On measurement, statistics, psychology, and lifelong learning

Sijtsma, K.

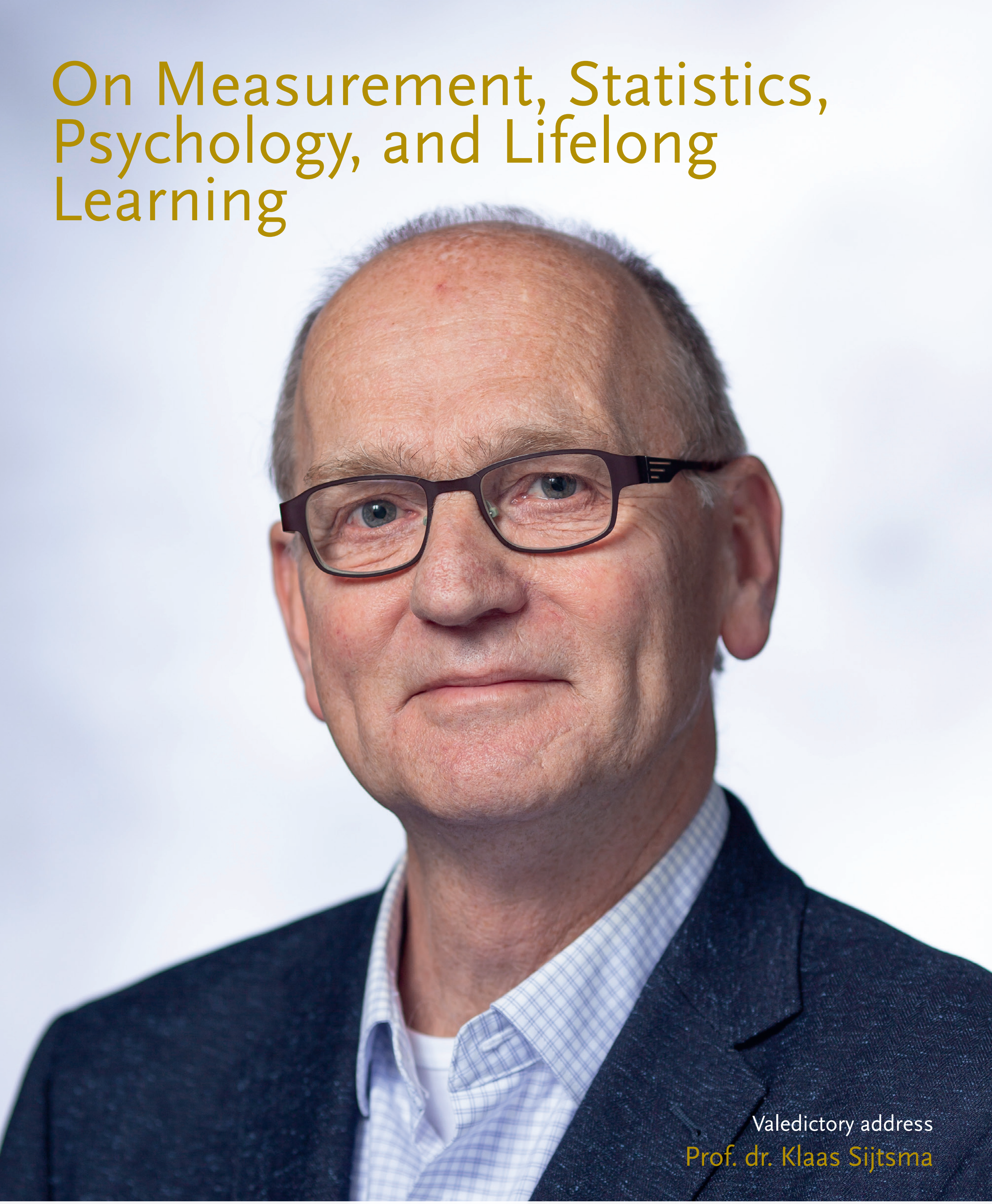
Publication date:
2022

Document Version
Publisher's PDF, also known as Version of record

Link to publication in Tilburg University Research Portal

Citation for published version (APA):
Sijtsma, K. (2022). *On measurement, statistics, psychology, and lifelong learning*.

On Measurement, Statistics, Psychology, and Lifelong Learning



Valedictory address

Prof. dr. Klaas Sijtsma

Tilburg University

Understanding Society

Klaas Sijtsma (1955) studied psychology from 1975 to 1982 at the University of Groningen. He graduated in personality psychology and statistics and measurement theory with a minor in mathematical methods. From 1981 to 1983, he was affiliated as a methodologist with the "Research Instituut voor het Onderwijs in het Noorden" (RION; Research Institute for Education in Northern Netherlands), which was associated with the University of Groningen. From 1983 to 1993, he was an assistant professor in psychometrics at the Vrije Universiteit Amsterdam, department of Work and Organizational Psychology of the Faculty of Psychology and Educational Sciences. He defended his PhD dissertation titled Contributions to Mokken's Nonparametric Item Response Theory, supervised by Ivo Molenaar (University of Groningen) and Pieter Drenth (Vrije Universiteit Amsterdam) in 1988 at the University of Groningen, his alma mater. From 1993 to 1997, Sijtsma was affiliated as associate professor with the department of Methodology and Statistics of the Faculty of Social Sciences at Utrecht University. In 1997, he was appointed full professor of Methods and Techniques of Psychological Research at Tilburg University. He was a member of the department of Methods and Techniques of Research, later the department of Methodology and Statistics of the Tilburg School of Social and Behavioral Sciences.

Sijtsma taught several courses in methodology, statistics, and psychometrics to graduate students and to PhD students and senior researchers, at home and abroad. His research focuses at psychometrics with an emphasis on statistical measurement models for psychological attributes. He has published over 200 articles and book chapters mostly in international journals and with renowned publishers, often with colleagues including many PhD students. In addition, he co-authored monographs on psychological test theory (1990, 2006; with Pieter Drenth; in Dutch), nonparametric item response theory (2002; with Ivo Molenaar; in English), and common measurement models for psychological attributes (2021; with Andries van der Ark, University of Amsterdam; in English). The data fraud affair concerning Diederik Stapel inspired Sijtsma to study the use of statistics as a source of questionable research practices. Currently, he is preparing a monograph on this topic. Sijtsma supervised many PhD students, of which 34 successfully defended their dissertation. Currently, a small number of PhD students is in the process of preparing their dissertation.

In addition to education and research, he was active as an administrator and a manager, both inside and outside the university. Two examples of the latter category are the membership of the Supervisory Board of Cito Arnhem (1998—2003, 2009—2020) and the chairmanship of the Dutch Committee on Tests and Testing (COTAN) of the Dutch Institute of Psychologists (NIP; 2005-2010). He was a member of the Management Team of the Faculty of Psychology and Educational Sciences at the Vrije Universiteit Amsterdam (1991—1993) and President of the Psychometric Society, the international learned society on psychometrics (2010—2011). At Tilburg University, Sijtsma was head of the department of Methodology and Statistics (1999—2010), dean of the Tilburg School of Social and Behavioral Sciences (2011—2017), and rector magnificus (2019—2020). He was recently appointed co-chair of the Committee on Research Integrity at Erasmus University Rotterdam.

On Measurement, Statistics, Psychology, and Lifelong Learning

Prof. Dr. Klaas Sijtsma

Valedictory address
delivered on occasion of his official retirement, May 20th 2022

ISBN: 978-94-6167-472-2

www.tilburguniversity.edu

On Measurement, Statistics, Psychology, and Lifelong Learning[1]

[1] I thank Jaap Denissen, Wilco Emons, Rob Meijer and Hester Sijtsma for their suggestions for improving the text.

1 A Short Personal History

I admit, my lecture has a rather ambitious title suggesting that I am grasping at straws to make my point. Perhaps I am, but instead of looking back and trying to have things my way, I will try to look into the future, ask some questions, and offer some suggestions. I hope you will feel good enough afterwards to enjoy a drink and alarmed enough to engage in discussion. I start with a short personal history.

Central in my scientific work has been a fascination with measurement. This fascination stems from my years as a student at the University of Groningen. I started there in 1974 as a student of pharmacy. This is the science of medical drugs and therapy. It is a mixture of many disciplines: mathematics, physics, biology, medicine, but most of all, chemistry. I am proud to say that I succeeded all the exams until May of the next year, when I decided to quit and find another study to pursue. To summarize my motives for quitting, I disliked working in a chemistry lab for most of the week and had not thought this over well enough when I chose to study pharmacy. As an 18-year-old, I had no idea about the real world and what was happening there outside school, football, pop music, and having fun with my friends.

However, I never looked back on pharmacy disappointed. One thing pharmacy taught me was working hard to accomplish something you want, and the other was that I liked science, I mean exact science, but it took me a few more years to find out. Before I did, I entered the psychology program in September 1975. Not unexpectedly, I liked especially the exact courses, like psychonomics, experimental social psychology, and methodology and statistics, but also the philosophy of science. The psychology program in Groningen was a six-year program of which six months had to be spent at another school. I chose the minor in mathematics and soon found myself in a course on linear algebra, which I liked.

I cannot know whether I would have become so deeply involved in measurement if pharmacy had not taught me the importance of it. Most of the time I spent in the chemistry lab, I did measurement but without realizing it. I learned there are two kinds of measurement. In Figure 1, the photo on the left, you see an example of *quantitative analysis*, which is determining the concentration of a chemical element in a liquid. The person in the photo is adding drops, half drops and even quarter drops until a drastic color change occurs indicating saturation, and from the collected data she will compute the concentration of the element of interest. In the photo on the right, you see a wire dipped in a residue held in a Bunsen burner producing a flame whose color identifies a metal. For example, the green flame suggests the presence of copper. This is *qualitative analysis*, directed at identifying chemical elements in a mixture; that is, categorizing by type. In the human sciences, nowadays we also say we are dealing with quantitative measurement versus qualitative measurement, determining an amount or a type.

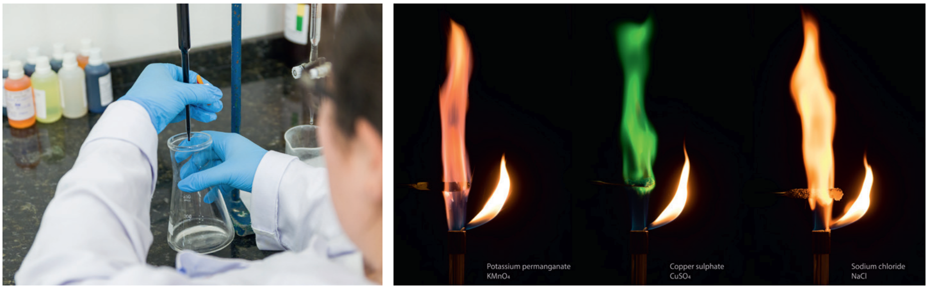
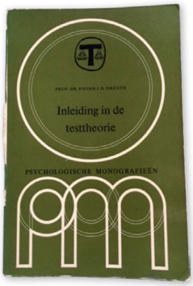


Figure 1: Quantitative analysis (left) and qualitative analysis (right).

It is imperative that you understand that the chemical analyses must be done with the greatest care and caution. In the quantitative analyses, contaminated or polluted mixtures may suggest the wrong concentration. In the qualitative analysis, a series of experiments must be done in a fixed order, and each inaccuracy may lead to the hiding of copper atoms in larger molecules, rendering it impossible later to identify the copper atoms; the flame will not turn green. Here and everywhere in science, ***standardization*** of the measurement procedure is key; without it, results are biased or meaningless. Now, why does a healthy young man become interested in such a dull topic as measurement?


PROF. DR. PIETER J. D. DRENTH
Inleiding in de testtheorie
PSYCHOLOGISCHE MONOGRAFIEËN

In my second year in psychology, I followed a compulsory course on psychological test theory. I had to study the book “Inleiding in de Testtheorie” (Introduction to Test Theory), written by Pieter Drenth (Drenth, 1975), then a professor at the Vrije Universiteit Amsterdam. The book was about the measurement of psychological attributes like intelligence and subattributes like spatial orientation, transitive reasoning, and word fluency, and personality traits like dominance, extraversion, and neuroticism. I was amazed to find that psychologists measured attributes, just like chemists did. And I was also fascinated by how they did this, using lists of problems one had to solve or questions one had to answer, and analyzing counts of the number-correct or counts of credit points earned. This looked very different from the chemical lab work I had done to determine amounts and types, and the computations needed to determine the former. I learned that psychological tests, which are the measurement instruments for psychological attributes, were used to select students for education, applicants for jobs, and patients for therapy, but also to assess education level, job performance, and therapy progress (e.g., Niessen & Meijer, 2017).

Figure 2 shows a task from a test measuring spatial orientation, an aspect of general intelligence. A typical test consists of several such tasks, often a few dozen, where each task represents a small measurement device. The person to which the task is presented is asked to respond. The response is correct or incorrect, and together the responses determine her measurement value for spatial orientation. There are two reasons why the test uses several tasks.

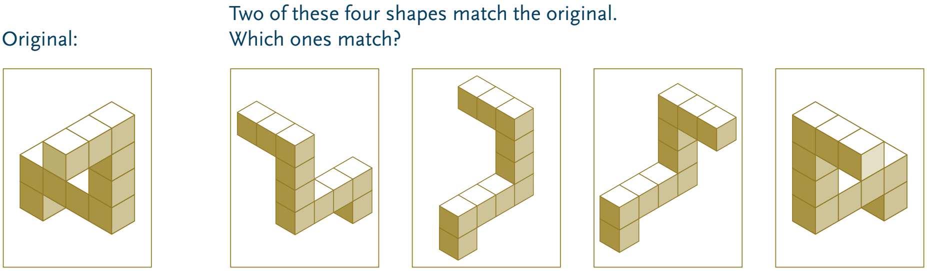


Figure 2: Example of a task for measuring spatial orientation.

First, different tasks are not fully interchangeable: Each task has peculiarities not shared with other tasks, and together the set of tasks better capture the attribute than a single task does. This difficult issue refers to the theory of the attribute and is known as the construct validity problem (Cronbach & Meehl, 1955; Markus & Borsboom, 2013). Construct validity is difficult to establish, often neglected but extremely important.

Second, responses to single tasks are unreliable in the sense that you cannot trust a person to give the same response again if you could retest this person under precisely the same conditions, assuming she did not recall the previous administration. We assume variation occurs at random. This is the reliability issue (Emons, Sijtsma, & Meijer, 2007; Lord & Novick, 1968). Several tasks together reduce the influence of randomness to a high degree and produce a reliable measurement value. Compare this to recording blood pressure repeatedly rather than once.

This is how I became interested in measurement. If my memory isn't accurate, at least it is a good story.

2 Measurement

I specialized in the discipline of psychometrics. Psychometrics studies the mathematical and statistical requirements data collected by means of a set of tasks must satisfy to conclude that we have measurement values; that is, values you can locate on a scale. To make the enterprise of scale construction successful, I assume a psychological theory of the attribute to be measured lay at the foundation of data collection by means of the test. Remember the answer to a task could be right or wrong. We can replace these answers by numbers 1 for a correct answer and 0 for an incorrect answer. Given that in the example task two answers were correct, we could also use scores 0 for both answers incorrect, 1 for one correct answer and 2 for two correct answers. Notice this is the beginning of quantification, replacing answers by numbers, and the numbers can be fed to a mathematical or a statistical measurement model. Table 1 is filled with zeroes and ones. Each line represents the scores for one person and each column represents the scores for one task. The first column shows the person numbers, which have no other function than to identify different persons. The numbers at the top of the columns show the item numbers or identifiers.

Table 1: Binary item scores, 0 (incorrect response) and 1 (correct response).

Person No.	Item No.										
	1	2	3	4	5	6	. . .	14	15	. .	25
1	0	0	1	0	1	1	. . .	1	1	. .	0
2	0	1	1	1	0	1	. . .	1	1	. .	1
3	1	1	0	1	1	1	. . .	1	0	. .	1
4	0	1	1	1	0	1	. . .	1	1	. .	1
5	0	0	1	1	1	1	. . .	0	1	. .	1
6	1	1	1	0	1	0	. . .	1	0	. .	1
.											
.											
124	0	0	0	1	0	1	. . .	0	0	. .	0
125	1	1	1	1	1	1	. . .	1	0	. .	1
.											
.											
364	0	1	1	0	1	1	. . .	1	1	. .	1

So, the numbers are fed to a measurement model and can be used to do computations. What is a measurement model? Measurement models are mathematical models that are defined by *assumptions* about the way people respond to tasks in a psychological test. The models define the requirements for measurement, but they are not theories about spatial orientation or other psychological attributes. From the models we can derive that test scores such as the number of correct answers to a set of spatial orientation tasks are located along one mathematical dimension, which can then serve as a scale, comparable with a yardstick or a thermometer but also a little different. That is, on the scale of a thermometer, the difference between 20 degrees Celsius and 15 degrees Celsius equals the difference between 15 degrees and 10 degrees, or between 17 and 12 degrees. This follows from the theory of temperature, which is at the basis of the construction of a thermometer. This is different in psychological measurement, not because we chose so but due to the state of knowledge, which is not as far as it is in physics and chemistry.

As a result, in psychology, the distances between the different scale locations do not have an obvious meaning. For example, if John had 20 tasks correct, Mary 15 tasks, and Roger 10 tasks, you cannot say that the difference in spatial orientation between John and Roger is twice that between Mary and Roger. The reason is that we do not know enough about how cognitive processes produce responses to spatial orientation and other tasks. That is, what we know about spatial orientation is not enough to derive tasks that provide exact information about this attribute. This is different from physics. Physics simply is ahead of psychology, but you would be surprised how much the measurement of temperature looked like psychological measurement a few centuries ago (Sherry, 2011). So, there is hope for psychological measurement!

What do we know about the scale values for psychological measurement? For a better understanding, let us have a look at two assumptions of measurement models:

- The first assumption is about the complexity of the measurement; that is, the number of attributes that influence the measurement. This is the assumption of the dimensionality of the data. For example, Figure 3 shows a task measuring transitive reasoning, a kind of logical reasoning, which also requires language skills. Consequently, the data—the table with the zeroes and the ones—are probably two-dimensional. Notice that the second dimension representing language skills is unwanted if you intend to measure transitive reasoning only. This multidimensionality, which can be represented mathematically using a measurement model, is typical of psychological measurement. Measuring only one attribute is problematic and confounding with other attributes is almost unavoidable.

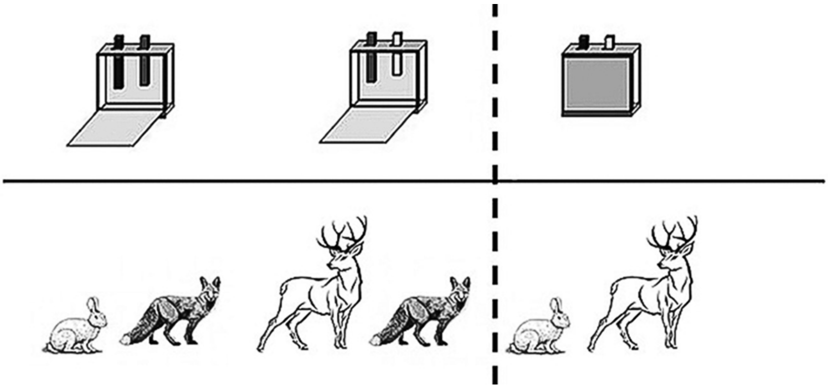


Figure 3[2]: Examples of transitive reasoning tasks for length (upper panel) and age (lower panel), showing two premises (left) and one inference task (right). Artwork by Samantha Bouwmeester.

- The second assumption concerns the relation of responses to problems in, for example, a transitive reasoning test with each of the dimensions. This relationship is called the task's response function, which usually is monotone. This reflects the idea that a higher position on the dimension—the scale—

[2] Reprinted with permission from the publisher.

implies a higher probability of giving the correct answer. Response functions can vary in different ways reflecting various properties of tasks presented to people, but I will leave this topic at rest.

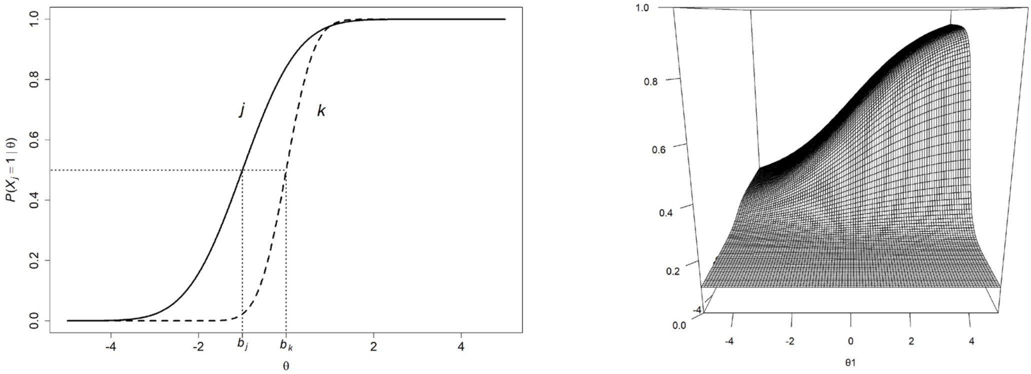


Figure 4[3]: Left: Two normal-ogive IRFs. For the solid IRF a_j=1, and b_j=−1; for the dashed IRF a_k=2, and b_k=0. Right: Item response surface of the two-dimensional 3-parameter logistic model with α_{j1}=0,5, α_{j2}=2, δ_j=0, and γ_j=.10, plotted in 3D perspective.

These assumptions and other assumptions together define a model, which mathematically implies one or more scales on which persons can be ***ordered*** by means of the total number of correctly solved tasks. Note that I say ordered, to emphasize that the distances between the scores cannot be compared as they can on the temperature scale. That is, John is better in transitive reasoning than Mary and Roger, and Mary is better than Roger. Sometimes, measurement takes the form of classification (e.g., Junker & Sijtsma, 2001; Magidson & Vermunt, 2004), comparable to qualitative analysis in chemistry.

How do you know that a measurement model is applicable to the measurement of a psychological attribute? The trick is that you can derive from the model how the data must look like to have a scale for measuring a psychological attribute. Only if the data predicted by the model and the real data collected with real people agree, can we say we have a scale. So-called goodness-of-fit research determines the degree of agreement and is key in data analysis aiming at constructing a scale.

[3] Reprinted with permission from the publisher.

Unfortunately, a typical finding of goodness-of-fit research is that the model is inconsistent with the data. When this happens, you do not have a scale. This is a handicap of research with human beings, which led the famous statistician, George Box, to his much-cited quote that all models are wrong. Are we on a hopeless mission? No, fortunately George added that some models are useful.

ALL MODELS ARE WRONG, BUT SOME ARE USEFUL

2-parameter normal-ogive model

$$P_j(\theta) = P(y_{ij} > 0) = P[\varepsilon_{ij} > -a_j(\theta_i - b_j)] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a_j(\theta_i - b_j)} \exp\,(-y^2/2)dy$$

general diagnostic model

$$P(X_j = x|\boldsymbol{\alpha}) = P_j(\boldsymbol{\alpha}) = \frac{\exp[\beta_{jx} + \sum_{k=1}^{K} \gamma_{jkx} h(q_{jk}, \alpha_k)]}{1 + \sum_{y=1}^{M_j} \exp[\beta_{jy} + \sum_{k=1}^{K} \gamma_{jky} h(q_{jk}, \alpha_k)]}$$

lognormal model for response time

$$f(t_{ij}; \mathcal{E}(\ln t_{ij}), \sigma_j^2) = \frac{1}{t_{ij}} . \frac{1}{\sigma_j \sqrt{2\pi}} \exp\left\{- \frac{[(\ln t_{ij} - \mathcal{E}(\ln t_{ij}))]^2}{2\sigma_j^2}\right\}$$

—George E. P. Box—

Figure 5: George Box' famous quote on the inconsistency between model and data illustrated by means of three psychometric models that—like all models—approximate the data structure at best, but when they do can be useful for the application envisaged.

Why are models wrong in principle? The reason is that models are simplifications of the phenomenon they intend to explain. Their ambition is to pick up the salient characteristics and ignore the details, and in doing that, a model that fits the data perfectly is impossible to attain. Look at it from the opposite side: Models *must* fail at describing the data; if they didn't, they would coincide with the data and provide no distinction between main principles and details. But without a correct guiding theory, for example, about transitive reasoning, models do not even catch all salient characteristics and will fail even more convincingly than when they only missed the details.

This situation is true in most of science. For example, it is also true for the models the RIVM uses to predict the course of the COVID-19 pandemic. So, if you felt the RIVM was not doing a good job because their predictions failed now and then, you must realize that reality is more complex than the model used to understand it. However, the RIVM models often were approximately correct and therefore they were useful. The proof of the pudding is in whether the less-than-perfect model improves making predictions compared to what we already know without the model.

The absence of a well-founded theory of the phenomenon of interest is an important cause of model-data misfit. If you want to measure, say, intelligence, you need a theory of intelligence firmly rooted in research. Such a theory tells you what the behavior is that is typical of spatial orientation or transitive reasoning, so that you know which behaviors you must assess to measure these attributes. The test tasks must be chosen such that they elicit this behavior and nothing else. The absence of well-founded theory for most attributes is the Achilles heel of psychological measurement. Without a well-founded theory, you must rely on experience, habit, tradition, and educated guesses. Clearly, this situation is far from ideal and stands in the way of scientific progress.

Theory-based measurement in the natural sciences has gone a long way and most measurement instruments were developed in the past century. Also typical of natural-science measurement is its emphasis on ***unobtrusive measurement***, eliminating all the disturbances that bias or attenuate measurement. My pharmacy lab work involved endlessly cleaning the materials that I used for the chemical measurements I carried out. I still have the habit of rinsing the coffee pot based on those lab rules before I prepare a fresh pot. Modeled after physics, psychology copied this practice a century ago but seems to have forgotten the importance of standardization. Today, data collection through the internet reduces control over different types of measurement error. Additional error sources are data-driven scale construction attempts downplaying the leading role theory must have, and financial arguments promoting measurement using only a few tasks thus reducing reliability.

You should have understood by now that measurement is ***problematic*** (also, see Borsboom, 2005; Briggs, 2022; Lord & Novick, 1968; Michell, 1999) and the construction of a valid and reliable scale a small miracle when it succeeds. The best piece of advice I have is that psychology must focus on the development of well-founded theory about attributes, so that theory guides the construction of measurement by means of tests and questionnaires. The work of Brenda Jansen (Jansen & Van der Maas, 1997) on proportional reasoning and Samantha Bouwmeester (Bouwmeester, Vermunt, & Sijtsma, 2007) on transitive reasoning may serve as guidelines. The focus on theory pushes validity and additionally encourages standardization and reliable measurement. Psychometrics tends to focus on developing measurement models but must help psychology to develop the substantive theories for psychological attributes. A situation in which measurement models are developed that can have no connection with underlying theories because the development of such theories has low priority, is highly undesirable. Without valid and reliable measurement, there can be no useful research.

3 Statistics

Statistics is the science of uncertainty. Statistical methods are designed for estimating the degree of uncertainty when making inferences from the sample about the population of interest. Statistics works best when the sample is a representation of the population that only differs from it by random fluctuation. Various sampling methods have been proposed to finetune the sample composition to the research question of interest, and statistical testing and model estimation often must be adapted to the way the different methods sample cases from the population. The question is always: What would happen when I do the sampling all over again?

Coincidence lies at the basis of sampling. We people are extremely bad at understanding coincidence or randomness. The reason is that we tend to see structure everywhere, even when it is absent. Assigning meaning to results from data analysis, even when the results are unexpected, implausible, unstructured, or pure noise, is something we simply are unable to avoid. The Dutch psychologist Linschoten (1964) warned already in 1964 about this tendency. Many great psychological researchers, such as Paul Meehl (Meehl, 1954), Robyn Dawes (Dawes, 1994), Amos Tversky and Daniel Kahneman (e.g., Tversky & Kahneman, 1974; also, see Kahneman, 2011) have asked our attention for the plethora of biases that inhabit our cognitive system including that of scientists, and statisticians such as David Hand (Hand, 2014) have explained why we do not understand probability. Let us have a look at a small example I borrowed from Daniel Kahneman (Kahneman, 2011, p. 115).

Suppose on a beautiful Wednesday, six babies, no twins among them, are born in the same hospital. When B stand for boy and G for girl, which chronological order of births do you consider random, and which provides evidence of a pattern?

BBBGGG
GGGGGG
BGBBGB

The intuitive answer is: The first two suggest patterns, the third is random. The correct answer is: They are all equally likely. The number of different patterns is $2^6=64$, and assuming boys and girls have the same probability and there are no direct genetic relations between different parent couples, each pattern has probability $\frac{1}{64}$. This outcome will fly in the face of intuition, which looks for patterns and believable stories, but intuition can be highly misleading. That is why we have science, and statistics in particular; they protect us from intuition.

Now let us have a look at the next problem. I generated artificial data for 25 variables using a computer and a statistical model. With 25 variables, we have $\frac{1}{2} \times 24 \times 25 = 300$ different correlations. I considered non-zero correlations that might be of interest when you come across them in a data analysis. For each sample size I considered, I generated 100 replications because one sample would allow my brain to play all the tricks on me that it does when I am confronted with data patterns. So, replications serve the role of finding out what would happen when I do it again; this is the central question in statistics, remember? Table 2, first line, shows that with sample size 50, each sample had at least one negative correlation between −.29 and −.60 and one positive correlation between .30, and .64. Correlations of this magnitude raise interest in almost any researcher!

There are a few things you should realize when looking at these results. First, for sample sizes of 100 and 500, correlations are smaller. This should tell you something, but what? Second, the second to last column tells you that across 100 replications, 15 correlations were significant. This is 5 percent of 300, so this looks like a result you expect when the null hypothesis of correlation zero is true and you test at a 5 percent significance level. Therefore, it looks as if in the population all 300 correlations are zero, and we have been looking at correlations that are non-zero due to random sampling error only. This is true: I generated the data using coin flipping with heads producing a score 1 and tails a score 0. As a model for the population, we have 300 zero correlations, and samples differ coincidentally from the model. I hope you feel cheated because this makes it easier for you to remember what I am saying.

Table 2: Sample Size, Results Averaged Across 100 Random Samples of Size N. Ranges of Minimum and Maximum Sample Correlations, Minimum, Maximum and Mean Numbers of Significant Results. Critical Values for Null Hypothesis Testing.

N	Range	Range	Min(Sign)	Max(Sign)	Mean(Sign)	Crit Value
	Min	Max				Corr
50	**-.60; -.29**	**.30; .64**	6	25	15.32	.28
100	-.42; -.21	.21; .40	7	23	15.34	.20
500	-.18; -.09	.10; .18	8	28	15.60	.09

Note: Wilco Emons programmed the example.

Based on the table, any statistician will tell you that there are statistical methods that help you to recognize situations like this and protect you from them. True, but the third thing you should realize, is that in real research you have only one data set available and not 99 additional replications that together reveal a pattern, here, a true pattern. And based on just one data set, you might find more than 15 large significant correlations that suggest more than there is! Not only that, not only coin flipping but many different population models could have produced that one sample you found, and coincidence prevents you from recognizing the correct model. The logical problem is here that the model implied the data, but the data do not imply the model. I will skip the possibility to estimate the most likely model; it takes a lot of expert knowledge to use this methodology and understand what the results mean.

The most compelling lesson from these examples and many others is that without a theory you will have a hard time finding out what the population model was that produced the data. Samples often are rather small and sampling error is difficult to capture unless the sample size is very large. I am worried by the many papers I review for journals that do not define the population of interest or only by crude approximation, and use samples that happened to be around, euphemistically called convenience samples. I know from my own experience that doing research, collecting real data with real people, is difficult and that textbook knowledge describing ideal situations is not always useful. But unfortunately, that does not help anyone unless more time and means are spent on developing theories for the phenomena of interest and then meticulously testing and improving the theories and with that, our understanding of the world around us. I hope what is nowadays called slow science contributes to this enterprise.

I jump to questionable research practices. They represent structural ways of working that produce invalid research outcomes. I will assume that researchers have no intention to do this but that the underlying cause is insufficient mastery of statistics. Combine this with a tendency in several research areas to work exploratory rather than confirmatory and you have a toxic mixture producing invalid results. Just think of the coin-flip data and results I just showed you. If your hypothesis would have been that the data are the result of coin flips, then the resulting sample correlation matrix would be supportive of your hypothesis. The reason is that before collecting the data and looking at the results, you made known what you expected to find. So, you narrowed down your chances considerably and this is what strengthens confirmatory research. But it still does not provide proof. Proof as in mathematical proof is impossible to attain with real-data research based on samples. But if replications consistently point to the expected outcome, you have got something on your hands!

If you first look at the data without expressing an unequivocal expectation, a correlation looks spectacular and your natural talent for storytelling will do the rest. But you do not have a clue where the correlation comes from. You simply never expected it, but now that it presents itself, albeit as a coincidental present, you find it difficult to ignore it. Perhaps you do not even try, and who can blame you. Understanding coincidence and probability are counterintuitive and statistics based on these concepts is therefore difficult. John Ioannidis (Ioannidis, 2005) was right when he warned about the frequent occurrence of invalid research results!

A focus on theory construction, testing, and improvement is one great way to protect yourself from questionable research practices. Publishing your data and all the information necessary for colleagues to replicate your findings in a publicly accessible repository is another way to reduce questionable research practices. Finally, I suggest working together with methodologists and statisticians who are more experienced than most researchers to recognize clues of cognitive deceit caused by coincidence. If you want to read more about it, there is a huge body of literature available, or you can wait another six months until my book on insufficient mastery of statistical reasoning and methods appears in a series of the American Statistical Association with Chapman & Hall/CRC (Sijtsma, in preparation).

4 Psychology

I started out as a student of psychology but was caught by measurement and statistics too soon to become a full-fledged psychological researcher. But my interest in the history and the development of psychology remained until the present day. This is the most pretentious part of my speech because I know the least of it, but as an outsider having a rather close relationship with psychology, I may have some observations that, if not correct at least I hope are entertaining more than irritating.

Psychology started inspired by the exact sciences. Many psychologists were physicists and introduced a strict, formal way of thinking into psychology. You simply can look at nineteenth and twentieth century psychology, and you will see what I mean (e.g., Murphy & Kovach, 1972). The tendency to resort to experimentation and borrowing theoretical models from physics, chemistry and medicine was not always successful, and I will not claim psychology has to return to these roots. However, I believe there was a stronger belief in theory construction as the cornerstone of a successful science discipline than there is today. Why is that?

I think several developments are responsible for this, and if I am wrong, I would appreciate knowing why. Three developments in chronological order are the following.

First, after the Second World War, the ***computer*** developed at an enormous pace (e.g., Dyson, 2012; Gleick, 2011), but it took until the late 1980s that we all got access to the personal computer and later the laptop computer, both with unbelievable computing power. The advent of the high-powered computer also inspired the development of new statistical methods not developed before because they were too complex to allow doing computations by hand. These complex statistical methods allow doing computations on large numbers of variables collected with samples of a size unheard before. The computer's development also allowed me to concoct the coin-flipping example, which would have been impossible a few decades ago. So, hurray for the computer. My point is that estimating the simultaneous relationships between many variables may distract from identifying causal relationships between fewer key variables, fundamental to understanding larger models. This stands in the way of theory development. Am I wrong?

Second, in the 1990s, ***fMRI scanners*** became affordable, allowing psychologists and other researchers to study neurological activity in response to psychological stimuli. This is the development I know least about, and I have my knowledge from reading non-technical books and a pile of the best master theses in neuroscience and neuropsychology. This is also the topic experts need to inform me about concerning the breakthroughs in psychology that have extended our knowledge of the structural foundation of our cognition and the workings of cognitive and other psychological processes. I know much effort has been spent on this kind of research and recognize the value of introducing insights from other disciplines into your own but would like to know the scientific return on investment. So, this is a request for information rather than a word of criticism.

Third, the availability of the ***Internet*** and tools such as wearables has facilitated the unstandardized collection of incredibly large data sets with hundreds, even thousands of variables and enormous samples. Several researchers seem to consider this Shangri-La for data analysis, facilitating the study of numerous models and hypotheses. I have three questions. First, given that beyond a couple of thousand cases, larger samples do not provide much additional statistical information, why would you want such enormous samples? In addition, given that populations are often ill-defined, what do these giant samples represent? Second, the large numbers of variables exponentially increase the finding of coincidental relations and fitting models to probability 1. There is always something to find in such huge data sets, but what have you found? Think of the coin-flipping example. Third, without a guiding theory, exploration will produce a large heap of incidental findings impossible to replicate. Why would a huge sample not based on expectations founded by theory contain anything useful?

The three developments are not typical of psychology, but they are hazardous as guiding principles the less a discipline relies on theory construction, testing, and improvement. My point is that each development represents a technological miracle, but is relatively ineffective when used without a plan, such as a theory. Of course, I know that I am presenting a one-sided discussion of these impressive developments, and I am sure that several applications of these technological innovations are highly supportive of the development of psychology. But looked at from a distance, I have the impression they are used without making much distinction between relevant and irrelevant applications. A running joke I heard too often not to take at least a little serious was that you were without much of a chance applying for a grant with NWO when you did not scan a few brains.

My point is not to refrain from complex statistical models, fMRI scans, and Internet data collection. That would be silly (e.g., Domingos, 2015; Myin-Germeys & Kuppens, 2022; Pearl & Mackenzie, 2018). But I do think we should use these technical innovations for theory construction, testing, and improvement. Exploration is very useful for generating new ideas, which should be put to the test in confirmatory research.

I think psychology should take a little more time to retrace its theoretical steps, and if they already did but I missed it, it should continue its course. Progress takes a lot of time and impatience leads to results that rarely persist. Psychology is not an old but also not a young discipline, mostly originating in the nineteenth century, just like chemistry, by the way. But many of the ideas central to the exact sciences date back much further. Take for example the idea that everything in the universe is composed of atoms (Rovelli, 2017). The idea goes back some 2,500 years to the ancient Greek thinker and early scientist, Democritus. And even though his ideas were in the right direction, it took us almost 2,500 years to make something useful of it. Science is a hurdle race and ideas are replaced all the time by other ideas, sometimes better and sometimes worse. But on average better, so that there is progress!

All of this is not to say that psychology has not provided important theoretical insights. On the contrary! I mention a few highlights familiar and important to me:

- First, I mention the effort put into classification of intelligence and personality in constituent components, as in Guilford's and Thurstone's intelligence models and the Big Five personality traits. Classification fosters a better understanding of the phenomenon of interest, an insight entertained in the 18th century by Carolus Linneaus when he classified the animal and plants kingdom.

- Second, I cherish the crucial insight Paul Meehl had that like all other people, experts are bad information processors to such a degree that a very simple additive model run on a computer will better predict future behavior than an expert who has all the information for a correct decision available.
- Third, I recommend reading the impressive studies Amos Tversky and Daniel Kahneman conducted, revealing that our judgment and decisions are riddled with cognitive bias and suffer greatly from reliance on intuition based on heuristics and the difficulty to reason rationally. Herbert Simon's contributions in these and other areas can hardly be overestimated.
- Fourth, psychonomics traditionally focused on sensory perception related to cognitive processes, and its focus has been theory-driven using experimental research as the main methodology. It has produced many important results, for example used in behavior genetics and ergonomics. In the same vein, I mention the work in language development and processing.

There is much more, of course, and all contributions share the preference for theory development based on sound empirical research.

I have already mentioned the great insight late 19th and early 20th century psychologists like Charles Spearman (1904a, b; 1910) and Alfred Binet (Binet & Simon, 1905) had that measurement is pivotal to scientific research and the development of theory. In this lecture, I have expressed concern that considering the rather unlimited expectation psychologists and many others have of technological innovations, psychology must not forget the crucial role measurement plays in the development of psychology as a science. My call to psychology is to continue paying attention to valid and reliable measurement as the basis of scientific research and theorizing, and invest in technological innovation when the theory-driven research calls for it.

5 Lifelong Learning

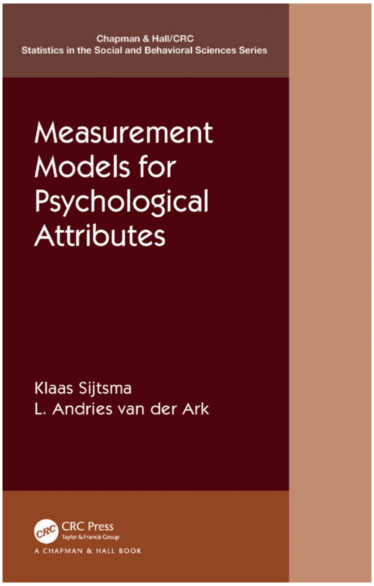
Or better: A life of learning. I first went to school in 1960 when I was four years old, almost five. Next week, after almost 62 years, I will leave school.

Aging has the disadvantage that you build up an immense collegial network, making it impossible to thank everyone when you retire. I will do this groupwise, a few exceptions noted.

First, I thank the several thousands of students that I tried to teach statistics and measurement, and who taught me that statistics is difficult, for everybody.

Second, I thank my PhD students, whom I tried to guide through the painful process of doing research and communicating results. They taught me that everybody needs a different approach and several surprised me with their outstanding, even brilliant ideas.

Third, I thank all my colleagues, everywhere, but Andries van der Ark in particular. In his inaugural lecture, Andries told the audience that he and I published 36 articles together. That number has grown to 50, including a book. I also thank the late Ivo Molenaar and Pieter Drenth, my PhD supervisors, Rob Meijer, Bas Hemker, Brian Junker, Jeroen Vermunt, Wilco Emons, and Julius Pfadt.



Chapman & Hall/CRC
Statistics in the Social and Behavioral Sciences Series

Measurement Models for Psychological Attributes

Klaas Sijtsma
L. Andries van der Ark

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

I fulfilled all the administrative and board positions a university has to offer and thank everybody who helped me to improve the Department of Methodology and Statistics, the School of Social and Behavioral Sciences, and Tilburg University. Even more, I thank those of you who felt you had to deal with me and succeeded.

I thank the dean of our School, Antoinette de Bont, and the rector magnificus, Wim van de Donk, for their generous support during the previous year. I mention former rector magnificus Philip Eijlander in particular for his support during the previous decade.

Finally, my deepest thanks go to my family, Marjon, Leonie, and Hester, and Romboud and Yarah, for their unconditional love and support.

Well, what can you say after 62 years in school? I will continue doing research and a few other things, but for now: School’s out!

Ik heb gezegd.

6 References

Binet, A., & Simon, Th. A. (1905). Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *L'Année Psychologique, 11,* 191-244.

Borsboom, D. (2005). *Measuring the mind. Conceptual issues in contemporary psychometrics*. Cambridge UK: Cambridge University Press.

Bouwmeester, S., Vermunt, J. K., & Sijtsma, K. (2007). Development and individual differences in transitive reasoning: A fuzzy trace theory approach. *Developmental Review*, *27*, 41-74.

Briggs, D. C. (2022). *Historical and conceptual foundations of measurement in the human sciences*. New York: Routledge.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302.

Dawes, R. M. (1994). *House of cards. Psychology and psychotherapy built on myth*. New York: The Free Press.

Domingos, P. (2015). *The master algorithm. How the quest for the ultimate learning machine will remake our world*. London, UK: Penguin Books.

Drenth, P. J. D. (1975). *Inleiding in de testtheorie* (*Introduction to test theory*). Deventer, the Netherlands: Van Loghum Slaterus.

Dyson, G. (2012). *Turing's cathedral. The origins of the digital universe*. London: Penguin Group.

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2007). On the consistency of individual classification using short scales. *Psychological Methods*, *12*, 105-120.

Gleick, J. (2011). *The information. A history. A theory. A flood*. New York: Vintage Books.

Hand, D. (2014). *The improbability principle. Why coincidences, miracles and rare events happen every day*. London, UK: Penguin Books.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Med*, *2*(8), https://doi.org/10.1371/journal.pmed.0020124

Jansen, B. R. J., & Van der Maas, H. L. J. (1997). Statistical test of the rule assessment methodology by latent class analysis. *Developmental Review, 17*, 321–357.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258-272.

Kahneman, D. (2011). *Thinking, fast and slow*. London, UK: Penguin Books.

Linschoten, J. (1964). *Idolen van de psycholoog* (*The psychologist's idols*). Utrecht, the Netherlands: Bijleveld.

Lord, F.M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison -Wesley.

Magidson, J., & Vermunt, J. K., (2004). Latent class models. In D. Kaplan (Ed.), *The Sage Handbook of Quantitative Methodology for the Social Sciences* (pp. 175-198). Thousand Oaks, CA: Sage.

Markus. K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: measurement, causation, and meaning*. New York, NY: Routledge.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis, MN: University of Minnesota Press.

Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept*. Cambridge, UK: Cambridge University Press.

Murphy, G., & Kovach, J. K. (1972). *Historical introduction to modern psychology*. London: Routledge & Kegan Paul Limited.

Myin-Germeys, I., & Kuppens, P. (2022). *The open handbook of experience sampling methodology*. The center for Research on Experience sampling and Ambulatory methods Leuven (REAL), Belgium. Downloaded from https://www.kuleuven.be/samenwerking/real/real-book/index.htm

Niessen, A. S. M., & Meijer, R. R. (2017). On the use of broadened admission criteria in higher education. *Perspectives on Psychological Science, 12*, 436-448. https://doi.org/10.1177/1745691616683050

Pearl, J., & Mackenzie, D. (2018). *The book of why. The new science of cause and effect*. Penguin Books UK.

Rovelli, C. (2017). *Reality if not what it seems. The journey to quantum gravity*. Pinguin Random House UK.

Sherry, D. (2011). Thermoscopes, thermometers, and the foundations of measurement. *Studies in History and Philosophy of Science, 42*, 509–524.

Sijtsma, K. *Never waste a good crisis. Lessons learned from data fraud and questionable research practices* (in preparation for Chapman & Hall/CRC).

Sijtsma, K., & Van der Ark, L. A. (2021). *Measurement models for psychological attributes*. Boca Raton, FL: Chapman & Hall/CRC.

Spearman, C. (1904a). Proof and measurement of association between two things. *American Journal of Psychology, 15*, 72–101.

Spearman, C. (1904b). 'General intelligence,' objectively determined and measured. *American Journal of Psychology, 15*, 201-293.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*, 271-295.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124–1131.

Colofon

design
Beelenkamp ontwerpers, Tilburg
photography cover
Maurice van den Bosch
layout and printing
Studio | powered by Canon