowski & Gescheider, 1991), and partly because of the reorientation in item response theory to the logistic psychometric function.

But the mathematical psychologists have kept the fire burning! For example, Luce (1959) has critically examined what different forms a functional law like Fechner's law can have, dependent upon the scale levels of the independent and dependent variable. Here, it should be noted that Rozeboom (1962) has shown that Luce's conclusions were too strong, because they were based on a principle that is dubious at best. On the positive side, Dzhafarov and Colonius (2011) have argued that a lot of criticisms on Fechner's work are based on misinterpretation (partly due to Fechner's own expository and terminological shortcomings) and that Fechner's law can be derived without the notion of a *jnd*. In addition, they indicated that if we replace the term *difference sensation* with a more modern-sounding term *subjective dissimilarity*, then this change of perspective leads to the conclusion that Fechner's theory has the *additivity property* of a *unidimensional distance* (Dzhafarov & Colonius, 2011, p. 129). They also give examples of generalizations to multidimensional Riemannian geometry.

Finally, Zudini (2011) has shown convincingly that Fechner's system satisfies the conditions posed by the principles of *classical measurement* in Book V of Euclid's *Elements*, which poses a theory of *proportions between magnitudes* (Euclid, 1956). It appears that time is coming for someone to write a unifying book entitled *Elements of Psychometrics*.

References

Attneave, F. (1949). A method of graded dichotomies for the scaling of judgments. *Psychological Review, 56*(6), 334–340. https://doi.org/10.1037/h0063110

Barnard, F. M. (1971). The “Practical Philosophy” of Christian Thomasius. *Journal of the History of Ideas, 32*(2), 221–246. https://doi.org/10.2307/2708278

Boake, C. (2002). From the Binet-Simon to the Wechsler-Bellevue: Tracing the history of intelligence testing. *Journal of Clinical and Experimental Neuropsychology, 24*(3), 383–405. https://doi.org/10.1076/jcen.24.3.383.981

Bock, R. D. (1997). A brief history of item response theory. *Educational Measurement: Issues and Practice, 16*(4), 21–33. https://doi.org/10.1111/j.1745-3992.1997.tb00605.x

Bock, R. D., & Jones, L. V. (1968). *The measurement and prediction of judgment and choice*. Holden-Day.

Böckenholt, U. (2004). Comparative judgments as an alternative to ratings: Identifying the scale origin. *Psychological Methods, 9*(4), 453–465. https://doi.org/10.1037/1082-989X.9.4.453

Bolanowski, S. J., & Gescheider, G. A. (1991). *Ratio scaling of psychological magnitude: In honor of the memory of S.S. Stevens*. Lawrence Erlbaum.

Boring, E. G. (1917). A chart of the psychometric function. *The American Journal of Psychology, 28*(4), 465–470. https://doi.org/10.2307/1413891

Boring, E. G. (1961). The beginning and growth of measurement in psychology. *ISIS, 52*(2), 238–257. https://doi.org/10.1086/349471

Bowman, M. L. (1989). Testing individual differences in ancient China. *American Psychologist, 44*(3), 576–578. https://doi.org/10.1037/0003-066X.44.3.576.b

Brown, S. R. (1991). William Stephenson (1902–1989): Obituary. *American Psychologist, 46*(3), 244. https://doi.org/10.1037/0003-066X.46.3.244

Brown, S. R. (1993). A primer on Q-methodology. *Operant Subjectivity, 16*(3/4), 91–138. https://doi.org/10.15133/j.os.1993.002

Brown, W., & Thomson, G. H. (1940). *The essentials of mental measurement*. Cambridge University Press.

Camic, C., & Xie, Y. (1994). The statistical turn in American Social Science: Columbia University, 1890 to 1915. *American Sociological Review, 59*(5), 773–805. https://doi.org/10.2307/2096447

Cattell, J. M. K. (1886). Psychometrische Untersuchungen, I. Apparate und Methoden [Psychometric Studies, I. Equipment and Methods]. *Philosophische Studien, 3*, 305–335.

Cattell, J. M. K. (1906a). A statistical study of American men of Science: The selection of a group of one thousand scientific men. *Science, 24*(621), 658–665. https://doi.org/10.1126/science.24.621.658

Cattell, J. M. K. (1906b). A statistical study of American men of science: II. The measurement of scientific merit. *Science, 24*(622), 699–707. https://www.jstor.org/stable/1634085

Cattell, J. M. K. (1910). A further statistical study of American men of science. *Science, 32*(827), 633–648. https://www.jstor.org/stable/1635729

Cattell, J. M. K. (1921). In memory of Wilhelm Wundt, by his American students, section II. *Psychological Review, 28*(3), 155–159. https://doi.org/10.1037/h0073437

Cattell, R. B. (1952). The three basic factor-analytic research designs—Their interrelations and derivatives. *Psychological Bulletin, 49*(5), 499–520. https://doi.org/10.1037/h0054245

Cattell, J. M. K., & Galton, F. (1890). Mental tests and measurements. *Mind, 15*(59), 373–381. https://www.jstor.org/stable/i339158

Cowan, R. S. (1970). *Introduction to the second edition of Galton (1874)*. F. Cass.

De Candolle, A. (1873). *Histoire des Sciences et des Savants depuis Deux Siècles: Suivie D’Autres Études sur les Sujets Scientifiques en Particulier sur la Sélection dans L’Espèce Humaine* [History of science and scholars for two centuries: Followed by other studies on scientific subjects, in particular on selection in the Human species]. H. Georg, Libraire-Editeur.

Dehue, T. (1997). Deception, efficiency, and random groups: Psychology and the gradual origination of the random group design. *ISIS, 88*(4), 653–673. https://doi.org/10.1086/383850

Doyle, K. O. (1974). Theory and practice of ability testing in ancient Greece. *Journal of the History of the Behavioral Sciences, 10*(2), 202–212. https://doi.org/10.1002/1520-6696(197404)10:2<202::AID-JHBS2300100208>3.0.CO;2-Q

Drechsler, W. (1997). Christian Wolff (1679–1754): A biographical essay. *European Journal of Law and Economics, 4*(4), 111–128. https://doi.org/10.1023/A:1008682025945

Drenth, P. J. D., & Sijtsma, K. (1990). *Testtheorie: Inleiding in de theorie van de psychologische test en zijn toepassingen* [Test theory. Introduction to the theory of the psychological test and its applications]. Bohn Stafleu Van Loghum.

Dzhafarov, E. N., & Colonius, H. (2011). The Fechnerian idea. *American Journal of Psychology, 124*(2), 127–140. https://doi.org/10.5406/amerjpsyc.124.2.0127

Edgeworth, F. Y. (1888). The statistics of examinations. *Journal of the Royal Statistical Society, 51*(3), 599–635. https://www.jstor.org/stable/2339898

Edgeworth, F. Y. (1890). The element of chance in competitive examinations. *Journal of the Royal Statistical Society, 53*(4), 599–635. https://www.jstor.org/stable/2979547

Euclid, (1956). *The Thirteen Books of Euclid's Elements, Vol. 2 (Books III-IX)* (L. Thomas Heath, Trans. & Ed.). Dover Publications.

Fechner, G. T. (1831). *Maßbestimmungen über die Galvanische Kette* [Dimensional determinations via the galvanic chain]. Brockhaus.

Fechner, G. T. (1859). Über ein wichtiges psychophysisches Grundgesetz und dessen Beziehung zur Schätzung der Sterngrössen [On an important psychophysical fundamental law and its relationship to the estimation of star sizes]. *Abhandlungen der Königliche Sächsische Gesellschaft der Wissenschaften, 6*(4), 455–532.

Fechner, G. T. (1860). *Elements of psychophysics*, volume I (Helmut E. Adler, Trans.). Holt, Rinehart and Winston, 1966.

Fechner, G. T. (1871). Zur experimentellen Ästhetik [On experimental aesthetics]. *Abhandlungen der Königliche Sächsische Gesellschaft der Wissenschaften, Mathematisch-physische Klasse, 9*(1), 555–635.

Fechner, G. T. (1882). *Revision der Hauptpunkte der Psychophysik* [Revision of the Main Points of Psychophysics]. Breikopf und Härtel.

Fechner, G. T. (1897). *Kollectivmasslehre* [*Collective Mass Doctrine*] (A. F. Lipps, Ed.). Engelmann.

Feuerhahn, W. (2004). Die Wolff'sche Psychometrie [Wolffian Psychometrics]. In O.-P. Rudolph & J.-F. Goubet (Eds.), *Die Psychologie Christian Wolffs: Systematische und historische Untersuchungen* (pp. 227–236). Max Niemeyer Verlag. https://doi.org/10.1515/9783110932317.227

Furr, R. M., & Bacharach, V. R. (2008). *Psychometrics: An introduction*. Sage.

Galton, F. (1869). *Hereditary genius: An inquiry into its laws and consequences*. MacMillan.

Galton, F. (1874). *English men of science: Their nature and nurture*. MacMillan.

Galton, F. (1885). On the anthropometric laboratory at the late international health exhibition. *The Journal of the Anthropological Institute of Great Britain and Ireland, 14*(1885), 205–221. https://www.jstor.org/stable/2841978

Godin, B. (2006). On the origins of bibliometrics. *Scientometrics, 68*(1), 109–133. https://doi.org/10.1007/s11192-006-0086-0

Godin, B. (2007). From eugenics to scientometrics: Galton, Cattell, and men of science. *Social Studies of Science, 37*(5), 691–728. https://doi.org/10.1177/0306312706075338

Green, C. D. (1996). All that glitters: A review of psychological research on the aesthetics of the golden section. *Perception, 24*(8), 937–968. https://doi.org/10.1068/p240937

Guilford, J. P. (1936). *Psychometric methods*. McGraw-Hill.

Hall, N. S. (2007). R.A. Fisher and his advocacy of randomization. *Journal of the History of Biology, 40*(2), 295–325. https://doi.org/10.1007/s10739-006-9119-z

Heiser, W. J. (2008). Psychometric roots of multidimensional data analysis in the Netherlands: From Gerard Heymans to John van de Geer. *Electronic Journal for History of Probability and Statistics, 4*(2), 1–25. https://www.jehps.net/Decembre2008/Heiser.pdf

Heiser, W. J., & De Leeuw, J. (1981). Multidimensional mapping of preference data. *Mathématiques et Sciences Humaines, 19*(1981), 39–96. http://www.numdam.org/item/MSH_1981__73__39_0/

Heymans, G., & Wiersma, E. (1906). Beiträge zur speziellen Psychologie auf Grund einer Massenuntersuchung [Contributions to differential psychology on the basis of a mass study]. *Zeitschrift für Psychologie, 42*(81–127), 258–301.

Hock, H. S., & Schöner, G. (2010). Measuring perceptual hysteresis with the modified method of limits: Dynamics at the threshold. *Seeing and Perceiving, 23*(2), 173–195. https://doi.org/10.1163/187847510X503597

Hornstein, G. A. (1988). Quantifying psychological phenomena: Debates, dilemmas, and implications. In J. G. Morawski (Ed.), *The rise of experimentation in American psychology* (pp. 1–34). Yale University Press.

Hunter, I. (2000). Christian Thomasius and the desacralization of philosophy. *Journal of the History of Ideas, 61*(4), 595–616. https://www.jstor.org/stable/3654071

Jones, L. V., & Thissen, D. (2007). A history and overview of psychometrics. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 1–27). Elsevier. https://doi.org/10.1016/S0169-7161(06)26001-2

Kelley, T. L. (1924). *Statistical method*. The Macmillan Company.

Kouwer, B. J. (1963). *Het spel van de persoonlijkheid: Theorieën en systemen in de psychologie van de menselijke persoon* [The game of personality: Theories and systems in the psychology of the human person]. Erven J. Bijleveld.

Leary, D. E. (1980). The historical foundation of Herbart's mathematization of psychology. *Journal of the History of the Behavioral Sciences, 16*(2), 150–163. https://doi.org/10.1002/1520-6696(198004)16:2<150::AID-JHBS2300160206>3.0.CO;2-1

Link, S. W. (1994). Rediscovering the past: Gustav Fechner and signal detection theory. *Psychological Science, 5*(6), 335–340. https://doi.org/10.1111/j.1467-9280.1994.tb00282.x

Lord, F. (1952). *A theory of test scores* (Psychometric Monograph, Number 7). Psychometric Corporation.

Luce, R. D. (1959). On the possible psychophysical laws. *Psychological Review, 66*(2), 81–95. https://doi.org/10.1037/h0043178

Lumsden, J. (1980). Variations on a theme by Thurstone. *Applied Psychological Measurement, 4*(1), 1–7. https://doi.org/10.1177/014662168000400101

Maydeu-Olivares, A. (2001). Limited information estimation and testing of Thurstonian models for paired comparison data under multiple judgment sampling. *Psychometrika, 66*(2), 209–227. https://doi.org/10.1007/BF02294836

McKeown, B., & Thomas, D. B. (2013). *Q-Methodology* (2nd ed.). Sage.

McReynolds, P., & Ludwig, K. (1984). Christian Thomasius and the origin of psychological rating scales. *ISIS, 75*(3), 546–553. https://doi.org/10.1086/353573

McReynolds, P., & Ludwig, K. (1987). On the history of rating scales. *Personality and Individual Differences, 8*(2), 281–283. https://doi.org/10.1016/0191-8869(87)90188-7

Mei, M. (2021). Wolff's idea of psychometria. In S. de Freita Araujo, T. C. R. Rereira, & T. Sturm (Eds.), *The force of an idea, studies in history and philosophy of science* (Vol. 50, pp. 89–103). https://doi.org/10.1007/978-3-030-74435-9_6

Mellenbergh, G. J. (2011). *A conceptual introduction to psychometrics*. Eleven International Publishing.

Michell, J. (2006). Psychophysics, intensive magnitudes, and the psychometrians' fallacy. *Studies in the History and Philosophy of Biological and Biomedical Sciences, 17*(3), 414–432. https://doi.org/10.1016/j.shpsc.2006.06.011

Miller, G. A. (1964). *Mathematics and psychology*. Wiley.

Mülberger, A. (2017). Mental association: Testing individual differences before Binet. *Journal of the History of the Behavioral Sciences, 53*(2), 176–198. https://doi.org/10.1002/jhbs.21850

Müller, G. E. (1878). *Zur Grundlegung der Psychophysik, Kritische Beiträge* [On the foundation of psychophysics, critical contributions]. Grieben.

Müller, G. E. (1879). Über die Maßbestimmungen des Ortsinnes der Haut mittels der Methode der richtigen und falschen Fälle [On measuring the spatial sense of the skin by means of the method of right and wrong cases]. *Archiv für die gesamte Physiologie des Menschen und der Tiere, 19*, 191–235. https://doi.org/10.1007/BF01639850

Murray, D. J. (2021). *The creation of scientific psychology* (S.W. Link, Ed.) (1st ed.). Routledge. https://doi.org/10.4324/9781315620985

Peirce, C. S., & Jastrow, J. (1885). On small differences in sensation. *Memoirs of the National Academy of Sciences for 1984*, Vol. III, 5th Memoir, 75–83.

Pratt, C. C. (1928). Bisection of tonal intervals larger than the octave. *Journal of Experimental Psychology, 11*(2), 17–26. https://doi.org/10.1037/h0075337

Ramul, K. (1960). The problem of measurement in the psychology of the eighteenth century. *American Psychologist, 15*(4), 256–265. https://doi.org/10.1037/h0047753

Ramul, K. (1963). Some early measurements and ratings in psychology. *American Psychologist, 18*(10), 653–659. https://doi.org/10.1037/h0040858

Rozeboom, W. W. (1962). The untenability of Luce's principle. *Psychological Review, 69*(6), 542–547. https://doi.org/10.1037/h0041419

Saffir, M. (1937). A comparative study of scales constructed by three psychophysical methods. *Psychometrika, 2*(3), 179–198. https://doi.org/10.1007/BF02288395

Sanford, E. C. (1898). *A course in experimental psychology: Part I, Sensation and perception*. Heath.

Schmiedek, F., Oberauer, K., Wilhelm, O., Süß, H. M., & Wittmann, W. W. (2007). Individual differences in components of reaction time distributions and their relations to working memory and intelligence. *Journal of Experimental Psychology: General, 136*(3), 414. https://doi.org/10.1037/0096-3445.136.3.414

Sheynin, O. (2004). Fechner as a statistician. *British Journal of Mathematical and Statistical Psychology, 57*(1), 53–72. https://doi.org/10.1348/000711004849196

Sijtsma, K. (2012). Psychological measurement between physics and statistics. *Theory & Psychology, 22*(6), 786–809. https://doi.org/10.1177/0959354312454353

Sijtsma, K. (2016). Invited discussion of Cronbach (1951 Coefficient alpha and the internal structure of tests). *Psychometrika, 81*(4), 1205–1208. https://doi.org/10.1007/s11336-016-9540-y

Sijtsma, K., & Junker, B. W. (2006). Item response theory: Past performance, present developments, and future expectations. *Behaviormetrika, 33*(1), 75–102. https://doi.org/10.2333/bhmk.33.75

Simonton, D. K. (2014). The personal characteristics of political leaders: Quantitative multiple-case assessments. In G. R. Goethals, S. T. Allison, R. M. Kramer, & D. M. Messick (Eds.), *Conceptions of Leadership* (Jepson Studies in Leadership) (pp. 53–69). Palgrave MacMillan. https://doi.org/10.1057/9781137472038_4

Spearman, C. (1904). General Intelligence, objectively determined and measured. *American Journal of Psychology, 15*(2), 201–293. https://www.jstor.org/stable/1412107

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*(3), 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

Stephenson, W. (1936). The foundations of psychometry: Four factor systems. *Psychometrika, 1*(3), 195–209. https://doi.org/10.1007/BF02288366

Stephenson, W. (1953). *The Study of Behavior: Q-Technique and its Methodology*. University of Chicago Press.

Stevens, S. S. (1961). To honor Fechner and repeal his law. *Science, 133*(3446), 80–86. https://www.jstor.org/stable/1706724

Stigler, S. M. (1978). Mathematical statistics in the early States. *The Annals of Statistics, 6*(2), 239–265. https://www.jstor.org/stable/2958876

Stigler, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900*. The Belknap Press of Harvard University Press.

Stigler, S. M. (1992). A historical view of statistical concepts in psychology and educational research. *American Journal of Education, 101*(1), 60–70. https://doi.org/10.1086/444032

Stigler, S. M. (2010). Darwin, Galton and the statistical enlightenment. *Journal of the Royal Statistical Society, Series A, 173*(3), 469–482. https://doi.org/10.1111/j.1467-985X.2010.00643.x

Sturm, T. (2006). Is there a problem with mathematical psychology in the eighteenth century? A fresh look at Kant's old argument. *Journal of the History of the Behavioral Sciences, 42*(4), 353–377. https://doi.org/10.1002/jhbs.20191

Szabó, A. T. (1985). Alphonse de Candolle's early scientometrics (1883, 1885) with references to recent trends in the field (1978–1983). *Scientometrics, 8*(1–2), 13–33. https://doi.org/10.1007/bf02025219

Takane, Y. (1989). Analysis of covariance structures and probabilistic binary choice data. In G. De Soete, H. Feger, & K. C. Klauer (Eds.), *New developments in psychological choice modeling* (pp. 139–160). North-Holland. http://takane.brinkster.net/Yoshio/p027.pdf

Thomasius, C. (1692a). *Die neue Erfindung einer wohlbegründeten und für das gemeine Wesen höchstnöthigen Wissenschaft das Verborgene des Herzens anderer Menschen auch wider ihren Willen aus der täglichen Conversation zu erkennen* [New Discovery of a Solid Science, Most Necessary for the Community for Discerning the Secrets of the Heart of Other Men from Daily Conversation, Even Against Their Will]. Christoph Salfeld.

Thomasius, C. (1692b). *Weitere Erleuterung durch unterschiedene Exempel des ohnelängst gethanen Vorschlags wegen der neuen Wissenschaft anderer Menschen Gemühter erkennen zu lernen* [Further Elucidation by Different Examples of the Recent Proposal for a New Science for Discerning the Nature of Other Men's Minds]. Christoph Salfeld.

Thomson, G. H. (1919). A direct deduction of the constant process used in the method of right and wrong cases. *Psychological Review, 26*(6), 454–464. https://doi.org/10.1037/h0070741

Thorndike, E. L. (1910). Handwriting. Part I. The measurement of the quality of handwriting: Criticisms of the scale. *Teachers College Record, 11*(2), 8–46.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *The Journal of Educational Psychology, 16*(7), 433–451. https://doi.org/10.1037/h0073357

Thurstone, L. L. (1927a). Psychophysical analysis. *The American Journal of Psychology, 38*(3), 368–389. https://www.jstor.org/stable/1415006

Thurstone, L. L. (1927b). A law of comparative judgment. *Psychological Review, 34*(4), 273–286. https://doi.org/10.1037/h0070288

Thurstone, L. L. (1929). Fechner's law and the method of equal appearing intervals. *Journal of Experimental Psychology, 12*(3), 221–238. https://doi.org/10.1037/h0070968

Titchener, E. B. (1905). *Experimental psychology: A manual of laboratory practice, Vol. II, Instructor's Manual*. MacMillan.

Torgerson, W. S. (1958). *Theory and methods of scaling*. Wiley.

Urban, F. M. (1907). On the method of just perceptible differences. *Psychological Review, 14*(4), 244–253. https://doi.org/10.1037/h0073288

Urban, F. M. (1908). *The application of statistical methods to the problems of psychophysics*. Psychological Clinic Press.

Urban, F. M. (1910). The method of constant stimuli and its generalizations. *Psychological Review, 17*(4), 229–259. https://doi.org/10.1037/h0074515

Urban, F. M. (1933). The Weber-Fechner law and mental measurement. *Journal of Experimental Psychology, 16*(2), 221–238. https://doi.org/10.1037/h0070805

Van der Heijden, P. G. M., & Sijtsma, K. (1996). Fifty years of measurements and scaling in the Dutch social sciences. *Statistica Neerlandica, 50*(1), 111–135. https://doi.org/10.1111/j.1467-9574.1996.tb01483.x

Vidal, F. (2011). *The sciences of the soul: The early modern origins of psychology*. University of Chicago Press.

Von Mises, R. (1912). Über die Grundbegriffe der Kollectivmasslehre [About the basic concepts of the collective mass doctrine]. *Jahresbericht der Deutschen Mathematiker- Vereinigung, 21*(1), 9–20. http://eudml.org/doc/145318

Von Plato, J. (1994). *Creating modern probability: Its mathematics, physics and philosophy in historical perspective*. Cambridge University Press.

Walker, H. M. (1929). *Studies in the history of statistical method, with special reference to certain educational problems*. The William & Wilkins Company.

Webster, D. S. (1985). James McKeen Cattell and the invention of academic quality ratings, 1903–1910. *The Review of Higher Education, 8*(2), 107–121. https://doi.org/10.1353/rhe.1985.0023

Wilhelm, O., Herzmann, G., Kunina, O., Danthiir, V., Schacht, A., & Sommer, W. (2010). Individual differences in perceiving and recognizing faces—One element of social cognition. *Journal of Personality and Social Psychology, 99*(3), 530–548. https://doi.org/10.1037/a0019972

Wilhelm, O., Hildebrandt, A. H., & Oberauer, K. (2013). What is working memory capacity, and how can we measure it? *Frontiers in Psychology, 4*(1), 433. https://doi.org/10.3389/fpsyg.2013.00433

Wissler, C. (1901). The correlation of mental and physical tests. *Psychological Review: Monograph Supplements, 3*(6), *i*–62. https://doi.org/10.1037/h0092995

Wolff, C. (1962a). Psychologia empirica, methodo scientifica pertractata [Empirical psychology, treated according to the scientific method]. In J. École (Ed.), *Christian Wolff: Vol. 5.Gesammelte Werke* (Psychologia empirica), Series II. Olms (Original work published 1732).

Wolff, C. (1962b). Philosophia prima, sive Ontologia, methodo scientifica pertractata, qua omnis cognitionis humanae principia continentur [First philosophy, or Ontology, treated according to the scientific method, in which all the principles of human cognition are contained]. In J. École (Ed.), *Christian Wolff: Vol. 3. Gesammelte Werke* (Ontologia), Series II. Olms (Original work published 1736).

Wolff, C. (1963). *Preliminary discourse on philosophy in general* (R. J. Blackwell, Trans.). Bobbs-Merrill (Original work published 1728).

Wright, W. F. (2016). Early evolution of the thermometer and application to clinical medicine. *Journal of Thermal Biology, 56*(1), 18–30. https://doi.org/10.1016/j.jtherbio.2015.12.003

Zhang, H. (1988). Psychological measurement in China. *International Journal of Psychology, 23*(1), 101–117. https://doi.org/10.1080/00207598808247755

Zudini, V. (2011). The Euclidean model of measurement in Fechner's psychophysics. *Journal of the History of the Behavioral Sciences, 47*(1), 70–87. https://doi.org/10.1002/jhbs.20472