1 Introduction

1.1 The measured mind

Psychological measurement plays an important role in modern society. Teachers have schoolchildren tested for dyslexia or hyperactivity, parents have their children's interests and capacities assessed by commercial research bureaus, countries test entire populations of pupils to decide who goes to which school or university, and corporate firms hire other corporate firms to test the right person for the job. The diversity of psychological characteristics measured in such situations is impressive. There exist tests for measuring an enormous range of capacities, abilities, attitudes, and personality factors; these tests are said to measure concepts as diverse as intelligence, extraversion, quality of life, client satisfaction, neuroticism, schizophrenia, and amnesia. The ever increasing popularity of books of the test-your-emotional-intelligence variety has added to the acceptance of psychological testing as an integral element of society.

When we shift our attention from the larger arena of society to the specialized disciplines within scientific psychology, the list of measurable psychological attributes does not become shorter but longer. Within the larger domain of intelligence measurement, we then encounter various subdomains of research where subjects are being probed for their levels of spatial, verbal, numerical, emotional, and perceptual intelligence; from the literature on personality research, we learn that personality is carved up into the five factors of extraversion, neuroticism, conscientiousness, openness to experience, and agreeableness, each of these factors themselves being made up of more specific subfactors; and in clinical psychology we discover various subtypes of schizophrenia, dyslexia, and depression, each of which can be assessed with a numerous variety of psychological tests. In short, scientific psychologists have conjured an overwhelming number of psychological characteristics, that can each be measured with an equally overwhelming number of testing procedures. How do these procedures work?

Consider, as a prototypical example, the measurement of intelligence. Intelligence tests consist of a set of problems that are verbal, numerical, or figural in character. As can be expected, some people solve more problems than other people. We can count the number of problems that people can solve and look at the individual differences in the computed scores. It so happens that the discovered individual differences are relatively stable across adulthood. Also, different tests for intelligence tend to be positively correlated, which means that people who solve more verbal problems, on average, also solve more numerical and figural problems. There is thus a certain amount of consistency of the observed differences between people, both across time periods and across testing procedures.

As soon as a way is found to establish individual differences between people, all sorts of correlations between the test scores and other variables can be computed. So, we can investigate whether people with higher intelligence test scores, when compared with people who obtain lower test scores, are more successful on a job; whether they make more money, vote differently, or have a higher life-expectancy. We can look into differences in intelligence test scores as a function of background variables like sex, race, or socio-economic status. We can do research into the association between intelligence test scores and neural speed, reaction time, or the amount of grey matter inside the skull. We find a diverse array of associations and mean differences. Some are large and stable, others small and difficult to replicate. And so the mechanism of science has been set in motion. Under which conditions, precisely, do certain effects occur? Which variables mediate or moderate relations between intelligence test scores and other variables? Are these relations the same in different groups of people? Once the scientific engine is running, more research is always needed.

However, a nagging question remains. Do such tests really measure something and, if so, what is it?

This book originates from my attempts to make sense of this question, which is encountered in virtually every field where psychological tests are used. In the past century, it has become known as the problem of test validity. Test validity has proven to be an elusive concept, which is illustrated by the fact that empirical validation research tends to be highly inconclusive. The issue remains problematic, in spite of the fact that psychological tests have come to belong to the standard equipment of social science research, and in spite of the enormous amounts of empirical data that have been gathered in psychological measurement. In fact, after a century of theory and research on psychological test scores, for most test scores we still have no idea whether they really measure something, or are no more than relatively arbitrary summations of item responses.

One of the ideas behind the present book is that part of the reason for this is that too little attention has been given to a conceptual question about measurement in psychology: what does it *mean* for a psychological test to measure a psychological attribute? The main goal of this book is to investigate the possible answers that can be given in response to this question, to analyse the consequences of the positions they entail, and to make an informed choice between them.

1.2 Measurement models

To the psychologist, to question what it means to measure a psychological attribute may seem somewhat odd. Many are under the impression that there exists considerable consensus among methodologists on how psychological measurement should be conceptualized, or even that there *is* only one theory of psychological measurement. One reason for this is that textbooks on research methods tend to propagate such a harmonious picture. However, this picture is inaccurate. Several distinct theories have been proposed for psychological measurement, and these sometimes differ radically in the way they conceptualize the measurement process. In order to make an informed judgment on the merits of psychological testing, it is critical to understand and evaluate these models.

There are three important types of measurement models, each of which is discussed in a separate chapter in this book. These are the classical test model, the latent variable model, and the representational measurement model. Each of these models proposes a way of conceptualizing theoretical attributes, and specifies how they relate to observed scores. They provide blueprints for ways of thinking about psychological measurement.

The *classical test theory model* is the theory of psychological testing that is most often used in empirical applications. The central concept in classical test theory is the true score. True scores are related to the observations through the use of the expectation operator: the true score is the expected value of the observed score. Thus, a researcher who sees intelligence as a true score on an intelligence test supposes that somebody's level of intelligence is his expected score on an IQ-test. Theories of psychological testing are usually related to a more or less coherent set of suggestions or prescriptions that the working researcher should follow, and classical test theory is no exception. For instance, a typical idea of classical test theory is that the main task of the researcher is to minimize random error. One way of doing this is by aggregating scores on different tests; the longer the test, the more reliable and valid it will be.

The *latent variable model* has been proposed as an alternative to classical test theory, and is especially popular in psychometric circles. The

central idea of latent variable theory is to conceptualize theoretical attributes as latent variables. Latent variables are viewed as the unobserved determinants of a set of observed scores; specifically, latent variables are considered to be the common cause of the observed variables. Thus, a researcher who views intelligence as a latent variable supposes that intelligence is the common cause of the responses to a set of distinct IQ-items or tests. A typical idea associated with latent variable theory is that the researcher should set up a statistical model based on a substantive theory, and that the fit of the model should be tested against observed data. Only if this fit is acceptable is the researcher allowed to interpret observations as measurements of the latent variables that were hypothesized.

Finally, the *representational measurement model* – also known as 'abstract', 'axiomatic', or 'fundamental' measurement theory – offers a third line of thinking about psychological measurement. The central concept in representationalism is the scale. A scale is a mathematical representation of empirically observable relations between the people measured. Such an empirical relation could be that John successfully solved items 1, 2, and 3 in an IQ-test, while Jane solved items 1 and 2, but failed on item 3. A mathematical representation of these relations could be constructed by assigning John a higher number than Jane, which indicates that he solved more items, and by assigning item 3 a higher number than items 1 and 2, which indicates that it was less often solved. In doing so, the researcher is constructing a scale on which persons and items can be located. Because this scale is a man-made representation, much like a map of an area, the researcher who views intelligence as a scale supposes that intelligence is a representation of observable relations between people and IQ-items. A typical suggestion associated with this theory, is that the task of the researcher is to establish these relations empirically, to prove that they can be represented in a formal structure with a certain level of uniqueness, and to find a function that gives a mathematical representation which is isomorphic to the empirically established relations.

As may be suspected from this discussion of measurement models, and as will certainly become clear in this book, classical test theory, latent variable theory, and fundamental measurement theory present radically different ideas on what measurement is and how it should be done. One of the aims of this book is to show how these theories differ, what these differences mean in the context of psychological measurement, and to evaluate their potential for furthering progress in psychological measurement.

What are the proper grounds for the evaluation of measurement models? At first glance, it may seem that the objective is to choose the 'correct'

or 'true' model. However, this is not the case. Measurement theories involve a normative component, in the sense that they suggest evaluative dimensions to assess the quality of testing procedures, but they cannot themselves be candidates for truth or falsity. The reason for this is that measurement models are not scientific theories. The relation that measurement models bear to scientific theories is very much like the relation a blueprint bears to a building. A blueprint may be invaluable in constructing a bar, but one cannot have a beer inside one. Analogously, a generic measurement model can be invaluable in setting up a testable substantive theory about how intelligence relates to IQ-scores, but it is not itself testable. For instance, a latent variable model for general intelligence could be tested by fitting a common factor model to IQ-scores, because this model yields testable consequences; however, if the model is rejected, this has consequences for the theory that intelligence is a latent variable underlying the IQ-scores, but not for the common factor model as a measurement theory. The common factor model would, at most, be useless as a measurement model for IQ-scores. It could never be true or false in itself.

Measurement models are blueprints for (parts of) scientific theories, they can suggest ways of thinking about scientific theories, but they are not themselves scientific theories. They have no empirical content of themselves. Therefore, the type of evidence that works in evaluating scientific theories, which is empirical, does not apply to conceptual frameworks for thinking about such theories; and the benchmarks of truth and falsity, which are invaluable in empirical research, are useless in evaluating measurement models. The proper ground for the evaluation of conceptual frameworks like measurement models lies not in their empirical implications, but in their philosophical consequences. We may find such consequences plausible, or they may strike us as absurd. In conceptual investigations like the one we are about to begin here, plausibility and absurdity play roles analogous to the roles of truth and falsity in empirical research; they show us which way to go.

1.3 Philosophy of science

The evaluation of measurement models in terms of their philosophical consequences implies that such models are at least in part based on philosophical theories. This is, indeed, one of the main assumptions behind this book. The central idea is that measurement models can be considered to be *local* philosophies of science. They are local in the sense that they are concerned with one specific part of the scientific process, namely measurement, rather than with general questions on the aims and

function of scientific theories. The latter type of philosophical theories could be called *global*.

Examples of global theories in the philosophy of science are theories like the logical positivism of the Vienna Circle or Popper's falsificationism. Such theories are based on general ideas on what scientific theories are, on how scientific research should proceed, and on what the best possible result of scientific research is. Global theories on philosophy of science and local theories on measurement models have something in common. Namely, in both cases a central question is how abstract scientific concepts, like intelligence, are connected to concrete observations, like IQ-scores. The difference is that in global theories, the connection between theory and observation is always related to more general ideas on the scientific enterprise, but hardly ever formalized, while in theories of measurement, the connection between theory and observation is always formalized, but hardly ever related to more general theoretical ideas. However, it would be surprising if such similar enterprises – formally specifying the relation between theoretical attributes and observations in a measurement model, and general theorizing on the relation between theoretical attributes and observations from a philosophical viewpoint – touched no common ground. And indeed they do. In fact, in this book I will consider measurement models to be local implementations of global philosophies of science within the context of measurement. The question then becomes: what philosophical viewpoint does a given measurement model implement?

To answer this question, global philosophies of science will be evaluated according to the status they assign to theoretical terms like 'intelligence', 'electron', or 'space-time'. I will divide these theories into two camps, corresponding to the answer they give to a simple question: does the value of scientific theories depend on the existence of the theoretical entities they mention? For instance: does it matter, for the theory of general intelligence, whether or not general intelligence exists?

Theories that answer this question positively are called *realist* theories. Realism gives the simplest interpretation of scientific theories, and it has therefore been described as science's philosophy of science (Devitt, 1991). For the realist, theoretical concepts refer directly to reality, so that intelligence and extraversion are conceptualized as having an existential status quite independent of the observations. The meaning of theoretical concepts derives largely from this reference to reality; general intelligence, for example, would be conceptualized by a realist as an unobservable, but causally relevant concept. We learn about intelligence through its causal impact on our observations, and when we use the term 'intelligence', it is this causally efficient entity we indicate. Such views are embodied in

the writings of many theorists in psychology (Jensen, 1998; Loevinger, 1957; McCrae and John, 1992; McCrae and Costa, 1997). In this book, I will argue that latent variable models are an implementation of this line of thinking.

Theories that answer the question whether theoretical attributes and entities exist negatively are sometimes called anti-realist theories (Van Fraassen, 1980; Devitt, 1991). There are many different theories that sail under this flag, like empiricism (Van Fraassen, 1980), instrumentalism (Toulmin, 1953), social constructivism (Gergen, 1985), and several variants of logical positivism (Suppe, 1977). Of course, when a realist scheme of thinking is denied, an alternative account for the meaning of theoretical terms, and for the successful role they play in the scientific enterprise, must be given; on this account such theories differ. For the reader to get some idea of the lines of reasoning that are commonly employed in anti-realist theories, it is useful to give a brief overview of some of them.

The logical positivists held that scientific theories could be partitioned into an observational and a theoretical part. In their view, a scientific theory could be represented as a set of sentences. Some of these sentences contained theoretical terms, like 'intelligence'. Such sentences were considered to be part of the 'theoretical vocabulary' (Suppe, 1997). The sentences that did not contain such terms, but only referred to observational terms, like 'John has given the response "13" when asked to complete the series 1, 1, 2, 3, 5, 8, . .', were thought to be directly verifiable; they formed the so-called 'observational vocabulary'. A set of 'correspondence rules' was considered to coordinate the theoretical and observational vocabularies (Carnap, 1956). These correspondence rules were supposed to play the role of a dictionary, which for every theoretical sentence returned a set of observation sentences that could be directly verified. If all observation sentences implied by a theory were verified, then the theory as a whole was thought to be confirmed. The emphasis on direct verification is the reason that logical positivism is also known as 'verificationism'. In logical positivism, theoretical attributes, properties, and entities are logical constructions, which are implicitly defined through the observational vocabulary by means of correspondence rules. Theoretical attributes have no referents in reality.

Instrumentalism (Toulmin, 1953) also has an anti-realist orientation, but for different reasons. In contrast to logical positivists, instrumentalists do not see it as the task of science to give us verified or true theories. From an instrumentalist viewpoint, theories are instruments that allow us to predict future events and to exercise control over our environment. Theoretical attributes are part of the predictive machinery that science

gives us. They allow us to make inferences to events that have not yet been observed. However, theoretical attributes need not refer to structures in the world. A theory may be useful without this being the case, and, according to instrumentalists, usefulness is the only appropriate criterion for scientific theories to satisfy. Thus, instrumentalists hold that the question whether theoretical attributes exist or not is both scientifically and philosophically unimportant.

Social constructivism (Gergen, 1985) is like instrumentalism in that it denies truth or verification to be a relevant criterion for theories to satisfy, and like logical positivism in that it views theoretical attributes as figments of the researcher's imagination. Researchers are not viewed as passive observers, but as actively involved in the construction of the world they study. Reality is thus seen not as a independent benchmark for scientific theories, but as a construction on part of the researcher, and so is truth. Social constructivists hold that theoretical attributes do not exist, but arise from a sort of negotiation or social exchange between researchers. The meaning of theoretical terms like 'intelligence' is actively constructed in this social exchange; in this respect, social constructivism is reminiscent of the ideas of the later Wittgenstein (1953) and of postmodern philosophy. However, it is unclear whether the word 'construction' is to be interpreted literally or metaphorically, especially because social constructivists do not tend to be very specific on how the construction process is carried out (Hacking, 1999). This contrasts sharply with logical positivism, in which the construction process is spelled out in detail.

Logical positivism, instrumentalism, and social constructivism differ in many respects. However, they share a fundamentally anti-realist view on the status of theoretical attributes or entities. In this view, one denies not only the existence of a theoretical attribute such as general intelligence, but also that the adequacy of the theory of general intelligence depends on its existence. This means that one is not merely an epistemological anti-realist (meaning that one does not know for sure whether, say, general intelligence exists; a position that, I think, every serious scientist should take) but that one is also a semantic anti-realist (meaning that one does not even think that whether general intelligence exists matters to the theory of general intelligence). I will generically indicate this line of reasoning as *constructivist*.

Like any philosopher of science, the constructivist needs an account of measurement. He cannot view the attributes measured in science as structures in reality that exist independently of our attempts to measure them. Thus, in order to remain consistent, the constructivist needs a measurement theory that squares with the idea that theoretical attributes, like general intelligence, are produced by the scientists who use intelligence

tests. One way to achieve this is by conceptualizing general intelligence as a representation of empirically observable relations. A mathematically rigorous model that can be used for this purpose is the representational measurement model. Thus, representational measurement theory implements a constructivist philosophy of science at the level of measurement. In particular, it will be argued that representational measurement theory is an incarnation of the logical positivist doctrine on the relation between theoretical and observational terms.

Summarizing, the argument to be made in the following chapters is that latent variable theory is an implementation of realist thinking, while representationalist theory implements a constructivist philosophy. Where does this leave the most influential of theories, classical test theory? Is it an implementation of constructivist, or of realist philosophy? It turns out that classical test theory is highly ambiguous with respect to this issue. It can, however, be clearly coordinated with the semantic doctrine of operationalism, which holds that the meaning of theoretical concepts is synonymous with the operations used to measure them. This idea has well-known implausible consequences, and these will be shown to apply to classical test theory with equal force. Thus, the most influential and widely used test theory in psychology may rest on defective philosophical foundations.

1.4 Scope and outline of this book

Classical test theory, latent variable theory, and representational measurement theory are evaluated in chapters 2, 3, and 4, respectively. In chapter 5, I will enquire whether it is possible to integrate these different measurement models and the associated philosophical positions into a larger framework. It is argued that, at least as far as their statistical formulation is concerned, the models can be viewed from a unified perspective. However, even though the models are closely connected to each other under a suitable choice of model interpretation, the focus of each model remains different. In particular, true score theory deals with error structures, fundamental measurement concentrates on the representation of observed relations, and latent variable models address the sources of variation in the test scores. Because latent variable theory is the only model that explicitly addresses the question where variation in scores comes from, the question of test validity is best considered within this framework. Chapter 6 develops a concept of validity that is based on a realist interpretation of psychological attributes closely connected to latent variable theory.

2 True scores

> Nothing, not even real data, can contradict classical test theory . . .
> Philip Levy, 1969

2.1 Introduction

In September 1888, Francis Ysidro Edgeworth read a paper before Section F of the British Association at Bath, in which he unfolded some ideas that would profoundly influence psychology. In this paper, he suggested that the theory of errors, at that point mainly used in physics and astronomy, could also be applied to mental test scores. The paper's primary example concerned the evaluation of student essays. Specifically, Edgeworth (1888, p. 602) argued that '. . . it is intelligible to speak of the mean judgment of competent critics as the true judgment; and deviations from that mean as errors'. Edgeworth's suggestion, to decompose observed test scores into a 'true score' and an 'error' component, was destined to become the most famous equation in psychological measurement: $O_{bserved} = T_{rue} + E_{rror}$.

In the years that followed, the theory was refined, axiomatized, and extended in various ways, but the axiomatic system that is now generally presented as classical test theory was introduced by Novick (1966), and formed the basis of the most articulate exposition of the theory to date: the seminal work by Lord and Novick (1968). Their treatment of the classical test model, unrivalled in clarity, precision, and scope, is arguably the most influential treatise on psychological measurement in the history of psychology. To illustrate, few psychologists know about the other approaches to measurement that are discussed here: you may be able to find a handful of psychologists who know of latent variables analysis, and one or two who have heard about fundamental measurement theory, but every psychologist knows about true scores, random error, and reliability – the core concepts of classical test theory.

The main idea in classical test theory, that observed scores can be decomposed into a true score and an error component, has thus proved a