Conceptual Issues in Psychological Measurement

Denny Borsboom

Conceptual Issues in Psychological Measurement

Denny Borsboom

ISBN 90-5470-103-X
Printed by PrintPartners Ipskamp, Enschede

Conceptual Issues in Psychological Measurement

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. mr. P.F. van der Heijden
ten overstaan van een door het college voor promoties ingestelde
commissie, in het openbaar te verdedigen in de Aula der Universiteit
op woensdag 11 juni, te 14:00 uur

door

Denny Borsboom

geboren te Den Haag

Promotores:	Prof. dr. G.J. Mellenbergh
	Prof. dr. J. Van Heerden

Faculteit Maatschappij- en Gedragswetenschappen
Afdeling Psychologie
Universiteit van Amsterdam

PREFACE

In hindsight, it is notoriously difficult to escape the impression that there have been certain turning points, or landmark events, in one's intellectual development. In my personal reconstruction of this process, two such events occurred when I was still a student, and both involved my later supervisors.

The first took place when I had just enrolled as a student at the Psychological Methods department. Like all new students in methodology, I followed a course on formal models which was given by Don Mellenbergh. Don introduced the classical test theory model by stating that he was going to do something strange. He was going to brainwash people inbetween testing occasions. When he then took the expectation over such replications, and subsequently introduced a population sampling scheme, he could drop the reference to the expectation of replications again, and all kinds of elegant expressions followed for important concepts like reliability. This seemed almost like a magic trick to me. It was the first time I started suspecting that there was something strange with the formal models that we use in the analysis of psychological test scores, and that it had to do with the application of the expectation operator.

The second event concerned a question posed by Jaap van Heerden at a technical colloqium about item bias. The speaker casually remarked that the latent variable could be considered the cause of its indicators. After the talk, Jaap looked over his glasses in a fashion that I later learned to be very characteristic of him, and said, "so, you would say that mentalistic concepts can play a causal role...". Then there was silence. When the speaker had regained his calm, he tried to defend his position by invoking the analogous thesis that mass was obviously the cause of readings on a balance scale. Jaap turned out to doubt this too. I remember thinking, "that man is crazy", but the question did not let go of me. I do not know whether Jaap's degree of inquisitiveness is to be considered pathological, but if it is, then I am afraid that I have developed a similar syndrom over time, so I hope not.

The combination of these two fundamental problems – that is, the interpretation of probability in psychometrics, and the theoretical status of concepts like latent variables – are the basis of this book. For when I became a Ph D. student with Don and Jaap, it seemed like a good idea to devote some attention to these issues. However, in searching the literature, I failed to find a thorough analysis of these problems, although possible conceptualizations were sometimes alluded to – usually in footnotes, or in paragraphs that were accompanied by notes like "reading of this section can be omitted without loss of continuity". This surprised me. I did find a wealth of literature on related questions in the philosophy of science, and I also found many psychological articles on the status of substantive traits, like general intelligence and the Big Five. None of these papers, however, was directly concerned with latent variable theory as it is used in psychometrics. Applying the frameworks of philosophy of science to the theoretical status of latent variables was, in my opinion, very clarifying, although the resulting paper (Chapter 3 in the present dissertation) can arguably be said to raise more problems that it solves. It turned out, however, that to apply philosophy of science to the question how technical concepts in psychometrics may relate to substantive concepts in psychology, was a very good way to elucidate the issues involved. This dissertation applies this line of thinking to different theories of psychological measurement. I have chosen to include the true score model, the latent variable model, and the representational measurement model in the analysis. The treatment must therefore be considered incomplete. Missing from the analysis are generalizability theory and multidimensional scaling. Also missing are chapters about the theoretical status of observed scores, and about a mysterious species of entities known as 'constructs'. Like life itself, however, a Ph. D. studentship is short, and choices have to be made.

Although the material included in this dissertation should be read as a book, the chapters are based on papers that have been published elsewhere or are currently submitted. Specifically, Chapter 2 is partly based on Borsboom & Mellenbergh (2002). Chapter 3 is a slightly adapted version of Borsboom, Mellenbergh, & Van Heerden (*in press*). Parts of Chapter 4 are based on Borsboom & Mellenbergh (*submitted*). The last part of Chapter 5 is based on Borsboom, Van Heerden, & Mellenbergh (*in press*); I must, however, say that I no longer subscribe to the conclusion of that paper, and my views on validity are now better represented in Chapter 6, which is based on Borsboom, Mellenbergh, & Van Heerden (*submitted*). An overview of and introduction to this material is given in the first chapter of the book. In addition, I have chosen to include two articles (Borsboom, Mellenbergh, & Van Heerden, 2002-a; Borsboom, Mellenbergh, & Van Heerden, 2002-b) as appendices.

CONTENTS

1. INTRODUCTION

The most ordinary things are to philosophy a source of insoluble puzzles. With infinite ingenuity it constructs a concept of space or time and then finds it absolutely impossible that there be objects in this space or that processes occur during this time...
– Ludwig Boltzmann, 1905

1.1 Philosophy of science's insoluble puzzle

Scientific theories say too little and they say too much. They say too little because they require abstraction: No theoretical explanation of a phenomenon includes the details necessary for a perfectly adequate account of empirical observations. Theories aspire to explain enough, rather than everything, and so effects of variables in which the researcher is not interested, or which he hypothesizes to be negligible, are usually left out of the theoretical model. For instance, a researcher who aims to explain how observed differences in scores on a test for spatial ability originate, may hypothesize that these are due to differences in general intelligence. Then he will set up a theoretical model to capture this relation, and in doing so he will exclude effects from other variables, insofar as he thinks this is justifiable. Thus, theoretical models are incomplete. Apart from leaving things out of the theory, however, the scientist also puts things into the theory. This is the process of idealization: In order to set up manageable models, the researcher will ascribe properties to theoretical entities that they could not really have. For instance, in testing his model, the researcher may assume that general intelligence is normally distributed on the continuum. This cannot actually be the case, because there are not enough people to realize such a distribution. However, the researcher may have a theoretical rationale that leads him to expect that the assumption will not be too far besides the truth. This process of idealization is the reason that theories say too much.

For the philosophically inclined mind, this situation is bound to be a source of theoretical problems – or 'insoluble puzzles', as Boltzmann refers to them in the above citation. One of the recurrent discussions concerns the status of the theoretical terms used. What do such terms mean? Do they aim to designate entities or structures that exist in reality? If this is the case, then we have a problem, because idealization ensures that theoretical terms will not exist in the way that the theory, interpreted literally, mentions them, and abstraction ensures that the theory will not be complete. This reasoning thus seems to lead to the conclusion that all theories are false by necessity, which is a problematic conclusion from both a scientific and a philosophical point of view. If one weights the objections to literal interpretations of theories heavily enough, however, one may draw this conclusion. Of course, this will require a substantial change of perspective on the meaning of theoretical terms. Theoretical terms must then be seen as instrumental concepts, which play a useful role in constructing economic ways of describing the empirical observations, but do nothing more. In such an interpretation, however, theories seem to lose the explanatory connotation that usually motivates their development in the first place.

The balance on which the gains and losses of such positions are weighted is the subject of the philosophy of science. All influential positions taken in this area of philosophy result more or less directly from the problem of interpreting theoretical concepts; from logical positivism, which required that all meaningful terms must be reducible to observation statements, to Popperian realism, which held that theories are bold guesses about the world and therefore must be interpreted literally, to Kuhnian relativism, which said that theories themselves carve up the world in digestible pieces, so that scientific progress is impossible because successive theories do not describe the same reality. No satisfactory solution to the general problem has, in my opinion, been formulated; but the questions asked are legitimate, and the philosophy of science has done important work in formulating the possible positions that can be taken with respect to the issue.

1.2 Theoretical terms in psychology

Questions concerning the meaning of theoretical terms are relevant to every scientific area, but in the case of psychology they seem especially pertinent, because psychology has not yet reached the point where these issues can be properly be left to philosophers. Scientific progress may be difficult to define, but it is certain that, if philosophy plays a substantial role in a field of inquiry, then research in that area has not progressed as far as it could have. Now, it is notoriously difficult to pin down the meaning of theoretical concepts in psychology – for instance, Neisser et al. (1996) mention a study in which a number of theorists were asked to define intelligence, and each of them gave a slightly different answer. Such a situation usually means that basic questions about the nature of theoretical concepts are the subject of discussion, and this, in turn, implies that philosophical considerations are still bound to be important, if not central. Thus, the proper conceptualization of the meaning of theoretical concepts in psychology has not yet become a purely philosophical problem; it is an issue that is relevant for theoretical psychology itself. Interestingly, however, the answers to the question, how theoretical terms like 'intelligence' should be interpreted, often exemplify positions taken in the philosophy of science. In psychology, the most important divide runs between realist, operationalist, and empiricist interpretations of these terms.

Realism gives the simplest interpretation of scientific theories, and it has 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; intelligence, for example, is conceptualized 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 (i.e., Jensen, 1999; Loevinger, 1957; McCrae & John, 1992; McCrae & Costa, 1997).

The empiricist denies the referential connection of theoretical terms to reality, and instead claims that theoretical concepts are functions of the observations. In this view, theoretical concepts are fictions, although they may be very useful ones. Once the referential connection to reality has been denied, however, the question becomes what determines the meaning of theoretical concepts. Multiple lines of reasoning may be followed in this respect, but two of them yield positions that have been important in the history of psychology.

The first is to identify the meaning of theoretical concepts with the way they are constructed, reflecting the positivist credo that the meaning of a proposition is its method of verification (Wittgenstein, 1922). In the philosophy of measurement, this position was translated into operationalism, a theory that states that theoretical concepts are synonymous with the set of operations by which they are measured (Bridgman, 1927). In this view, the meaning of the term 'intelligence' is synonymous with the set of operations leading to the score on, say, the Stanford-Binet. Operationalism has been of crucial importance to the development of psychology, because it formed the philosophical basis for behaviorism (Watson, 1913; Skinner, 1938). The crux of operationalism is that it denies that theoretical concepts have any surplus meaning over and above the observations.

A second option that may be taken is to locate the meaning of theoretical concepts in their connections to other concepts figuring in a theory. Since, in the positivist view, a theory is a set of propositions connected by covering laws (Hempel, 1965), this network of connections has become known as the nomological network ('nomos' is Greek for 'law'). In this view, intelligence has surplus meaning over and above the observations, but it derives this meaning from the nomological network in which it figures and not necessarily from a reference to reality. This viewpoint exerted its influence on psychology mainly through Cronbach & Meehl's (1955) formulation of construct validity, which draws heavily upon the notion of a nomological network.

Discussions on the interpretation of theoretical concepts continue to play an important role in many areas; they are the subject of enduring debates in fields like intelligence research (Neisser et al., 1996) and personality theory (Pervin, 1994). In my opinion, however, the generality of these discussions creates a serious problem. The problem is that it is not clear what the subject of the discussion is.

For what do we talk about when we discuss a theoretical concept such as general intelligence? Is it the vague notion researchers have prior to the formulation of a formal model? Possibly, but if this is the subject of our inquiry, we are not likely to come up with a consistent analysis: Researchers A and B may have a very different concept in mind, so that we would have to discuss the status of general intelligence for each researcher anew. Alternatively, we could take the topic of our inquiry to be the slightly less vague notion that figures in 'the' theory of general intelligence. This could be a fruitful approach (it is the one taken in the philosophy of science), apart from the slight problem that 'the' theory of general intelligence does not exist. We do have a largely unspecified system of relatively vague interconnections between rather loosely defined notions, but it does not resemble the well specified, coherent networks we encounter in physics, for example. In psychology, what is usually addressed as a nomological network is more accurately described as nomological sketchwork. The theoretical system is not precise enough to yield a fruitful analysis because, being vague, it is consistent with too many interpretations.

A final possibility is to analyze theoretical concepts as they figure in the theoretical system that meets the data in actual research. This does allow for a consistent analysis, because the 'gaps' in the theory need to be filled in and specified to generate hypotheses that allow for empirical tests, thus yielding a sufficiently precise theoretical framework. However, the fact that unspecified relations in psychological theories are 'filled in' through the application of a model suggests that the meaning of concepts may not be independent of that model. Viewing the issue in this way, therefore, requires an examination of the way theory and model interact in research.

1.3 The function of models in psychology

At first sight, the enterprise called science seems to display a considerable amount of homogeneity. A theory is formulated, hypotheses are derived, data are gathered, results interpreted, and implications for the theory are considered. The generality of this scheme of inquiry has been stressed in various philosophical treatises (De Groot, 1961; Popper, 1959, 1963; Hempel, 1965; Nagel, 1961), and this has prompted philosophers of science to consider the status of theoretical concepts in a similarly general scheme. Upon closer examination, however, there are significant differences between theoretical concepts in the various sciences; and, in psychology, there is a step in the research process that, in my opinion, has received too little attention. This is the step from substantive theory to formal model.

In the natural sciences, and especially in physics, this is such a small step that it hardly needs mentioning. To illustrate this, one need only consider the use of the term 'model'; the meaning of this term is virtually identical to the meaning it has in semantic logic. A theory is, in this view, an *already* formalized system that specifies a class of models. One may think of this in terms of a linear regression. The regression specifies a relation of the form $Y = a + bX$. In the semantic logic approach, this equation is not a model but a *theory*, and all realizations of the equation that are consistent with it (say, $Y = 2 + 3X$) are the *models* of the theory. In this line of thinking, therefore, the 'world' may be considered a particular realization of the equation and therefore as a model of the theory. Such approaches to the relation between theories and the world are not uncommon in the philosophy of science – for instance, Van Fraassen's (1980) constructive empiricism is built on this line of reasoning. However, the terminology will strike the psychologist as foreign.

In psychology, a theory is generally not a formalized system, but a system with a highly verbal, almost narrative, character. But because psychologists nevertheless have a preference for experimental research, rather than for interpretative traditions as common in the sociological and historical sciences, they tend to follow an approach to research that requires testing theories against experimental or quasi-experimental data. One pervasive characteristic of data is that there is, even in the simplest instances of research, too much of it too be processed by the human brain. This has nothing to do with whether or not the data are quantitative in nature; in qualitative research, one also tends to end up with substantive amounts of data, although these usually take the form of transcripts of interviews. In order to make sense of the data, one may sometimes need to start categorizing and counting even in qualitative research. So, most researchers end up with tables of counts of one type of another. Now how does one test a theory against tables of counts? This will almost invariably require setting up hypotheses that yield predictions about the structure of these tables. And this will in turn require the theory to be cast in some kind of formalized form. Usually, but not necessarily, this form will be inspired by statistics; however, note again that this has very little to do with the kind of research procedure that one follows when gathering the data (i.e., using psychological tests, questionnaires, observation techniques, or interviews). It is a problem that one always faces in psychological research.

Thus, testing a theory requires 'translating' the theory into a formal language. This translation is then called a model, and one examines whether the theory is consistent with empirical data indirectly, i.e., through an evaluation of the consistency between the model and the data. This is not to say that the procedure is inconsistent with semantic logic, of course, but to indicate a complication in research that philosophers of science easily overlook. In physics, what one calls a theory is nearly identical with a system of equations. It would be odd to ask whether Newton's $f = ma$ is an adequate translation of his theory concerning the relation between force, mass, and acceleration, because in an important sense this equation *is* Newton's theory. One may, of course, counter that this is a difference of degree, and insist that the concept of mass is just as well represented by the symbol m as intelligence is represented by the latent variable g in a structural equation model. This is probably true, but, to paraphrase Wilkes (1988), it is also true that there is a difference in degree between the moleheaps in my backyard and the himalayas; for the evaluation of theory and research in psychology, the fact that theories are not formalized systems, but relatively vague interconnections between loosely defined notions, does make a difference.

The reason for this is the following. Although the terminology of 'translating' a theory into a model is often used, it is an inadequate description of what really happens. It suggests, for example, that the model is constructed on the basis of the theory, so that the theory dictates the model: If this were the case, models would be tailored on individual theories. With the exception of a handful of research areas in mathematical psychology, psychophysics, and perception research, this is not what happens in psychology. The structure of psychology is such that there are a few widely used formalized systems, such as the generalized linear model (McCullagh & Nelder, 1989), the structural equation model (Jöreskog & Sörbom, 1993), and the generalized linear item response model (Mellenbergh, 1994), and when substantive theories meet the data, they are represented by a variant of one of these general models (no extensive modeling endeavors – e.g., those using path analysis – have to be imagined here; the analysis of variance model is also a model). Theories are, so to speak, forced into a prefab mold.

A direct consequence of this way of working is that it is not only the model that inherits its structure and specification from the theory; the theory inherits characteristics of the model just as well. To give an example, a theory may suggest a causal link between the level of extraversion and attractiveness, so that more extraverted people are considered more attractive. This theory could be tested using a structural equation model. In this case, the theory would inherit certain structures from the model that were not part of its initial specification. For example, if extraversion and attractiveness were conceptualized as latent variables, the researcher would have to assume that these characteristics are normally distributed, that they are linearly related to their indicators and to each other, that the observed variables follow a multivariate normal distribution, that errors are homoscedastic, and that the number of factor loadings equal to zero is large enough to identify the model. While some of the assumptions introduced by the model may be characterized as auxiliary, many cannot be dismissed as such. For example, the assumption of linearity concerns the very form of the relation between the variables in question. It requires, for instance, that the variables entering into such a relation have some kind of quantitative structure. Surely, this is not an auxiliary assumption, but a substantive theoretical one, even though it is brought in from the modeling perspective rather than dictated by the theory. Thus, what is usually called a translation of a theory into a model is more accurately described as an exchange process; the transference of structure works both ways and affects both the model and the theory. The structure that, in the end, meets the data is a *merger* of theory and model.

The situation as sketched above poses a problem for the analysis of theoretical concepts in psychology, as well as for the philosophy of science. Theorists who discuss the status of, say, general intelligence, have a tendency to neglect the way intelligence is conceptualized in formal systems. However, testing theories of intelligence against data requires important choices to be made, and it may well be that it is in these choices that the theoretical status researchers ascribe to intelligence becomes most salient. Likewise, philosophers of science like to see a kind of uniformity across different scientific research areas, because analyses of 'the' structure of science require a significant degree of abstraction from substantive theory. So, either science as a whole is to be conceived of in an empiricist fashion, or science as a whole is a realist enterprise. But it may be that some parts of scientific research are aptly described as realist, while other parts are more aptly described as empiricist. It may even be the case that, within a single discipline, some research strategies require realism about theoretical terms, while others resist such an interpretation.

Thus, philosophy of science, formal modeling approaches, and discussions on the theoretical status of psychological concepts, are highly relevant to each other; and one can doubt whether they can be studied in isolation.

1.4 Measurement models

An analysis of theoretical concepts as they meet the data in research requires that we analyze the mergers that result from the exchange of structure between theory and model. This, in turn, means we have to consider the viewpoint that is brought in from the modeling side, as well as the theory itself. And precisely because theoretical concepts are not uniquely tied to particular formalized concepts, we can expect that differences in formal models, and especially in the theoretical concepts they entertain, lead to differences in the status of the theoretical concepts that meet the data. In other words, intelligence-as-a-true-score may have a different meaning, and postulate a different ontology, than intelligence-as-a-latent-variable. The upshot of this line of thinking, of course, is that there is no such thing as 'the' meaning of intelligence – at least, not before the researcher has made a choice of model. For a substantial part of this meaning of theoretical terms is introduced by the chosen model and not by the theory.

How does a theoretical concept connect to the world? It is a dogma of empirical science that, at some point or another, a theoretical term must have something to do with observations. This is a risky formulation from a philosophical point of view, so I would like to neutralize possible philosophical quarrels about it right away. That theoretical terms be connected to observations does not imply that they must be defined in terms of observations, as the logical positivists demanded, and neither that observations must have immediate falsifying relevance for theoretical statements, as the falsificationists argued. It does not mean that observations are free of theory-ladenness, nor that we have the kind of incorrigible knowledge about sensory experiences that once went by the name of sense-data. Neither does it mean that there exists such a thing as objective knowledge about the world, or even that there is a final truth. It merely means that scientific tradition is such that a theorist, who invokes a theoretical concept, is expected to discuss possible observational implications of his theoretical concepts and the relations between them. He is not expected to come up with a ready-made set of hypotheses or an experimental setup that may falsify her theory. It is merely considered suspect to posit theoretical concepts with the accompanying note that no possible set of observations could ever be relevant to them. The scientific researcher is more or less obliged to think of ways to connect theories to data. It is in this sense that empirical science is empirical, and it is in this sense that theoretical terms are to be connected to data.

In psychology, models that take care of this connection are generally called 'measurement models'. Now I need to neutralize the strong connotations of the word 'measurement', for it has a tendency to get people up in arms (e.g., Michell, 1999). In psychology, the term 'measurement model' must not be interpreted as implying quantification. Measurement models may relate nominal observed variables to nominal latent variables, as is done, for instance, in latent class models, in which case quantification is achieved nor aspired. In psychology, the term 'measurement' is rather to be interpreted as an extended form of observation. Although measurement models are universally formal in character, they may be fully qualitative. Also, measurement models do not aspire to say everything there is to be said about people, so they do not try to 'catch people in numbers', as is sometimes thought. They do tend to abstract away from many features of human beings, and I suppose this is the reason that the above mistake is often made; however, the statement 'John is aggressive' just as well abstracts away from such features as the statement 'subject i is a member of latent class j' does. The purpose of measurement models in psychology is not necessarily to quantify, nor to yield a description of people that is in any way 'complete'. The purpose of measurement models is to connect theoretical concepts to observations, and it is for this reason that they are indispensable in psychology.

Because theoretical concepts are connected to observations through measurement models, these models should be the main focus of the analysis. Now, it may seem to the student of psychology (or even to the working researcher) that there is a considerable consensus on how psychological measurement should be conceptualized – or even that there *is* only one theory of psychological measurement. The main reason for this is that, just as textbooks on statistics tend to propagate a particular view of statistics as 'the' theory of statistics, thereby creating the false impression that statistics-is-statistics-is-statistics (Gigerenzer, 1993), textbooks on psychological measurement similarly display a psychometrics-is-psychometrics-is-psychometrics approach. For example, many psychologists are taught that reliability is an important feature of psychological tests, learn to equate the concept with the value of Cronbach's α, and adopt an attitude that can be described as 'the higher the better'; they are not informed of the highly distinct ways to conceptualize reliability (Lord & Novick, 1968; Mellenbergh, 1996; Brennan, 2002), of the existence of measurement models that may imply low internal consistency (Bollen & Lennox, 1989), or of the fact that there are arguments for abolishing the concept of classical reliability altogether (Lumsden, 1976). The case of reliability is not an exception. In fact, the very concept of psychological measurement, as well as the possible ways to address it, are the subject of enduring discussions among methodologists, mathematical psychologists, statisticians, and philosophers. Interestingly, these debates circle around the same themes as the ones we find in discussions between realists and empiricists in the philosophy of science.

For example, psychological measurement systems are often presented as methods for "measuring the degree of ability of the person" (Rasch, 1960, p.16). A whole research paradigm is based on this line of thinking, but for one of the most influential psychologists of the previous century, B.F. Skinner, this is a backward strategy. For Skinner, abilities and traits are not only fictions, they are useless fictions, impeding the progress of science: "Aqua regia has the ability to dissolve gold; but chemists will not look for an ability, they will look for atomic and molecular processes" (Skinner, 1987, p. 785). Jane Loevinger defends a diametrically opposed position, stating that what is at issue in psychological measurement is "...the validity of the test as a measure of traits which exist prior to and independently of the psychologist's act of measuring" (Loevinger, 1957, p. 642), where the term 'trait' is intended to designate exactly the kind of ability concept that Skinner finds ludicrous. Ebel (1956, p. 642-643) quotes[1] Kaiser (1960, p. 412), who is said to deplore this kind of "philosophically naive faith" which "went out of style in the nineteenth century", to underscore his own conviction that "those who think of a real trait which 'underlies' a test score" are philosophically immature, because they "...have not yet learned that realistic philosophy is productive mainly of verbal discourse, and that it must be shunned if mental measurement is to advance". Forty years after the major battles surrounding the cognitive revolution and the abandonment of behaviorism, the issue is still with us. Edwards & Bagozzi (2000, p.157) state that "we intend that constructs refer to phenomena that are real and exist apart from the awareness and interpretation of the researcher and the persons under study", but Michell (2000, p. 639) boldly qualifies a whole psychometric paradigm that is based on this line of thinking as resulting from a "methodological thought disorder".

Like the approaches in philosophy of science, the papers from which these quotes were taken are of a very general nature. There are treatises providing systematic overviews of realist, positivist, or empiricist stances toward psychological measurement (most notably Messick, 1981; 1989), but these tend to abstract away from the relation these viewpoints may have to the different models that can be used. Although now and then a reference to a specific type of model is made – for example, Michell (1999; 2000) makes a case for the additive conjoint model on the basis of philosophical considerations – there has, to my knowledge, been no systematic treatment of the way different measurement models relate to empiricist, realist, operationalist, or other philosophical viewpoints. The purpose of the present book is to study these connections. It will become apparent that statistical models are not philosophically neutral, as is sometimes thought; on the contrary, there turn out to be clear connections between measurement models and philosophical views. This implies that a researcher, who chooses a particular model to connect his theoretical terms to the observations, is taking a philosophical stance with respect to the status of the theoretical terms he uses. In fact, he not only chooses a model; he chooses a philosophy of science.

1.5 Outline of this book

The aim of this book is to evaluate different measurement models in terms of their connections to philosophical views, and to discuss the relevance of these views to psychology. I will focus on three measurement models that have been highly influential in the history of psychological measurement: classical test theory, latent variable theory, and representational measurement theory.

Chapter 2 covers the classical test model (Lord & Novick, 1968). Classical test theory is the most widely used model in psychology; it is the theory that provides well-known concepts like true scores and reliability. However, classical test theory

[1] In my opinion, this quote is out of context, for the sentence cited does not apply to realism about attributes, but to the conviction that attributes have an inherent scale of measurement, which is a completely different topic.

is an almost perfect instantiation of operationalism. The fact that operationalism is almost universally rejected by psychologists is, of course, inconsistent with the popularity of classical test theory, and it is argued that the conceptual framework of classical test theory is grossly misinterpreted in psychological research.

Chapter 3 examines the latent variable model, more specifically the generalized item response theory (GLIRT) model (Mellenbergh, 1994). This is actually a class of models which comprises, among others, the common factor model, the Rasch model, and the latent class model. It will be shown that these models require a realist interpretation of the latent structures they posit. Thus, a theorist who identifies a psychological concept with a latent variable buys into realism about theoretical terms. Special attention will be given to causal interpretations of latent structures, which prove to raise some interesting philosophical problems as well as psychological research questions.

Chapter 4 considers the representational measurement theory model as developed by Krantz, Luce, Suppes, & Tversky (1971). This model is, strictly speaking, more aptly characterized as an approach to measurement than as a model, and it is as philosophically explicit as it is mathematically rigorous. The central concept in representational theory is the measurement scale, which is widely known because of Stevens' (1946) typology of nominal, ordinal, interval, and ratio scales. It is argued that representational measurement theory implements an empiricist conception of measurement, and that measurement scales must be viewed as constructions. Therefore, a psychologist, who identifies a theoretical term with a scale, can no longer adhere to realism.

In Chapter 5, I will examine the relations between the different models. At a formal level, many such connections are known to exist from the psychometric literature. However, in terms of semantics, model interpretation, as well as ontology, these connections are not straightforward. In fact, I will show that whether any such relations can be taken to hold depends crucially upon the possibility to use what is known as a propensity interpretation of the probabilities figuring in the latent variable and true score models. If this interpretation is denied, the models must be viewed as strongly distinct. 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. The difference in theoretical status between true scores, latent variables, and measurement scales, remains regardless of the chosen probability semantics.

From Chapter 5, it will become evident that latent variable theory is the only model that explicitly addresses the question where variation in scores comes from. Second, it is the only model that explicitly incorporates the attribute to be measured in the formal structure of the model. And third, the relation between the attribute and the observations may be framed in terms of causality. These three ingredients are coupled in Chapter 6 to present an account of validity that is loosely inspired on the latent variable model, but can also be applied to other models. Validity is conceptualized in terms of a causal relation between the attribute to be measured and the observations. It will be argued that the primary source of the validity problem in psychological measurement is not that it is difficult to find out what is measured, but that it is difficult to find out what we intend to measure.

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 judgement 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 & 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 very attractive one. Actually, what was once an idea seems to have been transformed into a fact: There is no psychological measurement without error. This seems to be a safe position to take when applying psychological tests – after all, who would be so overconfident to claim that he could measure perfectly – but it is also counterintuitive. That is, if I endorse the item 'I like to go to parties', why would there necessarily be measurement error involved? Could it not be that I truly like to go to parties? What, then, *is* measurement error? Now, this question is not as easily resolved as it may seem to be. It is seductive to conclude that random error, for example, represents the impact of unsystematic, transient, factors on the observations (e.g., the subject had a headache at the testing occasion, or she was distracted by noise, etc.). However, we will see that this interpretation is not without problems in the classical test theory framework. More generally, it is exceedingly difficult to reconcile the formal content of classical test theory with common interpretations of terms such as 'random error'. The friction between intuitive interpretations of terms, and the way they are formally conceptualized, is particularly salient in the interpretation of classical test theory's central concept, the true score.

The true score is commonly introduced by using phrases such as "the true score is the construct we are attempting to measure" (Judd, Smith, & Kidder, 1991, p.49), or by stressing the distinction "between observed scores and construct scores (true scores)" (Schmidt & Hunter, 1999, p.189). This interpretation of true scores, as 'valid' or 'construct' scores, has been called the *platonic true score* interpretation (Lord & Novick, 1968, p. 39 ff.). Of course, the use of the adjective 'true' strongly invites such an interpretation, and as a consequence it is endorsed by many researchers and students. However, problems with the platonic interpretation of true scores have been exposed by several authors (Klein & Cleary, 1967; Lord & Novick, 1968; Lumsden, 1976). In particular, cases can be constructed where equating the true score with the construct score leads to violations of basic theorems in classical test theory. In these cases, the identification of true and construct scores will, for example, lead to correlations between true and error scores (Lord & Novick, 1968; Lumsden, 1976), while in the classical test theory model, these correlations are zero by construction.

These observations point to the conclusion that the conjunction of the platonic true score interpretation with the axiomatic system of classical test theory is, at least for some cases, untenable. The implication of such a conclusion would be that, in general, the true score does not admit a realist interpretation. It is argued here that this is indeed the case. Further, the factors that preclude such an interpretation are elucidated. It is argued that the problems can be traced back to the fact that the true score is syntactically defined in terms of a series of observations. This severely restricts the interpretation of the concept; for instance, the true score does not lend itself to an identification with Loevinger's (1957) traits, which are presumed to exist independently of the test scores. The reason for this is that true scores are conceptualized in terms of observed scores, and, as a result of the way classical test theory is constructed, have a highly restricted domain of generalization – namely, the domain of parallel tests. It is, however, also argued in this chapter that the entire idea, that two distinct tests could be parallel, is inconsistent. This essentially forces the conclusion that the true score can only apply to the test in terms of which it is defined. This, in turn, implies that a conceptualization of psychological constructs as true scores requires an operationalist position with regard to such constructs.

2.2 Three perspectives on the true score

The psychometric models discussed in this book are viewed from three perspectives: Formal, empirical, and ontological. The formal perspective consists of two parts. First, the model formulation, or syntax, is discussed. Second, the interpretation of the formal terms in the model, i.e., the model semantics, is evaluated. After clarifying the syntax and semantics of the model, I discuss it from an empirical perspective, by examining the way the model handles data in actual research. Finally, the ontological stance evaluates whether psychometric concepts such as the true score can be taken to refer to an external, objective reality, or must be considered to be products of the imagination of the researcher.

In the context of classical test theory, the formal stance will focus mainly on the syntactical definitions of true and error scores, which form the basis of the theory. The semantic interpretation of these concepts immediately takes us into philosophical territory, because it must be framed in terms of counterfactual premises. Specifically, classical test theory must rely on a thought experiment to establish a version of probability theory that applies to the individual subject; this version of probability theory is needed for a consistent interpretation of the true score. From an empirical perspective, the thought experiment does heavy work in the interpretation of concepts such as reliability. But from an ontological perspective, the fact that the true score is defined in purely syntactic terms, and moreover requires an interpretation in terms of counterfactuals, severely limits the interpretation of the concept. It is argued here that the true score is better conceptualized as an instrumental concept, that governs the interpretation of data analytic results in test analysis, than as an entity that exists independently of the researcher's imagination.

2.2.1 The formal stance

Syntax Classical test theory is syntactically the simplest theory discussed in this book. Virtually all theorems follow from just two definitions. First, classical test theory defines the true score of person i, t_i, as the expectation of the observed score X_i over replications:

$$t_i \equiv \mathcal{E}(X_i). \tag{2.1}$$

Second, the error score E_i is defined as the difference between the observed score and the true score:

$$E_i \equiv X_i - t_i. \tag{2.2}$$

The notation emphasizes that, while X_i and E_i are considered random variables, the true score t_i is by definition a constant. Note that the error scores have zero expectation by construction, since $\mathcal{E}(E_i) = \mathcal{E}(X_i - t_i) = t_i - t_i = 0$.

An extra source of randomness is introduced by sampling from a population of subjects. As a result, the true score also becomes a random variable and the theory generalizes to the familiar equation

$$X = T + E. \tag{2.3}$$

Lord & Novick (1968, p.34) note that no assumption concerning linearity needs to be made in order to derive Equation 2.3. The linear relation between true scores and observed scores follows directly from the definitions of true and error scores. Novick (1966) showed that all other required assumptions follow from the definitions of true and error scores for the individual, as given in Equations 2.1 and 2.2. For example, the above definitions ensure the independence of true and error scores, and imply that the error scores have zero expectation in the population (Mellenbergh, 1999).

Semantics The true score is defined as the expected value of the observed scores. However, the interpretation of the expectation operator immediately yields a problem, because the expected value of the observed score is conceived of at the level of the individual. This conceptualization is borrowed from the theory of errors (Edgeworth, 1888; see also Stigler, 1986, and Hacking, 1990), which has been fruitfully applied, for example, in astronomy. It is useful to briefly summarize this theory.

The theory of errors works as follows. Suppose that one wants to determine the position of a planet, and that the planet is sufficiently distant for its position to be considered a constant. Suppose further that multiple measurements of its position are made. These measurements, if made with sufficient precision, will not yield identical values (for most readers, this will not come as a surprise, but it was originally considered to be a tremendously shocking discovery; see Stigler, 1986). Now, the deviations from the true value may be interpreted as accidental disturbances, that is, as the aggregated effects of a large number of independent factors (e.g., weather conditions, unsystematic fluctuations in the measurement apparatus used, and the like). It is intuitively plausible that, if this is indeed the case, the observations will tend to produce a symmetrical, bell-shaped frequency distribution around the true value: Because they are accidental, deviations to either side of the true value are equally likely, and, further, larger deviations are less likely than smaller ones. A formal justification for this idea can be given on the basis of the central limit theorem, which states that the sum of independently distributed variables approaches the normal distribution as the number of variables of which it is composed gets larger. Indeed, in the context of astronomical observations, the repeated measurements were often observed to follow such a bell-shaped frequency distribution. The theory of errors combines these ideas: It conceptualizes accidental disturbances as realizations of a random error variable, which will produce a normal distribution of the observations around the true value. If this conceptualization is adequate, then it follows that random errors will tend to average out as the number of observations increases. Thus, in such a case it is reasonable to assume that the expectation of the errors of measurement equals zero. This, in turn, supports the use of the arithmetic mean over a series of measurements as an estimate of the true position, because the mean is defined as the point for which the sum of the deviations from that point equals zero. It takes but a small step to conceptualize the true position of the planet as the expected value of the measurements, for which the arithmetic mean is a maximum likelihood estimator.

If classical test theory dealt with series of repeated measurements for which an analogous line of reasoning could be maintained, there would be few problems in the interpretation of the theory. However, classical test theory does not deal with such series of measurements, but with measurements on a single occasion. Moreover, series of measurements for which the theory holds are not to be expected in psychological measurement. Such series must satisfy the axioms of classical test theory, which require that the replications are parallel. In a realistic interpretation, this would mean that replicated observations should be considered to originate from a stationary random process; Molenaar (personal communication) has observed that, in the terminology of time series analysis, one would refer to the observed score as a 'white noise' variable with nonzero expectation. A procedure that would approximately satisfy the assumptions involved could, for example, consist in repeatedly throwing dice. That throwing dice would conform to the requirements of classical test theory is no coincidence, for what is in fact required is a procedure that allows for the application of the probability calculus in a frequentist sense. In the context of psychological measurement, the stated assumptions are unrealistic, because human beings will remember their previous response, learn, get fatigued, and will change in many other ways during a series of repeated administrations of the same test. Thus, even if the observed scores could be appropriately characterized as originating from a random process (which could be doubted in itself), this random process would not be stationary, which implies that the repeated measurements would not be parallel. It is clear, therefore, that classical test theory a) is not concerned with series of measurements, and b) could not concern itself with such series in the first place, because *actual* repeated measurements cannot be expected to conform to the assumptions of the theory. Still, the syntactical formulation of the theory uses the expectation operator at an essential point in the development of the theory – namely in the definition of its central concept, the true score. What is to be done about this awkward situation?

Introducing Mr. Brown It is useful to put oneself in Lord & Novick's shoes in order to appreciate the problems at hand[1]. First, Lord & Novick want to use a probability model based on Kolmogorov's (1933) axioms, but are unable to give this model a strong frequentist interpretation, which would make it comply with the dominant view of probability at the time (e.g., Neyman & Pearson 1967), because no actual series of repeated measurements will allow for such an interpretation. A subjectivist interpretation (De Finetti, 1974) is conceptually difficult; of course, the true score of subject i could be conceptualized as the expected value of the researcher's degree-of-belief distribution over the possible responses of subject i, but this view will not match the average researcher's idea of what constitutes a true value. For example, in psychological testing, the researcher will often not have any knowledge of subject i prior to test administration. In such cases, the Bayesian view would motivate the use of a noninformative prior distribution, which would moreover be the same across subjects. But this would imply that every subject has the same true score prior to testing. This is not unreasonable within the Bayesian paradigm, but it is squarely opposed to the way the average researcher thinks of

[1] The development commented on here can be found in Lord & Novick, 1968, Chapter 2.

measurement[2]. As a consequence, the application of the probability calculus has to be justified in a different manner.

Second, Lord & Novick want to reason along the lines of the theory of errors, but they cannot do this because the *assumption* that errors will average out in an *actual* series of repeated observations, and that the arithmetic mean of that series will therefore be a reasonable estimate of the theoretical construct in question, is in flagrant contradiction with the basic fact that human beings, unlike coins and dices, are capable of learning and inclined to do so. Moreover, Lord & Novick do not want to restrict the theory to continuous variables with normally distributed error scores, which, in the theory of errors, are critical for motivating the interpretation of the expected value as the true score. On the contrary, they want to generalize the theory to categorical observed variables, because, in psychological testing, these are far more common than continuous observed variables. For example, intelligence tests work with items that are scored dichotomously (as correct of incorrect), and Lord & Novick surely want their theory to cover such situations.

Third, Lord & Novick need to do something with the individual, but this does not mean that they want to take such an undertaking serious. Classical test theory has no business with the peculiar idiosyncratic processes taking place at the level of the individual: The probability model is merely needed to allow for the formulation of concepts such as reliability and validity, both of which are defined at the population level. A serious attempt at modeling individual subjects (e.g., through time series analysis) would, in all likelihood, not even yield results consistent with classical test theory. So, the subject must receive a probability distribution, but only in order to make him disappear from the analysis as smoothly as possible.

Lord & Novick's response to these problems may either be characterized as a brilliant solution, or as a deceptive evasion. In either case, their approach rigorously disposes of all problems in a single stroke: Lord & Novick simply delete subjects' memory by brainwashing them. Naturally, they have to rely on a thought experiment to achieve this. This thought experiment is taken from Lazarsfeld (1959):

> 'Suppose we ask an individual, Mr. Brown, repeatedly whether he is in favour of the United Nations; suppose further that after each question we 'wash his brains' and ask him the same question again. Because Mr. Brown is not certain as to how he feels about the United Nations, he will sometimes give a favorable and sometimes an unfavorable answer. Having gone through this procedure many times, we then compute the proportion of times Mr. Brown was in favor of the United Nations.' (Lazarsfeld, 1959; quoted in Lord & Novick, 1968, pp. 29-30)

Through the application of this thought experiment, the replications are rendered independent as a result of the brainwashing procedure. The resulting hy-

[2] Application of Bayes' theorem would also involve a term denoting the expected value of the observed score, conditional on the true score, and it is not unlikely that the interpretation of this term would still require a thought experiment similar to Lord & Novick's (1968, p.29), to be described hereafter. In this context, it is interesting that Novick, Jackson, & Thayer (1971) do not address this issue, while Novick & Jackson (1974) seem to retain this thought experiment in their Bayesian account of test theory

pothetical series of observations allows for the application of standard probability theory, a quasi-frequentistic conception of probability, and a syntactical definition of the true score which has at least *a* semantic interpretation: In the particular case of Mr. Brown, the true score equals the probability of him giving a favorable answer, which is estimated by the proportion of times he was in favor of the United Nations.

Propensities? Interestingly, Lord & Novick call the probability distribution characterizing this counterfactual series of replications a *propensity* distribution. This may be after Popper (1963), who proposed the propensity theory of probability as an objectivist alternative to Von Mises' conception of probability as relative frequency (Van Lambalgen, 1990). The propensity view holds that probability is not a relative long run frequency, but a physical characteristic of an object like a coin, or, more accurately, of the object and the chance experimental setup (Hacking, 1965). Lord & Novick's reference to the propensity view is remarkable because, in the thought experiment, they seem to introduce a limiting frequency view of probability. However, the limiting frequency and propensity interpretations of probability do not, in general, coincide. This is because propensities, by themselves, do not logically entail anything about relative frequencies. For example, a coin may have a propensity of .5 to fall heads; then it is possible, although perhaps unlikely, that it will forever fail to do so. In this case, the limiting relative frequency equals zero and thus deviates from the propensity. Because propensities are, in contrast to relative frequencies, logically disconnected from empirical observations, but are nevertheless supposed to conform to Kolmogorov's axioms, they have been said to operate under the 'conservation of mystery' (Kelly, 1996, p. 334). So, strictly speaking, the true score as a limiting frequency in the thought experiment is not logically connected to the true score as a propensity, because the propensity view and the relative frequency view are not logically connected.

Thus, Lord & Novick's reference to the propensity interpretation of probability is intriguing, especially in view of the fact that they are going through so much trouble in order to generate a relative frequency interpretation for the observed score distribution. One reason for their referencing the propensity view may be that it is the only objectivist theory of probability that allows one to ascribe probabilities to unique events. It is not improbable that Lord & Novick mention the term 'propensity' because they are aware of the fact that they are actually doing just this, and therefore cannot use a relative frequency account. But why, then, introduce the thought experiment in the first place? Why not settle for the propensity interpretation and let the relative frequencies be?

My guess is that the reason for this move is twofold. First, propensities are logically disconnected from relative frequencies (i.e., they are not defined in terms of such frequencies), but they are not fully disconnected either. It is in fact obvious that the propensity of a coin to fall heads is related to its behavior in repeated coin tossing. One could say that propensities should be viewed as dispositions to behave in a certain way; a propensity of .5 to fall heads, as ascribed to a coin, could then be viewed as expressing the conditional 'if the coin were tossed a large number of times, the relative frequency of heads would approximately be .5'. Because ascribing a disposition generally involves a prediction of this kind, Ryle (1949) has called dispositional properties 'inference tickets'. So, if Mr. Brown's true score is to be conceptualized in a similar way, the frequency behavior for which it would be an inference ticket must involve replicated measurements. Actual replicated measurements, however, are not generated by stationary random processes, and so it is likely that the propensities will not predict the actual relative frequencies at all. This would render Ryle's inference ticket useless. The inference ticket would, however, apply to the replicated measurements with intermediate brainwashing.

Second, we must not forget that Lord & Novick are forging an account of psychological measurement; and although they know that they cannot follow the line of reasoning that is the basis for the theory of errors, they do want to stay close to it. The theory of errors is clearly based on an observation concerning the behavior of scores in a long run of replicated measurements. Moreover, it is essential for these series themselves that they are unsystematic, i.e., that they are random. If they were not, there would be little reason to attribute the fact, that repeated measurements are not identical, to unsystematic fluctuations, and to view such disturbances as random error. Again, actual replications are unlikely to produce such series; these will neither be stationary, nor random. Hence, the need for Mr. Brown's being brainwashed inbetween the replications.

The conclusion must be that Lord & Novick do not need the thought experiment for the application of the probability calculus itself; this could be done solely on the basis of the propensity view. Moreover, the propensity view seems more appropriate because classical test theory is largely concerned with probability statements concerning unique events. Lord & Novick need the thought experiment to maintain the connection between probability theory and the theory of errors, that is, to justify the definition of the true score as the expected value of the observed scores, and to defend the view that deviations from that value are to be interpreted as to random error.

Thought experiments The brainwashing thought experiment could be cal-led successful, for it is used in many psychometric models. Models that use it are said to follow a *stochastic subject* interpretation (Holland, 1990; Ellis & Van den Wollenberg, 1993). A stochastic subject interpretation of psychometric models must, in general, rely upon a thought experiment like the above. The thought experiments are needed to provide an interpretation that is in line with both the probability calculus and the typical idea of random error, and could be said to function as a 'semantic bridge'. This property distinguishes them from other kinds of thought experiments, which are usually directed at a theory, rather than part of a theory (Brown, 1991; Sorensen, 1992). For this reason, it has been proposed to treat these thought experiments as a distinct class of 'functional' thought experiments (Borsboom, Mellenbergh, & Van Heerden, 2002-a[3]).

Classical test theory requires such a functional thought experiment, but this does not mean that it must take the particular form in which Lord & Novick present

[3] this paper is included in this dissertation as Appendix A

it. Any thought experiment that provides an interpretation consistent with the syntax of the theory could, in principle, do. Rozeboom (1966-a, p.387) considers, for example, that "we may fantasize an experiment in which each member i of P has been replicated p times and each replica (...) is tested (...), so that if p is large the frequency of a particular observed value X among among i's replicas approaches the probability of this observed score for i". This thought experiment thus considers a probability distribution over a very large number of replicas of Mr. Brown, every one of which is asked whether he is in favor of the United Nations. Still another form of the thought experiment is in terms of an infinite series of administrations of distinct parallel tests. In this case, we would not ask Mr. Brown the same question repeatedly, but we would present him with different questions that are parallel to the original question, that is, with a series of questions that all have the same expected value and error variance as the original question. Probably, many other forms of the thought experiment could be imagined. These thought experiments have in common that, as Rozeboom (1966-a, p. 385) puts it, they "try to convey some feeling for how sense can be made of the notion that a given testing procedure determines a probability distribution over potential test scores specific to each individual who might so be tested". It should be noted, however, that such thought experiments do little more than convey some feeling. Basically, the classical test theorist is trying to sell you shoes of which it is already obvious that they are three sizes too big.

How definitions replaced assumptions Lord & Novick swiftly go over the construction of true and error scores based on this thought experiment, and manage to dispose of the individual subject in exactly six pages (Lord & Novick, 1968, p. 28-34). In the remainder of their treatment of classical test theory, the focus is on between-subjects results and techniques. At the basis of the theory, however, remains the true score, defined through this peculiar thought experiment.

It is illustrative to recapitulate what has happened here. Lord & Novick have managed to put the theory of errors on its head. Recall that this theory is based on the idea that accidental errors will average out in the long run. The statistical translation of this notion is that accidental error scores can be viewed as realizations of a random variable with zero expectation. The zero expectation of measurement errors must therefore be viewed as an assumption (i.e., its truth is contingent upon the actual state of affairs in the world). On the basis of this assumption, the expectation of the measurements can be conceptualized as an estimate of the true score. Since Lord & Novick are not in a position to use anything resembling an actual series of replications, and therefore are not in possession of a suitable long run, they create one for themselves. However, because their long run is constructed on counterfactual premises, it must remain thought experimental. It is obvious that, upon this conceptualization, the zero expectation of error scores can no longer be taken serious as an assumption, because it applies to a counterfactual state of affairs. As a result, there is no empirical basis for taking the expected value of the measurements as an estimate of the true score. Now, Lord & Novick's response to this problem is remarkable. Instead of taking the zero expectation of errors as an *assumption* on which one can base the *hypothesis* that the expectation of the observed scores is equal to the true score, they *define* the true score as the expected value of the observed scores and then *derive* the zero expectation of errors as a consequence. Where the theory of errors observes irregularities in measurement, and then proposes statistical machinery to deal with those, classical test theory proposes the statistical machinery, and then hypothesizes the irregularities that would conform to it. The identity of expected observed score and true score is thus transformed from a hypothesis into a definition; and the assumption that error scores have zero expectation becomes a necessary truth. Following these moves, one can see the circle close: The theory becomes a tautology. The price that is paid consists in the fully syntactical definition of the true score.

2.2.2 The empirical stance

If the applications of classical test theory were as esoteric as its theoretical formulation, nothing could be done with it. However, classical test theory is without doubt the most extensively used model for test analysis. What, then, does it actually do in test analysis? How does it relate to empirical data?

At this point, it is important to distinguish between how the classical model could be used in test analysis, and how the model is typically used. The basic axioms of classical test theory imply nothing about the data, and are therefore permanently immune to falsification: The adequacy of the posited decomposition of observed scores in true and error scores cannot, for any given item, be checked. Thus, this part of the model is untestable. This does not mean, however, that classical test theory could not be used to formulate testable hypotheses at all. However, to formulate such hypotheses requires extending the model with additional assumptions. These additional assumptions concern relations between true scores on different test forms, or items. Three such relations are commonly distinguished: parallelism, tau-equivalence, and essential tau-equivalence. Two tests x and x' are parallel in a population if they yield the same expected value and the same observed score variance for every subpopulation (including subpopulations consisting of a single subject). If distinct tests are assumed to be parallel, they must have equal means and variances; in addition, all intercorrelations between tests must be the same. Two tests are tau-equivalent if they yield the same expected values, but different error variances; and they satisfy essential tau-equivalence if they neither yield identical expected values, nor identical observed score variances, but the expected values are linearly related through the equation $\mathcal{E}(X) = c + \mathcal{E}(X')$, where c is constant over persons. For a given set of items, all three of these relations can readily be tested. For example, as Jöreskog (1971) has observed, when the classical model is extended with any one of the above relations, the model can be formulated as an identified factor model, and the implied covariance matrix can be fitted to the observed covariance matrix. Thus, commonly invoked assumptions about relations between true scores do have testable consequences. At least some parts of the so extended model could be tested.

This is how the model *could* be applied. It is safe to say, however, that classical test theory is never applied in this way. The common applications of classical test theory do not involve testing the model assumptions. The cause of this neglect is probably historical, but will not concern us here. Rather, we will be concerned with the function classical test theory fulfills in applications. The strategy that is followed is highly indirect, and works via the estimation of reliability. It is important to review this process extensively, for it contains the basis for many misinterpretations of what classical test theory is about.

Reliability

Reliability is a population dependent index of measurement precision (Mellenbergh, 1996). It indicates the fraction of observed variance that is systematic, as opposed to random, in a given population. In classical test theory, reliability is the squared population correlation, ρ^2_{XT}, between true and observed scores. This equals the ratio of true score variance to observed score variance:

$$\rho^2_{XT} = \frac{\sigma^2_T}{\sigma^2_X} = \frac{\sigma^2_T}{\sigma^2_T + \sigma^2_E}. \tag{2.4}$$

This equation has intuitive appeal: In a given population, the value of the reliability coefficient will decrease as the error variance increases. If there is no error variance, reliability is perfect and equals unity. Note that this definition of reliability is population dependent (Mellenbergh, 1996). The reason for this is that reliability is defined in terms of the population model in Equation 2.3. This is reflected in the random variable notation for the true score in the definition of reliability, i.e., in Equation 2.4 the true score is denoted as T and not as t. A well-known implication of this definition is that reliability becomes smaller, if the true score variance in a population approaches zero while the error variance remains constant. As a consequence, for any individual subject i the reliability of a test equals zero, because by definition $\sigma^2_{t_i}$ equals zero for all i. Because reliability is a population dependent concept, it can be meaningfully considered only when interpreted in terms of individual differences in a specific population.

Of course, the formula for reliability contains the true score, which is unobservable. The conceptual strategy of classical test theory consists in rewriting the formula for reliability in terms of potentially observable terms. Lord & Novick (1968) discuss the matter on p. 58-59; what follows here could be viewed as a conceptual reconstruction of this development.

First, suppose that we had the ability of brainwashing subjects inbetween measurements. In this case, determining reliability would pose no difficulties. The determination of true score variance would still be impossible at any given time point, but because replications would be parallel by definition, we could use the correlation between the observed scores on two administrations, X and X', as follows. Assume, without loss of generality, that the expected value of the test scores in the population is zero. The correlation between the observed scores at two time points would equal:

$$\rho_{XX'} = \frac{\sigma_{XX'}}{\sigma_X \sigma_{X'}} = \frac{\mathcal{E}(TT')}{\sigma_X \sigma_{X'}}. \tag{2.5}$$

See Lord & Novick, 1968, p. 58, for the details of the derivation. This almost equals Equation 2.4, which defines reliability. All that remains to be done is to rewrite the term $\mathcal{E}(TT')$ as σ_T^2, and the term $\sigma_X\sigma_{X'}$ as σ_X^2. If this step can be justified, the quantity σ_T^2/σ_X^2, which is unobservable in principle, has been rewritten as the quantity $\rho_{XX'}$, which is observable in principle. This would create a possible connection to the analysis of empirical data. Thus, what we have to do is to interpret a covariance between two variables as the variance of a single variable, and the product of two standard deviations of different variables as the variance of a single variable. This requires that the two variables in question are one and the same. That is, we need to be able to say not only that $T = T'$, in the sense of being numerically equal, but that $T \equiv T'$, in the sense that T and T' are synonymous. The reason for this is not primarily syntactical: $\rho_{XX'}$ will be numerically equal to ρ^2_{XT} as soon as the true scores and error variances on two tests x and x' are numerically equal for each subject, even if this is by accident. For a consistent interpretation of the theory, however, these quantities have to be equal by necessity.

As an illustration of this point, consider the following situation. Suppose that it were the case that height and weight correlated unity in a population of objects, and that these attributes were measured on such a scale that the expected value of the measurement of weight with a balance scale, and the expected value of the measurement of length with a centimeter, happened to always be numerically equal. One could then use the correlation between height and weight as an estimate of the reliability of the balance scale. As a pragmatic empirical strategy, this could work. But theoretically, one cannot admit such a situation in definitions and derivations like the above, because it would not be a necessary, but a contingent fact that the expectations of the measurement procedures were equal; they might very well not have been. Thus, from a semantic perspective, equating the correlation between parallel tests with the reliability of a single test makes sense only if the two tests measure the same true score. This requires that the true scores on the first and second administration are not merely numerically equal, but synonymous.

Can we take the required step while retaining a consistent semantic interpretation of the theory? It is one of the intriguing aspects of classical test theory that this can be done. The reason for this is that the true scores in question are not only syntactically, but also semantically indistinguishable. This is because, for subject i, both t_i and t'_i are defined as the expected value on test x, where the expectation is interpreted in terms of repeated administrations with intermediate brainwashing. It may seem that, because t_i is the expected value of the observed scores on the first administration of test x, and t'_i is the expected value of the observed scores on the second administration of test x, t_i and t'_i are distinguishable with respect to their temporal position. But the role of time in the brainwashing thought experiment is a peculiar one. The thought experiment uses the term 'replications' in order to make the application of the expectation operator to the individual subject a little more digestible than it would otherwise be, but the idea that we are talking about replications in the actual temporal domain is an illusion. This may be illustrated through the classical test theory models for change (Mellenbergh & Van den Brink, 1998). In such models, the difference between subject i's observed scores on administrations 1 and 2 of the same test, $X_{i2} - X_{i1}$, must be considered to be an estimator of i's true gain score, defined as $t_{i2} - t_{i1}$. Each of the true scores is thus defined as the expected value at a single time point. Although the thought experiment creates the impression that the expectation can be interpreted in terms of temporally separated replications of the same test, the term 'brainwashing' must be taken to mean that the subject is restored to his original state – *not only with respect to memory, learning, and fatiguing effects, but with respect to time itself.* Otherwise, classical test theory concepts such as the true gain score would be completely uninterpretable. Within the brainwashing thought experiment, the true scores on replications must be considered synonymous. Thus, Lord & Novick are justified in stating that $T \equiv T'$, and are able to write

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2} = \rho_{XT_X}^2, \tag{2.6}$$

which completes the first part of their mission.

Obviously, the development sketched above only takes us halfway in making the connection between classical test theory and the analysis of empirical data. What we want is not to express reliability in terms of counterfactual relations, which involve brainwashing entire populations, but to express it in terms of actual relations between observed variables in real data. So, Lord & Novick's brainwash has had its best time; it has been crucially important in deriving the main psychometric concepts in classical test theory, but now it has to go. Can we get rid of it? The answer is: yes and no. An exact estimate of reliability cannot be obtained from empirical data, so in this sense there is no way to get around the issue. We can, however, settle for lower bounds on reliability, which can be estimated from the data under rather mild conditions. In the final analysis, however, the true score must be invoked again to conceptualize what such a lower bound is a lower bound *for*.

Constructing empirical estimates of reliability

The first option for constructing estimates of reliability is to neglect the conditions, that preclude the interpretation of actual repeated measurements as identical with the thought experimental replications, by simply ignoring the problem. This can be done in two ways: either we may assume that two actual replications of the same test are parallel, or we may assume that two distinct tests are parallel. The first of these methods is known as the test-retest method, and the second forms the basis of the parallel test method, the split-halves method, and the internal consistency method.

Test-retest reliability The test-retest method is based on the idea that two administrations of the same test may be regarded as one administration of two parallel tests. If this were the case, the population correlation between the scores on these administrations would be equal to the reliability of the test scores. However, the assumption that repeated administrations are parallel introduces a substantial assumption into the technicalities of classical test theory, namely that the trait in question is stable. On the basis of this observation, it has been suggested that the test-retest correlation should be called a 'stability coefficient'. It should be noted, however, that the between-subjects correlation cannot distinguish between situations where individual true scores are stable and situations where they increase or decrease by the same amount. Therefore, the term 'stability' can only be taken to refer to the stability of the ordering of persons, not to the stability of the construct itself. Note also that the method necessarily confounds differential change trajectories and unreliability. We do not know, for most constructs, whether change trajectories are homogeneous or heterogeneous across subjects. This, of course, poses a problem for the interpretation of the test-retest correlation as a reliability estimate.

A second problem is that, in contrast to the thought experimental replications, actual replications are temporally separated, which creates the problem of choosing an appropriate spacing of the replications. Is reliability to be estimated by test-retest correlations based on immediate retesting? Retesting after a day? A month? A year? Since classical test theory cannot provide an answer to these questions, the test-retest scheme must introduce decisions which are, from a methodological perspective, arbitrary. However, these arbitrary decisions concerning the spacing of the replications will generally influence the value of the test-retest correlation. Does this mean that there is a distinct reliability for each choice of temporal spacing? Or should we consider the approximation to reliability to be systematically affected by temporal spacing, so that, for example, the estimate becomes better as we wait longer with retesting? Or does the approximation decrease with the time elapsed since the first administration? Or is this relation curvilinear so that, for example, the approximation is optimal after 1.2 weeks? And should we consider the relation between the quality of the reliability estimate and elapsed time to be the same across testing situations? Across groups? Across constructs? Why? It seems that these issues cannot be satisfactorily addressed, either from a psychological, a philosophical, or a methodological perspective.

In view of these issues, it is interesting that the test-retest method has recently been defended by Brennan (2001), on the grounds that reliability is intelligible only when interpreted in terms of replications of full test forms. This is plausible, but the concept of reliability should be considered within the definitions of classical test theory. Classical test theory defines the true score in terms of a thought experiment, and since the syntactical notation of reliability contains the true score as one of its elements, this definitional issue carries over to the interpretation of reliability. Upon a consistent interpretation of classical test theory, reliability is the proportion of variance in observed scores that would be attributable to variance in true scores; for the test-retest correlation to be an estimate of this proportion, the entire population of subjects must be brainwashed inbetween repeated administrations. Therefore, reliability must conceptually be interpreted in terms of the brainwashing thought experiment; it cannot be defined in terms of actual replications because these simply will not behave according to the axioms of classical test theory. Practically, of course, one may suppose that the actual test-retest correlation is an estimate of the thought experimental one, but in this case it has to be assumed that relevant characteristics of the thought experimental replication are retained in an actual replication. Unfortunately, the essential characteristics involve parallelism and independence of repeated measurements, i.e., the assumption that the replications could be viewed as realizations of a stationary random variable. This is extremely unrealistic. Thus, the interpretation of the test-retest correlation as reliability (i.e., as the concept is defined in classical test theory through equation 2.4) requires a substantial leap of faith.

Using correlations between distinct tests The second strategy, which encompasses the methods of parallel tests, split-halves, and internal consistency estimates, is based on the idea that two distinct tests could be parallel. First, consider the parallel test method. This method assumes that a simultaneous administration of two different tests could be viewed as approximating two thought experimental replications of a single test. In case we had distinct parallel tests, the correlation between them could then be taken to be a direct estimate of the reliability of the test scores. There are two problems with this method.

The first is a practical problem, namely that the search for parallel test forms has been unsuccessful to date; this is not surprising, because the empirical requirements for parallelism (equal means, variances, and covariances of observed scores) are rather demanding. Further, there is no substantial psychological reason for assuming that two tests for, say, spatial reasoning, should have equal means and variances; nor is there a reason for regarding such tests as theoretically superior to tests that are not parallel.

The second problem is of a theoretical nature, namely that the idea that two distinct tests could be parallel seems semantically inconsistent. We have seen, in Section 2.2.1, that classical test theory interprets the true score on a test x as the expected value on a number of repeated independent administrations of that test. That is, the true score is explicitly defined in terms of the test in question. If we now turn to a distinct test y, the true score on this test is semantically interpreted in terms of repeated independent administrations of test y. Earlier in this section, we have seen that, to interpret the correlation between parallel test scores as a reliability estimate, the covariance between the two true scores on these measures must be interpreted as the variance of one true score, that is, it must be assumed that $T \equiv T'$. This can be done within the counterfactual state of affairs, defined in Lord & Novick's brainwashing thought experiment, exactly because T and T' are synonymous. However, the true scores on distinct tests x and y are semantically distinguishable, simply because they are defined with respect to different tests. They may be empirically equal, but this does not make them logically identical. This is to say that the identity of the true scores on repeated administrations with intermediate brainwashing, as used in the derivation of equation 2.4, is a necessary truth; but the empirical equality of expected values on distinct tests is a contingent truth (if it is a truth at all). This may be illustrated by noting that the former equivalence will hold by definition (one does not even have to administer the test to find out), while the observation that the latter holds in the present testing occasion does not guarantee that it will hold tomorrow.

The problem here is not so much that, as a hypothesis formulated independently of the classical test theory model, two distinct tests could not be taken to measure the same attribute; this hypothesis could certainly be added, and would in effect specify a latent variable model. The problem is rather that classical test theory itself has insufficient conceptual power to do the trick. The syntax of classical test theory cannot express what it means for two distinct tests to measure the same attribute, if the attribute is identified with the true score. It is only possible to write down, syntactically, that two tests measure the same true score. However, semantically, this makes sense only if these two 'tests' are in fact replicated administrations of the same test, as they are in the brainwashing thought experiment. But of course the brainwashing thought experiment is completely unrealistic. This is why the theory must take recourse to the strange requirement of tests that are distinct and yet parallel. What the syntactical derivations, as well as the semantics, of classical test theory imply is that parallel measurements consist in two independent administrations of the same test. A procedure that could reasonably be said to conform to the requirement of parallelism is, for example, the replicated length measurement of a number of rods with the same centimeter. With two distinct psychological items or test scores, however, this logic is, at best, artificial and contrived; at worst, it is inconsistent. Thus, it is difficult to see how the method could yield theoretically interesting results, since it seems built on a contradiction. It is also obvious that the method has no practical value, because tests that satisfy at least the empirical equivalence needed for exact reliability estimates to work, are hard to come by. The parallel test method is thus useful for only one purpose, namely for the derivation of reliability formulae. It cannot be taken serious as an empirical method.

In the pursuit of exact reliability estimates, two methods have been proposed that may serve as alternatives to the parallel test method. These are the split-halves and internal consistency methods. The split-halves method splits a test in two subtests of equal length, assumes that the subtests are parallel (or constructs them to be nearly so; Gulliksen, 1950; Mellenbergh, 1994), computes the correlation between the total scores on subtests, and yields an estimate of the reliability of total test scores by using the Spearman-Brown correction for test lengthening. Internal consistency formulae such as the KR_{20} and coefficient α extend this method. They can be interpreted as the average reliability coefficient as derived from the split-halves correlation, where the average is taken over all possible split-halves. If the split-halves are parallel, the resulting quantity yields an exact estimate of the reliability of the total test scores. Since parallelism is as troublesome for split-halves as it is for full test forms, these methods fail for the same reasons as the parallel test method.

Lower bounds The exact estimation of reliability from observed data is thus impractical and theoretically questionable. This has prompted classical test theorists to look at worst-case scenarios, and to search for lower bounds for reliability (Guttman, 1945; Jackson & Agunwamba, 1977). For instance, it can be proven that, if test forms are not parallel, but satisfy weaker assumptions such as essential tau-equivalence, reliability estimates like Cronbach's α provide a lower bound on reliability. Thus, if α equals .80 in the population, then the reliability of the test scores is at least .80. This is a clever strategy, and the researcher who follows it seems to be fairly safe. In essence, the reasoning which could be followed is: no matter how bad things may be, the reliability of my test is always higher than the (population) value of the lower bound that is computed. This is probably the most viable defense that could be given for the standard practice in test analysis.

Note, however, that the true score does not do any work in the computation of any of the statistics discussed. The test-retest correlation is, well, a test-retest correlation, and internal consistency is just a transformation of the average split-half correlation. Both could be used in test analysis, and judged for their merits, without recourse to classical test theory as a theory of measurement. The statistical machinery will do just fine. However, this does not mean that classical test theory is irrelevant to the way the analyses are used. For the *interpretation* of test-retest correlations or average split-halves correlations as reliability estimates does involve classical test theory. What is obtained in the analysis is a test-retest or average split-halves correlation, but when these are interpreted in terms of reliability, they are interpreted as estimates of, or lower bounds for, the quantity denoted as $\rho^2_{XT_X}$, and this quantity does involve the true score as defined in classical test theory. Thus, what we observe here is an inference from empirical relations (involving only observables) to theoretical relations (involving observables and unobservables). This type of inference is, of course, nothing new, for it is the gist of science. What is typical and unusual here, is that the inference does not come at a price. The researcher gets the theoretical interpretation in terms of unobservable true scores for free. The question, however, is what this theoretical interpretation is worth: What is it exactly, that we are informed about? What is the status of the true score?

2.2.3 The ontological stance

Of all psychometric concepts, reliability plays the most important role in practical test analysis. Of course, all researchers pay lip service to validity, but if one reads empirical research reports, reliability estimates are more often than not used as a primary criterion for judging and defending the adequacy of a test. In this sense, reliability is the poor man's validity coefficient, as Rozeboom (1966-a) has observed. I think that the analysis presented above casts doubt on whether reliability deserves this status. The theoretical acrobatics necessary to couple empirical quantities, like test-retest correlations, to reliability, as defined in classical test theory, are disconcerting. Coupled with the fact that these coefficients are used and interpreted rather uncritically, the observation that "classical measurement theory [is] the measurement model used in probably 95% of the research in differential psychology" (Schmidt & Hunter, 1999, p. 185) seems to be a cause for concern, not for celebration. The problems grow even deeper when one considers that 95% of the researchers involved in research in differential psychology are probably not doing what they think they are doing. For no concept in test theory has been so prone to misinterpretation as the true score.

As has been noted earlier in this chapter, it is tempting to think that the distinction between true scores and observed scores is the same as the distinction "between observed scores and construct scores" (Schmidt & Hunter, 1999, p.189), or that "the true score is the construct we are attempting to measure" (Judd, Smith, & Kidder, 1991, p.49), or that it is the score "that would be obtained if there were no errors of measurement" (Nunnally, 1978, p. 110). This is the way the matter is often explained to students, and it is the way many researchers think about psychological measurement. However, the identification of the psychological construct with the true score of classical test theory is not without problems.

There are two problematic assumptions underlying the platonic interpretation of the true score. The first assumption underlying the idea that the true score is the real score on a psychological construct is the result of a confound of unreliability with invalidity. This is a recognized fallacy, but it is so common and persuasive that it deserves a thorough treatment. The second assumption concerns the ontological status of the true score itself. It will be argued here that the entire idea that a person has a true score, as defined in classical test theory, is unintelligible – except when interpreted in a thought experimental sense. So interpreted, it has the status of a dispositional concept, but, oddly enough, it specifies dispositional properties with respect to an impossible sequence of situations; namely, the thought experimental replications. The true score is therefore best thought of as a fiction. Finally, in contrast to psychological constructs, the true score cannot be conceptualized independently of the test in question. This is why the true score must be seen as a concept that is best interpreted in an operationalist sense.

True scores as construct scores

The idea that true scores are valid construct scores can be seen as a confound of reliability and validity. These are qualitatively different concepts: Reliability has to do with the precision of the measurement procedure, while validity involves the question whether the intended attribute is indeed being measured. For the simple reason that no formal model can contain its own meaning (it cannot itself say what it is a model for), it seems obvious that this interpretation is incorrect from the outset. However, although various authors have warned against it, the platonic true score interpretation is like an alien in a B-movie: No matter how hard you beat it up, it keeps coming back. A recent revival has, for example, been attempted by Schmidt & Hunter (1999; see Borsboom & Mellenbergh, 2002, for a criticism). True scores are not valid construct scores, and neither do they necessarily reflect construct scores.

At the present point in the discussion, the concept of validity is introduced, and therefore the relation of measurement has become important. In itself, it is interesting that, in the entire discussion so far, the term 'measurement' has remained unanalyzed. We have been able to review the assumptions, semantics, and empirical applications of classical test theory without making the meaning of this concept explicit. This is typical of classical test theory and contains an important clue as to why the identification of true scores with psychological constructs is so problematic. To see this, take it as given that the objective of psychological testing is to measure constructs, or, if you like, the phenomena to which constructs refer. If true scores could be taken to be identical to construct scores, then it should be possible for classical test theory to rewrite the relation of measurement, interpreted as a relation between observed scores and construct scores, as a relation between observed scores and true scores. It turns out that classical test theory cannot do this. The reason for this is that, because the theory is statistical in nature, it is natural to conceive of the relation between observed scores and construct scores statistically. This is also the position taken by Lord & Novick (1968, p. 20), who say that '... an observable variable is a measure of a theoretical construct if its expected value is presumed to increase monotonically with the construct' and '... to be primarily related to construct being defined'. This is similar to the measurement relation as conceived in item response models, where the expected value on items is related to the position on the latent variable. It follows from this conceptualization, however, that true scores cannot play the role of construct scores. This is because the true score is itself defined as the expected value on a test, so that identifying true scores with construct scores and substituting this in Lord & Novick's conception of measurement leads to the following definition: '... an observable variable is a measure of a [true score] if its [true score] is presumed to increase monotonically with the [true score]'. This can hardly be considered enlightening.

In contrast to, for example, latent variable models, classical test theory does not have the conceptual power to represent the construct in the model. The relation of measurement must thus be seen as a relation between true scores and something else. This is in perfect accordance with the way validity is treated in classical test theory, namely as the correlation between the true scores on the test in question and an external criterion. However, it is inconsistent with the idea that true scores are construct scores. It is actually rather strange that this misconception occurs at all, because classical test theory defines the true score without ever referring to psychological constructs or a measurement relation. The theory does not contain the identity of true scores and construct scores - either by definition, by assumption, or by hypothesis. Moreover, it is obvious from the definition of the true score that classical test theory does not assume that there is a construct underlying the measurements at all. In fact, from the point of view of classical test theory, literally every test has a true score associated with it. For example, suppose we constructed a test consisting of the items "I would like to be a military leader", "$.10/\sqrt{.05+.05} = ..$ ", and "I am over six feet tall". After arbitrary - but consistent - scoring of a person's item responses and adding them up, we multiply the resulting number by the number of letters in the person's name, which gives the test score. This test score has an expectation over a hypothetical long run of independent observations, and so the person has a true score on the test. The test will probably even be highly reliable in the general population, because the variation in true scores will be large relative to the variation in random error (see also Mellenbergh, 1996). The true score on this test, however, presumably does not reflect an attribute of interest. The argument shows that it is very easy to construct true scores that have no substantial meaning in terms of scientific theories, and are therefore invalid upon any reasonable account of validity.

It is also very easy to construct situations in which there is a valid construct score, while that score differs from the true score as classical test theory defines it. Consider, for example, the following example, which is based on an example by Lord & Novick (1968, p.39 ff.). At present, whether a patient has Alzheimer's disease or not cannot be determined with certainty until the patient is deceased and autopsy can be performed. In other words, the diagnostic process, taking place while the patient is still alive, is subject to error. We can conceptualize the diagnostic process as a test, designed to measure a nominal variable with two levels ('having the disease' and 'not having the disease'). Because this variable is nominal, we may assign an arbitrary number to each of its levels. Let us assign the number '1' to a patient who actually has Alzheimer's, and the number '0' to a patient who does not. This number represents patient i's construct score c_i on the nominal variable 'having Alzheimer's'. Thus, a patient who actually has Alzheimer's has construct score $c_i = 1$, and a patient who does not have Alzheimer's has construct score $c_i = 0$.

In practice, the construct score cannot be directly determined. Instead, we obtain an observed score, namely the outcome of the diagnostic process. This observed score is also nominal, so we may again assign an arbitrary number to each of its levels. Let us code patient i's observed score x_i as follows. The value $x_i = 1$ indicates the diagnosis 'having Alzheimer's', and the value $x_i = 0$ indicates the diagnosis 'not having Alzheimer's'.

The diagnostic process is imperfect and therefore the test scores are subject to error. Now suppose that the test is valid, so that misclassifications are due solely to random error, for example, to equipment failures that occur at random points in time. This renders the observed score a random variable X. What is the true score on the test? It is tempting to think that patient's i's true score, t_i, on the diagnostic test is equal to the construct score (i.e., $t_i = c_i$). Specifically, the infelicitous use of the adjective 'true' suggests that a patient who actually has Alzheimer's, i.e., a patient with construct score $c_i = 1$, also has a true score of $t_i = 1$ on the test. For this indicates the diagnosis 'having Alzheimer's', and it is, after all, true that the patient has that disease.

This interpretation of the true score is not, in general, consistent with classical test theory. For suppose that the sensitivity of the diagnostic test is .80. This means that the probability that a patient who actually has Alzheimer's will be correctly diagnosed as such is .80. Now consider the true score of a patient who has Alzheimer's , i.e., a patient with construct score $c_i = 1$. This patient's true score is not $t_i = 1$, because the true score of classical test theory is equal to the expectation of the observed score, which is $t_i = E(X_i \mid c_i = 1) = .80$. Suppose further that the sensitivity of the test is .70. This means that the probability that a patient who does not have Alzheimer's will be correctly diagnosed is .70. For a patient who does not have Alzheimer's (i.e., a patient whose construct score is $c_i = 0$), the true score is equal to $t_i = E(X_i \mid c_i = 0) = .30$. In both cases the true score and construct score yield different values[4].

It can now be seen why the identification of true scores with construct scores is logically inconsistent with classical test theory in general. If the test in the

[4] Note that the argument implicitly uses a latent class formulation, where the construct score indicates class membership; this suggests that latent variables can be used to extend the model in the required direction. It will be argued in the next chapter that this is indeed the case.

example contains error, this means that there is misclassification; and if there is misclassification, the expected value of the observed score can never be equal to the construct score. So, if measurements contain random error, the identification of true scores with construct scores is logically inconsistent with classical test theory in general. It should be noted that Lord and Novick (1968) themselves were thoroughly aware of this, since they explicitly state that "in general the two concepts and definitions [of true scores and construct scores] do not agree" (p. 41).

It is clear that the identification of construct scores with true scores is fundamentally incorrect. The objective of psychological measurement is to measure psychological constructs, but classical test theory cannot express the relation of measurement as a relation between observed and true scores. Rather, the theory must conceptualize the measurement relation as a relation between true scores and psychological constructs, which shows that these should not be considered identical. This conclusion is strengthened by the observation that we can easily construct cases where a true score 'exists', but where it is invalid in that it does not have substantial meaning in terms of a theory. We can also construct cases where there is a valid score, but where that score is not the true score.

In view of these problems, it is interesting and elucidating to inquire under what conditions the true score and the construct score could be taken to coincide. It seems that the situation, in which this would be the case, is exactly the situation as the theory of errors portrays it. Namely, if the validity of the test has been ascertained, the observations are continuous, the attribute in question is stable, and deviations from the true value over actual replications are produced by a large number of independent factors. In this case, the axioms of classical test theory will be satisfied by actual, rather than thought experimental, replications – in fact, there would be no need for a thought experiment. It also seems that the number of psychological measurement procedures, for which these assumptions could be taken to hold, equals zero. Thus, it is safe to conclude that, in psychological measurement, the true score cannot be taken to coincide with the construct score.

Do true scores exist?

The identification of true scores with constructs is a serious mistake that, unfortunately, permeates much of the literature on psychological measurement. The fact that true scores cannot be considered in this way does not, however, entail that true scores cannot exist. We may suppose that true scores and construct scores both exist, but are not identical; for example, we could imagine true scores to exist quite independently of the construct, but to be systematically related to that construct. This is the way Lord & Novick construct the relation of measurement, as we have seen, and it is also the way that latent variable models sometimes formulate the situation. The question then becomes how the existence of true scores could be interpreted. Is there a plausible interpretation that could locate the true score in reality, i.e., conceive of it as an objectively existing entity, without becoming inconsistent or downright absurd? It is argued in this section that such a realist interpretation is unreasonable. When the classical test theorist invites us to imagine the existence of a true score, most of us will be inclined to grant him this much. We will see, however, that it is completely unclear what we are supposed to imagine. The reason for this is that it is difficult, or even impossible, to give a serious account of the distribution on which the true score is defined. The problem is that the thought experiment, that is supposed to define this distribution, does not specify sources of random error, and that the almost universally endorsed interpretation of random error is circular. Moreover, the assumption that true scores exist in reality does not lead to testable predictions, which strongly invites the application of Occam's razor – especially because the true score leads to a needless multiplication of theoretical entities, which is undesirable.

Where does error come from? The conceptualization of the true score as an expected value is ill-defined. For it is entirely unclear under what circumstances the replications mentioned in Lord & Novick's brainwashing thought experiment should occur. The primary problem is that it is unclear where the random variation is supposed to come from. This issue is usually circumvented in treatises on psychological measurement. These suggest that random error is due to unsystematic factors affecting the observations. For example, the typical examples of unsystematic errors are: Mr. Brown had a headache at the particular testing occasion; Mr. Brown accidentally filled in "yes", while he intended to fill in "no"; Mr. Brown was distracted by the noise of schoolchildren playing nearby, etc. However, identifying this, in itself reasonable, conceptualization of random error with the formal term indicated by E_i is circular.

To see this, first recall that the true score cannot be conceptualized as the average score over *actual* replications. This would violate the basic assumptions of the model, especially those concerning independence and parallelism of repeated measurements. For the same reason, error cannot be conceptualized as the lump sum of all variables that cause variation in the observed scores over actual replications: The true score is defined through a thought experiment, and so is the error score. Further, we have seen that the semantics of classical test theory do not only require that Mr. Brown is brainwashed inbetween measurements, but also that Mr. Brown takes a trip in a time-machine inbetween measurements, because the true score must be conceptualized as being instantiated at a particular time point. What, then, is supposed to cause the fluctuations, that might generate the probability distribution on which the true score is defined, on this particular time point? In other words: What varies in the replications under consideration?

There are three possible answers to this question. The first is: Nothing. In this interpretation, we have a quite mysterious source of randomness, which is supposedly inherent to Mr. Brown himself. Test theorists holding this interpretation should definitely have a chat with people working in quantum mechanics, for it would follow that human beings and quarks have more in common than one might think. But certainly, the random error would not come from variations in 'irrelevant' variables, because there would not be variation at all. This interpretation does therefore not return the typical idea of random error as discussed above.

The second answer to the question is: Everything. Now we imagine Mr. Brown taking his United Nations test not only in the original testing situation, but also in the jungle, in space, under water, while playing a game of tennis, and so on. This interpretation, however, neither returns the typical idea of random error. For nothing prohibits Mr. Brown's constitution to be changed in such a way that, say, his social desirability level goes down, or, more drastically, he turns deaf, or, still more dramatically, he becomes identical to a different person (say, Kofi Annan). Therefore, this interpretation forces us to include under the header 'random error' factors that we do not usually view as such – social desirability, for instance, is the classic example of a variable that is supposed to influence test scores systematically, not randomly.

The third answer that we may give is: Some things will change, and some will not. This, however, requires that we distinguish between factors that are variable across replications, and factors that are constant. Doing this allows us to create the desired interpretation of random error, but at the price of circularity. For which things are supposed to change in order to return the desired interpretation of random error? Well, those things that are supposed to be unsystematic. Which things are that? Supposedly, Mr. Brown's headache, schoolchildren playing nearby, etc. But why these things? Because they are influential and change across replications. And why do they change? Because we have included them as varying in the thought experiment. Now we are back at square one. Thus, a platonic conception of error, as reflecting unsystematic influences on the observed testscore, involves a circularity in reasoning. It actually allows us to create any interpretation of random error we desire, by incorporating the factors we want to subsume under that header as variable in the thought experimental replications. Nothing is gained in this interpretation.

Clearly, the true score is ill-defined as an expected value, because the distribution that is supposed to generate it cannot be characterized - not even roughly. The thought experiment that should do this does not specify the conditions under which replications should occur, except for the fact that these should be statistically independent, which, ironically, is exactly the reason that such replications cannot in general be equated with actual replications. Moreover, there is a serious problem in the interpretation of the thought experimental replications. Not only does classical test theory fail to provide grounds for choosing between the above accounts of random error, but the available accounts are either mysterious, inadequate, or circular. The thought experiment does not elucidate the situation. Mr. Brown's brainwash adds little to the syntactic formula $t_i = \mathcal{E}(X_i)$, but rather obscures the fact that taking the expectation of a distribution, which is defined at a particular moment on a particular person, is a doubtful move. Thus, when Lord & Novick invite the reader to assume the existence of a true score, it is not at all clear what the reader is supposed to believe in.

The multiplication of true scores The true score is ill-defined, but this, in itself, is not sufficient reason for rejecting the realist interpretation. Many concepts lack an unambiguous definition; surely, most psychological constructs do. The inability to define a construct unambiguously does not force us to the conclusion that the phenomena denoted by that construct therefore cannot exist. In many cases, definitions are the result of doing research, not a prerequisite for it. Indeed, much scientific progress can be described in terms of a continuous redefining of scientific constructs.

However, what we may require from a realist interpretation of true scores is some kind of testability. This does not mean that theories must be falsifiable in the strict sense of Popper (1959) - in psychology, this would probably leave us with no theories at all - but there must be some kind of connection to observations that takes the form of a prediction. In theories of psychological measurement, this connection usually takes the form of discriminative hypotheses. For example, the intelligence tester may concede that he cannot give a definition of intelligence, but he can formulate the hypothesis that the number series '1 1 2 3 5 8 ..' does measure intelligence (in a population of normal adults), while the item 'I like to go to parties' does not. This is, for example, the way that constructs are related to testable predictions in latent variable modeling. In the case of true score theory, no such connection can be made. There are two reasons for this. First, according to the classical test model, a distinct true score exists for literally every distinct test. Second, the theory cannot say what it means for two distinct tests to measure the same true score, except through the awkward requirement of parallelism. Therefore, the true score hypothesis does not yield testable predictions in the discriminative sense discussed above.

Consider the first point. The definition of the true score as an expected value leaves no room for saying that some tests do measure a true score, and some do not: We may always imagine a series of thought experimental replications and define the true score as the expected value of the resulting distribution. This means that every imaginable test has an associated true score, as has been illustrated in the previous section. Admitting the true score into reality thus forces the conclusion that every person is a walking collection of infinitely many true scores - one for every imaginable testing procedure. It would seem that, in this way, reality gets rather crowded.

Second, classical test theory cannot posit the true score as a hypothesis generating entity. This could, in principle, be done if it were reasonable for, say, the intelligence tester, to say that a number series item measures the same true score as a Raven item, similar to the way different items can be related to a single latent variable in item response models. Within true score theory, the only way to say that two tests measure the same true score is by saying that the tests are parallel. However, there is absolutely no reason to suppose that two distinct items that measure the same construct should be empirically parallel. Moreover, it has been shown in section 2.2.2 that the very idea, that two items that are empirically parallel measure the same true score, is inconsistent in its own right: The only item that could be said to measure the same true score as the number series item '1 1 2 3 5 8 ..' is the number series item '1 1 2 3 5 8 ..' itself. Of course, one could reason that two items that measure the same construct should have, for example, perfectly correlated true scores. This does yield testable predictions, but these do not result from the true score hypothesis itself. Rather, they result from a hypothesis concerning relations between true scores; a hypothesis that, in turn, is based on the idea that the items measure the same construct – in fact, it is based on a latent variable hypothesis and specifies Jöreskog's (1971) congeneric model. The construct theory can specify testable discriminative hypotheses ('these items measure intelligence, but those do not'), but the hypothesis that there exists a true score for a given measurement procedure cannot.

Thus, upon a realist interpretation, the true score is a metaphysical entity of the worst kind: Posing its existence does not lead to a single testable hypothesis. This does not mean that true scores, or classical test theory, are useless; obviously, the true score may figure in a set of hypotheses based on substantive theory, as it does in the congeneric model. It means that the true score hypothesis in itself is not capable of generating testable predictions.

Operationalism and true score theory

Two conclusions must be drawn. First, it is unclear what a true score is, because the probability distribution that is supposed to generate it lacks sufficient specification. Second, the true score hypothesis, in itself, does not lead to predictions. Note that these conclusions are not problematic for the true score concept, or for classical test theory in general. They are only problematic for a full-blown realist interpretation of classical test theory. It seems that such an interpretation is untenable.

However, just like the lack of correspondence between construct scores and true scores should not, in itself, bother the classical test theorist, the fact that a realist conception of true scores is problematic does not pose a problem for classical test theory either. It does suggest that we consider different ways of conceptualizing the true score's ontological status. Since there are enough alternatives to realism, the question becomes within which of these the true score could find a home. An adequate account of the theoretical status of true scores also illuminates what kind of philosophical outlook would be consistent with an identification of true scores with construct scores. The observation, that such an identification is completely unreasonable, suggests that the philosophical viewpoint that is consistent with it will also be completely unreasonable. In fact, the philosophical viewpoint that is consistent with classical test theory (as well as with the identification of constructs with true scores) is the most unreasonable of all, namely operationalism.

Operationalism (Bridgman, 1927) holds that the meaning of a theoretical term is synonymous with the operations by which it is measured. Interestingly, we have seen that the true score is defined without reference to anything but a measurement process. The true score is thus completely defined in terms of a series of operations: It is the proportion of times Mr. Brown would be in favour of the United Nations if he were tested infinitely many times. That the operations in question are hypothetical, and cannot be carried out, is a peculiar feature of the true score, but it does not preclude the conclusion that the true score is defined in terms of these operations, which is consistent with operationalism.

The true score also has some typical problematic aspects that are essentially identical to those faced by the operationalist philosophy of measurement. It has been argued against that view, for example, that it leads to a multiplication of theoretical terms (Suppe, 1977). For example, suppose that the meaning of the theoretical term 'intelligence' is equated with the set of operations that lead to an IQ-score on the Stanford-Binet. It immediately follows that the WAIS, the Raven, or any other intelligence test cannot also measure intelligence, because each test specifies a distinct set of operations. So, each measurement procedure generates a distinct theoretical concept. It is therefore conceptually difficult, if not impossible, for an operationalist to say what it means for two tests to measure the same construct.

In classical test theory, we face essentially the same problem. The true score is defined in terms of the expected value on a particular test, and since each test generates a distinct expected value, it generates a distinct true score. Moreover, when the classical test theorist tries to express the idea that two tests measure the same true score, he runs into troubles that are comparable to those facing the operationalist. The only option, that comes close to saying that two tests measure the same true score, is to invoke the idea of distinct yet parallel tests. This is not only a highly contrived and unnecessarily strict requirement, that in no way matches the intended meaning of the proposition that the Stanford-Binet and the WAIS measure the same construct; it is essentially a concealed way of saying that the only test that measures the same construct as test x is test x itself. The same conclusion would be reached by an operationalist.

The operationalist view also resolves some of the problems surrounding the true score, as exposed in previous sections. For instance, the operationalist does not refer to a construct or attribute score as something that exists in objective reality, and certainly not as something that exists independent of the measurement process. Thus, we cannot speak of length without mentioning a centimeter; we cannot speak of weight without referring to a balance scale; and we cannot consider intelligence apart from the IQ-test. To do this is, in the eyes of the operationalist, to indulge in intolerable metaphysical speculation. A difficulty for operationalism may be created by countering that objects would still have a definite length if there were nobody to measure it, and that people already possessed intelligence before the advent of IQ-tests. This argument, which of course invites realism about such attributes, can be deflected by invoking scores on attributes as dispositional properties. Rather than accepting my height, which is about 185 cm, as a property which I have independent of any measurement apparatus used, the proposed solution is that height is a dispositional property with respect to a measurement procedure: *if* I were measured with a centimeter, *then* the indicated height would be about 185 cm.

It is interesting to inquire whether the true score could also be interpreted as a disposition, and, if so, what kind of disposition it is. In this context, Van Heerden & Smolenaars (1989) propose a taxonomy of dispositions by cross-classifying them with respect to the question whether they concern heterogeneous or homogeneous behavior, and whether this behavior is recurrent or unique. An example of a disposition that refers to the unique, non-recurrent, presentation of a specific piece of homogeneous behavior is 'being mortal'. Saying that John is mortal expresses the fact that, upon the appropriate operations (e.g., poisoning him), John will decease. Dispositions may also refer to the recurrent presentation of homogeneous behavior. For example, the proposition 'Professor G. likes to go out with his Ph. D. students at conferences' refers to professor G.'s tendency to end up in a bar, in the company of his Ph. D. students, at conferences – usually at closing time. A proposition that refers to the recurrent presentation of a heterogeneous set of behaviors is 'Professor J. is forgetful'. This proposition refers to professor J.'s tendency to display a wide variety of behaviors recurrently; for example, professor J. regularly finds himself in the secretary's office without remembering what he went there for, he forgets to remove the keys from his bicycle, etc. Finally, a disposition can refer to the unique presentation of a heterogeneous set behaviors. Van Heerden & Smolenaars (1989) give the example of 'being intelligent', which can be viewed as a disposition that may manifest itself once by completing the different (heterogeneous) tasks of an intelligence test.

The latter example is interesting, because Rozeboom (1966-a) treats test scores in a similar manner. He observes that, in order to conceptualize the idea that 'every person has an IQ-score', we must not interpret this sentence realistically (in which case it is obviously false), but in terms of the dispositional 'for every person it is true that, if that person were tested, he or she would obtain an IQ-score'. In a similar vain, we can interpret the sentence 'every person has a true IQ-score' as 'for every person it is true that, if he or she were tested infinitely many times (with intermediate brainwashing and time-travel), the resulting observed score distribution would have an expected value'. In the terminology of Van Heerden & Smolenaars (1989), the behavior under consideration could be heterogeneous (if it refers to total test scores) or homogeneous (if it refers to item scores), but it would certainly be recurrent, although with respect to thought experimental replications[5]. The true score, as classical test theory defines it, may thus be considered to be a disposition, which specifies how a person will behave in a counterfactual state of affairs.

The overburdening of reality, which follows from the realist interpretation of true scores, dissolves upon this view. I am not a walking collection of true scores on all imaginable tests; but it is the case that, *if* I were repeatedly administered an arbitrary test (with the appropriate brainwashing and timetravel), *then* my true score on that test would take on a definite value. Such dispositions are thus not located in reality, but rather specify characteristics of subjunctive, or in our case counterfactual, conditionals (Rozeboom, 1966-a). It is certainly the case that the true score, with its thought experimental character, must be considered an oddball among recognized dispositional properties (e.g., solubility, fragility), because these usually specify responses, of the object to which the dispositional property is ascribed, in situations that could actually occur. The concept of fragility is considered to be dispositional, because it is characterized by conditionals such as 'this vase is fragile, for if it were dropped, it would break'; and we may check this by actually dropping the vase. Similarly, all of the examples of Van Heerden & Smolenaars (1989) refer to actually realizable behaviors in realistic situations. This is the reason that Ryle (1949) characterizes dispositions as inference tickets. Rozeboom (1973) argues that such dispositional properties involve a realist commitment to underlying characteristics that generate the dispositions, and Van Heerden & Smolenaars

[5] It is interesting to observe that, upon a dispositional interpretation of true scores, the stochastic aspect of classical test theory may receive a new interpretation. For instance, the value of the observed score variance for a given subject (normally interpreted as an index of measurement precision) would in this case reflect the strength of the disposition.

(1989) interpret dispositions as promissory notes (i.e., they involve a promise that, if one looks into dispositional properties, one could discover more fundamental laws that govern dispositional behavior).

The true score is not open to such an interpretation as long as the thought experiment is retained. For it refers to what would happen in an impossible state of affairs, namely to the observed score distribution that Mr. Brown would generate in the brainwashing thought experiment. So, the thought experimental character of the true score makes it quite useless as an inference ticket in the common interpretation of that term. A possible response to this problem, is to dispose of the thought experiment and to interpret true scores as propensities, without reference to limiting frequencies. I have already discussed this possibility in 2.2.1. Doing this, however, would preclude an interpretation of true score theory as a theory of measurement, for the only remaining connection to the theory of errors would disappear. In fact, classical test theory would become a statistical theory of dispositions; interpretations of observed score variance as 'random error', for example, would be out of the question, and the interpretation of the squared correlation between observed and true scores as 'reliability' would hardly make sense. Certainly, the founders of the theory did not intend classical test theory in this way, and its users do not interpret it so; in fact, it is likely that, interpreted as a statistical theory of dispositions for unique events, classical test theory would not appeal to those involved in psychological testing at all.

2.3 Discussion

Classical test theory was either one of the best ideas in 20th century psychology, or one of the worst mistakes. The theory is mathematically elegant and conceptually simple, and in terms of its acceptance by psychologists, it is a psychometric success story. However, as is typical of popular statistical procedures, classical test theory is prone to misinterpretation. One reason for this is the terminology used: If a competition for the misnomer of the century existed, the term 'true score' would be a serious contestant. The infelicitous use of the adjective 'true' invites the mistaken idea that the true score on a test must somehow be identical to the 'real', 'valid', or 'construct' score. This chapter has hopefully proved the inadequacy of this view beyond reasonable doubt.

The problems with the platonic true score interpretation were, however, seen to run deeper than a confound of validity and unreliability. It seems that the entire idea, that true scores are real entities, leads to a metaphysical explosion of reality. It was therefore argued that true scores are not to be granted a place in reality, but rather should be seen as a particular kind of score. And just as we do not say that every person has an IQ-score, but rather that every person would receive an IQ-score if tested, we have to refrain from saying that every person has a true score. What we can say is that every person would have an expected value, if he or she were repeatedly tested in a long run of testing occasions with intermediate brainwashing and time travel. The true score must thus be considered to be a dispositional concept, in particular as a disposition to generate specific long run frequencies.

The fact that not even such frequencies can be granted a place in reality, but must be considered in terms of a thought experiment, further degenerates the connection that the true score may bear to the real world.

A philosophy of measurement that could accommodate for these problems is operationalism. The true score has some problems that are similar to those of operationalism, and, at the same time, other problems can be resolved by introducing arguments that are similar to the operationalist defense. An operationalist interpretation could also restore the identity of constructs and true scores; this time, however, not by upgrading true scores, but by degrading constructs. Such an interpretation should therefore not be considered a platonic true score interpretation, but rather a nonplatonic construct interpretation. That is, if one is willing to give up the realist semantics of psychological constructs such as intelligence, extraversion, or attitudes, and to conceive of them in an operationalist fashion (with a touch of fictionalism to accommodate for the thought experimental character of the true score), then the true score could be a candidate for representing these constructs in a measurement model. The fictionalist element, which must be introduced because the true score, as a disposition, generalizes over a domain of impossible replications, precludes the interpretation of the true score as an inference ticket (Ryle 1949), or a promissory note (Van Heerden & Smolenaars, 1989). It also deprives the concept of the possible realist semantics that may be introduced for dispositions (Rozeboom, 1973), unless the entire idea that we are dealing with a theory of measurement is given up, by dismissing the thought experiment and disconnecting the theory from the theory of errors. I suspect that no classical test theorist will be willing to do this; classical test theory is intended, formulated, and used as a theory of measurement, and I do not expect classical test theorists to revert their self-image from 'expert in measurement theory' to 'expert in dispositional psychology'. However, retaining the idea that we are dealing with a theory of measurement requires abandoning a realist interpretation of the true score, and taking an operationalist perspective.

Of course, since the true score is appropriately characterized as a product of the test theorists imagination, and therefore does not obtain a realist ontology, this is not a particularly pressing philosophical problem. At least, it is better than the kind of realism needed to localize the true score in the world. It is a pressing theoretical problem for psychology, however, because I do not think that many researchers in psychology are particularly interested in a true score, which specifies a disposition with respect to a set of impossible situations. So, once again we see the fundamental tension that Lord & Novick have introduced through their axiomatic treatment of test theory: The theory is constructed in such a way that it always works, but at the price of losing the natural interpretation of its central concepts. A psychologically meaningful interpretation of true scores and random error, as reflecting a stable characteristic and unsystematic variation respectively, is philosophically untenable. A philosophically acceptable interpretation of these concepts, as products of the imagination which refer to recurrent dispositions in a counterfactual state of affairs, is psychologically unattractive. Classical test theory systematically falls between these two stools.

It is my understanding that few, if any, researchers in psychology conceive of psychological constructs in a way that would justify the use of classical test theory as an appropriate measurement model. Why, then, is the classical test theory model so immensely successful? Why is it that virtually every empirical study in psychology reports values of Cronbach's α as the main justification for test use? I am afraid that the reason for this is entirely pragmatic, and has been given in section 2.2.2: The common use of classical test theory does not involve testing the model assumptions. The lower bound strategy always returns a value for the internal consistency coefficient. In fact, this value can be obtained through a mindless mouse-click. Inserting the lower bound into formulae for disattenuating correlations between test scores, as advocated by Schmidt & Hunter (1999), will further allow one to boost validity coefficients to whatever level is desired. All this will come at no additional costs, for it does not require any of the tedious work involved in latent variable models, which moreover have a tendency to prove many of the commonly held interpretations of test scores illusory. Applying classical test theory is easy, and a commonly accepted escape route to avoid notorious problems in psychological testing, such as constructing unidimensional tests. The model is, however, so enormously detached from common interpretations of psychological constructs, that the statistics based on it appear to have very little relevance for psychological measurement. Coupled with the unfortunate misinterpretations of the true score as the construct score, of random error as irrelevant variation, and of reliability as some kind of fixed characteristic of tests, instead of as a population dependent property of scores, it would seem that large parts of the psychological community are involved in self-deception. Wishful thinking, however, is not a particularly constructive scientific procedure, and mystifying test theoretical concepts is certain to obstruct, rather than stimulate, progress in psychology. I therefore hope that the analysis reported here has added to the understanding and demystification of classical test theory concepts, and has made clear that much more is needed for an adequate treatment of psychological test scores.

3. LATENT VARIABLES

Once you have formed the noun 'ability' from the adjective 'able', you are in trouble.
– B.F. Skinner, 1987

3.1 Introduction

In the previous chapter, I have argued that the classical test theory model is unsatisfying for a number of reasons. Most important is the fact that the attribute to be measured is not adequately represented in the model. The reason for this is that the true score is an operationalist concept, and can only represent a psychological attribute if this attribute is similarly defined in an operationalist fashion. In fact, unless one holds a strongly operationalist view of the measurement process, it is difficult to maintain even that classical test theory is a theory of measurement in the first place.

A view of measurement that does represent the attribute explicitly in the model formulation can be based on latent variable theory. In latent variable models, one sets up a formal structure that relates test scores to the hypothesized attribute, deduces empirical implications of the model, and evaluates the adequacy of the model by examining the goodness of fit with respect to empirical data. Because the latent variable model has to be restricted to make empirical tests possible, a theoretical justification of the model structure is, in general, required. Latent variable theory thus goes beyond classical test theory in that it attempts to construct a hypothesis about the data generating mechanism, in which the attribute is explicitly represented as a latent variable.

Historically, the conceptual framework originates with the work of Spearman (1904), who developed factor analytic models for continuous variables in the context of intelligence testing. In the twentieth century, the development of the latent variable paradigm has been spectacular. The factor analytic tradition continued with the work of Lawley (1943), Thurstone (1947) and Lawley & Maxwell (1963), and entered into the conceptual framework of confirmatory factor analysis (CFA) with Jöreskog (1971), Wiley (1973), and Sörbom (1974). In subsequent years, CFA became a very popular technique, largely because of the LISREL program by Jöreskog & Sörbom (1993). In a research program that developed mostly parallel to the factor analytic tradition, the idea of latent variables analysis with continuous latent variables was applied to dichotomous observed variables by Guttman (1950), Lord (1952; 1980), Rasch (1960), Birnbaum (1968) and Mokken (1970). These measurement models, primarily used in educational testing, came to be known as Item Response Theory (IRT) models. The IRT framework was extended to deal with polytomous observed variables by Samejima (1969), Bock (1972), and Thissen & Steinberg (1984). Meanwhile, in yet another parallel research program, methods were developed to deal with categorical latent variables. In this context, Lazarsfeld (1950), Lazarsfeld & Henry (1968), and Goodman (1974) developed latent structure analysis. Latent structure models may involve categorical observed variables, in which case we speak of latent class analysis, or metrical observed variables, giving rise to latent profile analysis (Bartholomew, 1987). After boundary-crossing investigations by McDonald (1982), Thissen & Steinberg (1986), Takane & De Leeuw (1987), and Goldstein & Wood (1989), Mellenbergh (1994) connected some of the parallel research programs by showing that most of the parametric measurement models could be formulated in a common framework.

At present, there are various developments that emphasize this common framework for latent variables analysis, cases in point being the work of Muthén & Muthén (1998), McDonald (1999), and Moustaki & Knott (2000). Different terms are used to indicate the general latent variable model. For example, Goldstein & Wood (1989) use the term Generalized Linear Item Response Model (GLIRM), while Mellenbergh (1994) speaks of Generalized Linear Item Response Theory (GLIRT), and Moustaki & Knott (2000) follow McCullagh & Nelder (1989) in using the term Generalized Linear Model (GLIM). I will adopt Mellenbergh's terminology and use the term GLIRT, because it emphasizes the connection with IRT, and, in doing so, the fact that the model contains at least one latent variable. Now, at the beginning of the twenty-first century, it would hardly be an overstatement to say that the GLIRT model, at least among psychometricians and methodologists, has come to be the received view in the theory of psychological measurement – notwithstanding the fact that classical test theory is still the most commonly used theory in test analysis.

The growing use of latent variables analysis in psychological research is interesting from a philosophical point of view, exactly because latent variable theory, in contrast to classical test theory, is typically aimed at constructing an explanatory model to account for relations in the data. This means that explanations that make use of unobservable theoretical entities are increasingly entertained in psychology. As a consequence, the latent variable has come to play a substantial role in the explanatory structure of psychological theories. Now, concepts closely related to the latent variable have been discussed extensively. These concepts include the meaning of the arrows in diagrams of structural equation modeling (see, for example, Sobel, 1994; Pearl, 1999; Edwards & Bagozzi, 2000), the status of true scores (Klein & Cleary, 1967; Lord & Novick, 1968; Lumsden, 1976), definitions of latent variables (Bentler, 1982; Bollen, 2002), specific instances of latent variables such as the Big Five Factors in personality research (Lamiell, 1987; Pervin, 1994), and the trait approach in general (Mischel, 1968; 1973). Also, the status of unobservable entities is one of the major recurrent themes in the philosophy of science of the past century, where battles were fought over the conceptual status of unobservable entities such as electrons (see Cartwright, 1983, Hacking, 1983, Van Fraassen, 1980, and Devitt, 1991, for some contrasting views). However, the theoretical status of the latent variable as it appears in models for psychological measurement has not received a thorough and general analysis as yet.

Questions that are relevant, but seldomly addressed in detail, are similar to the questions addressed in the previous chapter. For instance, should we assume that the latent variable signifies a real entity, or conceive of it as a useful fiction, constructed by the human mind? Should we say that we measure a latent variable in the sense that it underlies and determines our observations, or is it more appropriately considered to be constructed out of the observed scores? What exactly constitutes the relation between latent variables and observed scores? Is this relation of a causal nature? If so, in what sense? And, most importantly, is latent variable theory neutral with respect to these issues? In the course of discussing these questions, we will see that latent variable theory is not philosophically neutral; specifically, it will be argued that, without a realist interpretation of latent variables, the use of latent variables analysis is hard to justify. At the same time, however, the relation between latent variables and individual processes proves to be too weak to defend causal interpretations of latent variables at the level of the individual. This observation leads to a distinction between several kinds of latent variables, based on their relations with individual processes.

3.2 Three perspectives on latent variables

The syntax, semantics, and ontology of latent variable models are substantially different from those used in classical test theory. Syntactically, the model relates expected item responses to a latent variable by specifying an appropriate item response function. This function formulates a regression of the item score on a latent variable. Semantically, the expected item response may be interpreted in two ways: As a true score, in which case we follow a stochastic subject interpretation, or as a subpopulation mean, in which case we follow a repeated sampling interpretation. From an ontological viewpoint, the model is most naturally interpreted in a realist fashion. This probes the question what constitutes the nature of the relation between latent variables and observed scores. It is argued that this relation can be constructed as a causal one, but only when the latent variable is interpreted as the cause of differences between subpopulations.

3.2.1 The formal stance

Syntax In modern test theory models, such as the various IRT-models or confirmatory factor models, the relation between the latent variable and the observed scores is mathematically explicit. In GLIRT, the form for this relation is a generalized regression function of the observed scores on the latent variable, although this regression may differ in form. The model relates an observed item response variable U to a latent variable θ via a function of the form

$$g\left[\mathcal{E}(U_{ij})\right] = \beta_j + \alpha_j \theta_i, \tag{3.1}$$

where g is a link function, $\mathcal{E}(U_{ij})$ is interpreted either as the expected item response of subject i on item j, or as the expectation of the item response in a population of subjects with position θ_i on the latent variable, and α_j and β_j are an item-specific regression weight and intercept term, respectively.

Some specific forms of the model will be relevant in the following chapters. First, in item response theory for dichotomous items and continuous latent variables, the link function is often taken to be the logit transformation (the natural logarithm of the odds ratio). In this case we have a model of the form

$$\ln\left[\frac{\mathcal{E}(U_{ij})}{1-\mathcal{E}(U_{ij})}\right] = \beta_j + \alpha_j\theta_i. \tag{3.2}$$

The intercept term β is then usually interpreted as item difficulty, because it refers to the location of the item response function on the θ-scale, and α is interpreted as item discrimination, because it refers to the slope of the item response function. If all item discrimination parameters are assumed equal, then we have an additive model, because item and subject effects are independent (i.e., they do not interact, where the interpretation of 'interact' is the same as in analysis of variance). This form of the model is known as the Rasch model (Rasch, 1960). Allowing the discrimination parameters to vary gives the less restrictive two-parameter logistic model introduced by Birnbaum (1968). This model can be viewed as incorporating a person × item interaction term.

If item responses are continuous, and the function g is taken to be the identity link, we arrive at Jöreskog's (1971) congeneric model, better known as the common factor model:

$$\mathcal{E}(U_{ij}) = \beta_j + \alpha_j\theta_i. \tag{3.3}$$

Finally, if the latent variable is categorical, we can formulate the latent class model (if item responses are dichotomous) or the latent profile model (if item responses are continuous) by dummy coding the latent variable. Various other models can be arrived at by introducing appropriate restrictions and transformations (Mellenbergh, 1994), but the models discussed above are the most important ones for the present discussion.

It is important to realize that, despite the intricate mathematics that sometimes accompanies the literature on latent variable theory, the basic form of the model is very simple. For instance, in a factor model for general intelligence, the model says that an increase of n units in the latent variable leads to an increase of n times the factor loading in the expected value of a given item. So, formally, the model is just a regression model, but the independent variable is latent rather than manifest. The ingenious idea in latent variable modeling is that, while the model cannot be tested directly for any given item because the independent variable is latent, it can be tested indirectly through its implications for the joint probability distribution of the item responses for a number of items. Specifically, in the standard latent variable model the item responses will be independent, conditional on the latent variable, which means that the items satisfy local independence.

Now there are two things we can do on the basis of our set of assumptions. First, we can determine how observed scores would behave if they were generated under our model (this applies not only to mathematical derivations but also to simulation studies). Second, we can develop plausible procedures to estimate parameters in the model on the basis of manifest scores, given the assumption that these scores were generated by our model. It is sometimes implicitly suggested that the formal derivations tell us something about reality, but this is not the case. Each supposition 'inside' the formal system is a tautology, and tautologies in themselves cannot tell us anything about the world. So this is all in the syntactic domain, that is, it has no meaning outside the formal theory. Let us denote the latent variable as it appears in this formal stance (that is, the concept indicated by θ, in the IRT literature, or by ξ, in the SEM literature) as the formal latent variable.

Semantics The syntax of latent variable theory specifies a regression of the observed scores on the latent variable. What are the semantics associated with this relation? In other words: how do we interpret this regression?

Of course, as is the case for classical test theory, the syntax of latent variables analysis is taken from statistics, and so are its semantics. And, like classical test theory, latent variable theory needs an interpretation for the use of the expectation operator in the model formulation. Because it is not at all clear why a response to an item, say, the item '2 + 2 = ..', should be considered a random variable, it is important to interpret the item response in such a way as to justify this approach. The problem faced here is similar to that faced by the classical test theorist in the definition of the true score, but the latent variable theorist has a considerably greater freedom of interpretation.

The first interpretation, known as the stochastic subject interpretation, uses the same line of reasoning as classical test theory, and views the expectation as applying to the individual subject. This implies a series of hypotheticals of the form 'given that subject i has value θ_i on the latent variable, i's expected item response equals $\mathcal{E}(U_{ij}|\theta_i)$', where $\mathcal{E}(U_{ij}|\theta_i)$ is the expectation of the item response as given by the item response function. Supposing that the imaginary subject John takes an intelligence test item, this would become something like 'given that John's level of intelligence is two standard deviations below the population mean, he has a probability of .70 to answer the item '2 + 2 = ..' correctly'. For subjects with different positions on the latent variable, different parameters for the probability distribution in question are specified. So, for John's brighter sister Jane we could get 'Given that Jane's level of intelligence is one standard deviation above the population mean, Jane has a probability of .99 to answer the item correctly'. The item response function (i.e., the regression of the item response on the latent variable) then specifies how the probability of a correct answer changes with the position on the latent variable. The stochastic subject interpretation requires a thought experiment similar to that used in classical test theory, and in this interpretation the expected value of subject i on item j, $\mathcal{E}(U_{ij})$, can be considered to be identical to subject i's true score on item j if the latent variable model is true.

In contrast to classical test theory, however, the model can also be formulated without the brainwashing thought experiment. This requires conceptualizing the model in terms of a repeated sampling interpretation, which is more common in the literature on factor analysis (see, for example, Meredith, 1993) than in the literature on IRT. This is a between-subjects formulation of latent variables analysis. It focuses on characteristics of populations, instead of on characteristics of individual subjects. The probability distribution of the item responses, conditional on the latent variable, is conceived of as a probability distribution that arises from repeated sampling from a population of subjects with the same position on the latent variable. In particular, parameters of these population distributions are related to the latent variable in question.

Thus, the repeated sampling interpretation is in terms of a series of sentences of the form 'the population of subjects with position θ_i on the latent variable follows distribution f over the possible item responses u_{ij}; the expected item response $\mathcal{E}(U_{ij}|\theta_i)$ is the expectation of the item responses in the subpopulation of subjects with position θ_i on the latent variable'. Now, the probability distribution over the item responses, that pertains to a specific position θ_i on the latent variable, arises from repeated sampling from the population of subjects taking this position; the expectation may then be interpreted as a subpopulation mean. In this interpretation, the probability that John answers the item correctly does not play a role. Rather, the focus is on the probability of drawing a person that answers the item correctly from a population of people with John's level of intelligence, and this probability is .70. In other words, 70% of the population of people with John's level of intelligence (i.e., a level of intelligence that is two standard deviations below the population mean) will answer the item correctly; and 30% of those people will answer the item incorrectly. There is no random variation located within the person.

The difference between the stochastic subject and repeated sampling interpretations is substantial, for it concerns the very subject of the theory. The two interpretations entertain different conceptions of what it is we are modeling: in the stochastic subject formulation, we are modeling characteristics of individuals, while in the repeated sampling interpretation, we are modeling subpopulation means. However, if we follow the stochastic subject interpretation and assume that everybody with John's level of intelligence has probability .70 of answering the item correctly, then the expected proportion of subjects with this level of intelligence that will answer the item correctly (repeated sampling interpretation) is also .70. The assumption that the measurement model has the same form within and between subjects has been identified as the local homogeneity assumption (Ellis & Van den Wollenberg, 1993). Via this assumption, the stochastic subject formulation suggests a link between characteristics of the individual and between-subjects variables. Ellis & Van den Wollenberg (1993) have shown, however, that the local homogeneity assumption is an independent assumption that follows in no way from the other assumptions of the latent variable model. Also, the assumption is not testable, because it specifies what the probability of an item response would be in a series of independent replications with intermediate brainwashing in the Lord & Novick (1968; p. 29) sense. Basically, this renders the connection between within-subject processes and between subjects variables speculative (in the best case). In fact, it will be argued later on that the connection is little more than an article of faith: the standard measurement model has virtually nothing to say about characteristics of individuals, and even less about item response processes. This will prove crucially important for the ontology of latent variables, to be discussed later in this chapter.

3.2.2 The empirical stance

Because a latent variable model has testable consequences at the level of the joint distribution of the item responses, it is possible to test the adequacy of the model against the data. In contrast to classical test theory applications, such model tests are commonly carried out in latent variables analysis. Like many testing procedures throughout science, however, such model fit tests suffer from the problem of underdetermination of theory by data. This means that many data generating mechanisms can produce the same structure in the data as the hypothesized model. So, if observed variables behave in the right way, a latent variable model will fit, but this does not imply that the model is correct.

The issue that is called underdetermination in the philosophy of science is called statistical equivalence in the modeling literature (see, for example, Hershberger, 1994). In this context it has, for instance, been shown by Bartholomew (1987; see also Molenaar & Von Eye, 1994) that a latent profile model with p latent profiles generates the same first and second order moments (means, variances, and covariances) for the observed data as a factor model with $p-1$ continuous latent variables. These models are conceptually different: the factor model posits continuous latent variables (i.e., it specifies that subjects vary in degree, but not in kind), while the latent profile model posits categorical latent variables at the nominal level (i.e., it specifies that subjects vary in kind, but not in degree). This suggests, for example, that the five factor model in the personality literature corresponds to a typology with six types. Moreover, on the basis of the covariances used in factor analysis, the Big Five Factors would be indistinguishable from the Big Six Types. The fact that theoretically distinct models are practically equivalent in an empirical sense urges a strong distinction between the formal and empirical structure of latent variables analysis.

This point is important because it emphasizes that the attachment of theoretical content to a latent variable requires an inferential step, and is not in any way 'given' in empirical data, just as it is not 'given' in the mathematical formulation of a model. The latent variable as it is viewed from the empirical stance, i.e., the empirical entity that is generally presented as an estimate of the latent variable, will be denoted here as the operational latent variable. Note that there is nothing latent about the operational latent variable. It is simply a function of the observed variables, usually a weighted sumscore (that the weights are determined via the theory of the formal latent variable does not make a difference in this respect). Note also that such a weighted sumscore will always be obtained, and will in general be judged interpretable if the corresponding model fits the data adequately. The foregoing discussion shows, however, that the fit of a model does not entail the existence of a latent variable. A nice example in this context is given by Wood (1978), who showed that letting people toss a number of coins (interpreting the outcome of the tosses as item responses) yields an item response pattern that is in perfect agreement with the Rasch model. A more general treatment is given in Suppes and Zanotti (1981) who show that, for three dichotomous observed variables, a latent variable can be found if and only if the observed scores have a joint distribution. The developments in Bartholomew (1987) and Molenaar & Von Eye (1994) further show that model fit does not entail the form (e.g., categorical or continuous) of the latent variable, even if its existence is assumed a priori.

The above discussion shows that the connection between the formal and operational latent variable is not self-evident. In order to make that connection, we need an interpretation of the use of formal theory in empirical applications. This, in turn, requires an ontology for the latent variable.

3.2.3 The ontological stance

The formal latent variable is a mathematical entity. It figures in mathematical formulae and statistical theories. Latent variable theory tells us how parameters that relate the latent variable to the data could be estimated, if the data were generated under the model in question. The 'if' in the preceding sentence is very important. It points the way to the kind of ontology we have to invoke. The assumption, that it was this particular model that generated the data, must precede the estimation process. In other words, if we consider the weighted sumscore as an estimate of the position of a given subject on a latent variable, we do so under the model specified. Now this weighted sumscore is not an estimate of the formal latent variable: we do not use an IQ-score to estimate the general concept usually indicated by the Greek letter θ, but to estimate intelligence. Thus, we use the formal side of the model to acquire knowledge about some part of the world; then it follows that we estimate something which is also in that part of the world. What is that something?

It will be clear that the answer to this question must consider the ontology of the latent variable, which is, in quite a crucial way, connected to its theoretical status. An ontological view is needed to connect the operational latent variable to its formal counterpart, but at first sight there seems to be a considerable freedom of choice regarding this ontology. I will argue that this is not the case.

There are basically three positions one can take with respect to this issue. The first position adheres to a form of entity realism, in that it ascribes an ontological status to the latent variable in the sense that it is assumed to exist independent of measurement. The second position could be coined 'constructivist' in that it regards the latent variable as a construction of the human mind, which need not be ascribed existence independent of measurement. The third position maintains that the latent variable is nothing more than the empirical content it carries – a 'numerical trick' used to simplify our observations: This position holds that there is nothing beyond the operational latent variable and could be called operationalist. Strictly taken, operationalism is a kind of constructivism, but the latter term is intended to cover a broader class of views (for example, the more sophisticated empiricist view of Van Fraassen, 1980). In fact, only the first of these views can be consistently attached to the formal content of latent variable theory.

Operationalism and the numerical trick

It is sometimes heard that the latent variable is nothing but the result of a numerical trick to simplify our observations. In this view, the latent variable is a (possibly weighted) sumscore and nothing more. There are several objections that can be raised against this view. A simple way to see that it is deficient is to take any standard textbook on latent variable theory and to replace the term 'latent variable' by 'weighted sumscore'. This will immediately render the text incomprehensible. It is, for example, absurd to assert that there is a sumscore underlying the item responses. The obvious response to this argument is that we should not take such texts literally; or, worse, that we should maintain an operationalist point of view. Such a move, however, raises more serious objections.

If the latent variable is to be conceived of in an operationalist sense, then it follows that there is a distinct latent variable for every single test we construct. This is consistent with the operationalist view of measurement (Bridgman, 1927) but not with latent variable theory. To see this, consider a simple test consisting of three items j, k, and l. Upon the operationalist view, the latent variable that accounts for the item responses on the subtest consisting of items j and k is different from the latent variable that accounts for the item response pattern on the subtest consisting of items k and l. So, the test consisting of items j, k, and l does not measure the same latent variable and therefore cannot be unidimensional. In fact, upon the operationalist view, it is impossible even to formulate the requirement of unidimensionality; consequently, an operationalist would have a very hard time making sense of procedures commonly used in latent variable theory, such as adaptive testing, where different tests are administered to different subjects with the objective to measure a single latent variable. Note the striking difference with classical test theory, which suffers from exactly the opposite problem, because it cannot say what it means for two tests to measure the same attribute. Where classical test theory and operationalism go hand in hand, operationalism and latent variable theory are fundamentally incompatible.

In a line of reasoning that is closely related to operationalism, it can be argued that the use of latent variable theory is merely instrumental, a means to an end. This would yield an instrumentalist point of view (Toulmin, 1953) which is akin to operationalism. In this view, the latent variable is a pragmatic concept, a 'tool', that is merely useful for its purpose (the purpose being prediction or data reduction, for example). No doubt, methods such as exploratory factor analysis may be used as data reduction techniques and, although principal components analysis seems more suited as a descriptive technique, are often used in this spirit. Also, such models can be used for prediction, although it has been forcefully argued by several authors (e.g., Maxwell, 1962) that the instrumentalist view leaves us entirely in the dark when confronted with the question why our predictive machinery (i.e., the model) works. We do not have to address such issues in detail, however, because the instrumentalist view simply fails to provide us with a structural connection between the formal and operational latent variable. In fact, the instrumental interpretation begs the question. For suppose that we interpret latent variable models as data reduction devices. Why, then, are the factor loadings determined via formal latent variable theory in the first place? Obviously, upon this view, no weighting of the sumscore can be structurally defended over any other. Any defense of this position must therefore be as ad hoc as the use of latent variables analysis for data reduction itself[1].

Realism and constructivism

So, if there is more to the latent variable than just a calculation, used to simplify our observations, what is it? We are left with a choice between realism, maintaining that latent variable theory should be taken literally - the latent variable signifying a real entity - and constructivism, stating that it is a fiction, constructed by the human mind.

The difference between realism and constructivism resides mainly in the constructivist's denial of one or more of the realist claims. Realism exists in a number of forms, but a realist will in general maintain one or several of the following theses (Hacking, 1983; Devitt, 1991). First, there is realism about theories: the core thesis of this view is that theories are either true or false. Second, one can be a realist about the entities that figure in scientific theories: the core thesis of this view is that at least some theoretical entities exist. Third, realism is typically associated with causality: theoretical entities are causally responsible for observed phenomena. These three ingredients of realism offer a simple explanation for the success of science: we learn about entities in the world through a causal interaction with them, the effect of this being that our theories get closer to the truth. The constructivist, however, typically denies both realism about theories and about entities. The question is whether a realist commitment is implied in latent variables analysis. It will be argued that this is the case: latent variable theory maintains both theses in the set of assumptions underlying the theory.

Entity realism is weaker than theory realism. For example, one may be a realist about electrons, in which case one would maintain that the theoretical entities we call 'electrons' correspond to particles in reality. This does not imply a full-blown realism about theories: for example, one may view theories about electrons as abstractions, describing the behavior of such particles in idealized terms (so that these theories are, literally taken, false). Cartwright (1983) takes such a position. Theory realism without entity realism is much harder to defend, for a true theory that refers to non-existent entities is difficult to conceive of. I will first discuss entity realism, before turning to the subject of theory realism.

Entity realism

Latent variable theory adheres to entity realism, because this form of realism is needed to motivate the choice of model in psychological measurement. The model that is customary in psychological measurement is the model in the left panel of

[1] This should not be read as a value judgement. Data reduction techniques are very important, especially in the exploratory phases of research. The fact that these techniques are important, however, does not entail that they are not ad hoc.

Figure 3.1. (The symbolic language is borrowed from the structural equation modeling literature, but the structure of the model generalizes to IRT and other latent variable models.) The model specifies that the pattern of covariation between the indicators can be fully explained by a regression of the indicators on the latent variable, which implies that the indicators are independent after conditioning on the latent variable (this is the assumption of local independence). An example of the model in the left panel of the figure would be a measurement model for, say, dominance, where the indicators are item responses on items like 'I would like a job where I have power over others', 'I would make a good military leader', and 'I try to control others'. Such a model is called a reflective model (Edwards & Bagozzi, 2000), and it is the standard latent variable model in psychology - employed in prominent models such as the general intelligence and Big Five models. An alternative model, that is more customary in sociological and economical modeling, is the model in the right panel of Figure 3.1. In this model, called a formative model, the latent variable is regressed on its indicators. An example of a formative model is the measurement model for social economic status (SES). In such a model a researcher would, for example, record the variables income, educational level, and neighborhood as indicators of SES.

The models in Figure 3.1 are psychometrically and conceptually different (Bollen & Lennox, 1991). There is, however, no a priori reason why, in psychological measurement, one should prefer one type of measurement model to the other[2]. The measurement models that psychologists employ are typically of the reflective kind. Why is this?

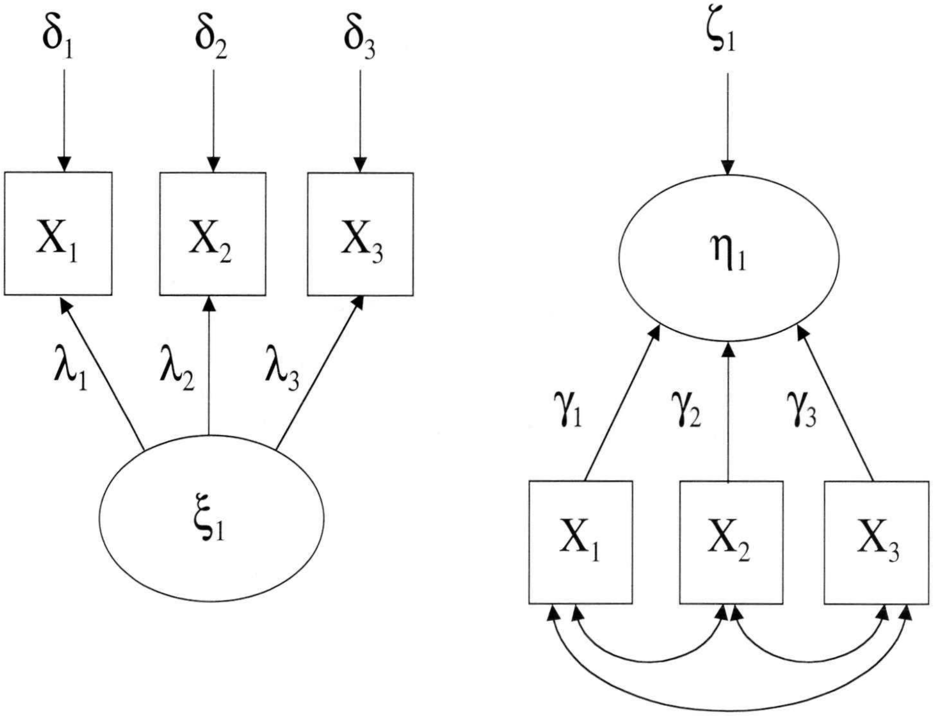
The obvious answer is that the choice of model depends on the ontology of the latent variables it invokes. A realist point of view motivates the reflective model, because the response on the questionnaire items is thought to vary as a function of the latent variable. In this case, variation in the latent variable precedes variation in the indicators. In ordinary language: dominant people will be more inclined to answer the questions affirmatively than submissive people. In this interpretation, dominance comes first and 'leads to' the item responses. This position implies a regression of the indicators on the latent variable, and thus motivates the choice of model. In the SES example, however, the relationship between indicators and latent variable is reversed. Variation in the indicators now precedes variation in the latent variable: SES changes as a result of a raise in salary, and not the other way around.

Latent variables of the formative kind are not conceptualized as determining our measurements, but as a summary of these measurements. These measurements may very well be thought to be determined by a number of underlying latent variables (which would give rise to the spurious model with multiple common causes of Edwards & Bagozzi, 2000), but we are not forced in any way to make such an assumption. Now, if we wanted to know how to weight the relative importance of each of the measurements comprising SES in predicting, say, health, we could use a formative model like that in the right panel of Figure 3.1. In such a model, we

[2] It is in itself an interesting (and neglected) question where to draw the line separating these classes of models at a content-level. For example, which of the formal models should be applied to the relation between diagnostic criteria and mental disorders in the DSM?

could also test whether SES acts as a single variable in predicting health. In fact, this predictive value would be the main motivation for conceptualizing SES as a single latent variable. However, nowhere in this development have we been forced to admit that SES exists independent of our measurements.

Figure 3.1. Two models for measurement. The figure in the left panel is the reflective measurement model. The X's are observed variables, ξ is the latent variable, λ's are factor loadings and the δ's are error terms. The right panel shows the formative model. The latent variable is denoted η, the γ's are the weights of the indicators, and ζ is a residual term.



The formative model thus does not necessarily require a realist interpretation of the latent variable that it invokes. In fact, if a realist interpretation were to be given, it would be natural to conceptualize this as a spurious model with multiple common causes in the sense of Edwards and Bagozzi (2000). This would again introduce a reflective part in the model, which would correspond to that part of the model that has a realist interpretation. Thus, the realist interpretation of a latent variable implies a reflective model, whereas constructivist, operationalist, or instrumentalist interpretations are more compatible with a formative model.

In conclusion, the standard model in psychological measurement is a reflective model that specifies that the latent variable is more fundamental than the item responses. This implies entity realism about the latent variable, at least in the hypothetical side of the argument (the assumptions of the model). Maybe more important than this is the fact that psychologists use the model in this spirit. In this context, Hacking's (1983) remark that "the final arbitrator in philosophy is not how we think but what we do" (p. 31) is relevant: the choice for the reflective measurement model in psychology expresses realism with respect to the latent variable.

Theory realism

Theory realism is different from entity realism in that it concerns the status of the theory, over and above the status of the entities that figure in the theory. It is therefore a stronger philosophical position. The realist interpretation of theories is naturally tied to a correspondence view of truth (O'Connor, 1975). This theory constructs truth as a 'match' between the state of affairs as posed by the theory and the state of affairs in reality, and is the theory generally endorsed by realists (Devitt, 1991). The reason why such a view is connected to realism is that, in order to have a match between theoretical relations and relations in reality, these relations in reality have to exist quite independent of what we say about them. For the constructivist, of course, this option is not open. Therefore, the constructivist will either deny the correspondence theory of truth and claim that truth is coherence between sentences (this is the so-called coherence theory of truth), or deny the relevance of the notion of truth altogether, for example by posing that not truth, but empirical adequacy (consistency of observations with predictions) is to be taken as the central aim of science (Van Fraassen, 1980).

The formal side of latent variable theory, of course, does not claim correspondence truth; it is a system of tautologies and has no empirical content. The question, however, is whether a correspondence type of assumption is formulated in the application of latent variable theory. There are three points in the application where this may occur. First, in the evaluation of the position of a subject on the latent variable; second, in the estimation of parameters; and third, in conditional reasoning based on the assumption that a model is true.

In the evaluation of the position of a subject on the latent variable, correspondence truth sentences are natural. The simple reason for this is that the formal theory implies that one could be wrong about the position of a given subject on the latent variable, which is only possible upon the assumption that there is a true position. To see this, consider the following. Suppose you have administered an intelligence test and you successfully fit a unidimensional latent variable model to the data. Suppose that the single latent variable in the model represents general intelligence. Now you determine the position on the latent variable for two subjects, say John and Jane Doe. You find that the weighted sumscore (i.e. the operational latent variable) is higher for John than for Jane, and you tentatively conclude that John has a higher position on the trait in question than Jane (i.e., you conclude that John is more intelligent). Now could it be that you have made a mistake, in that John actually has a lower score on the trait than Jane? The formal theory certainly implies that this is possible (in fact, this is what much of the theory is about; the theory will even be able to specify the probability of such a mistake, given the positions of John and Jane on the latent variable), so that the answer to this question must be affirmative. This forces commitment to a realist position because there must be something to be wrong about. That is, there must be something like a true (relative) position of the subjects on the latent trait in order for your assessment to be false. You can, as a matter of fact, never be wrong about a position on the latent variable if there is no true position on that variable. Messick (1989) concisely expressed this point when he wrote that "one must be an ontological realist in order to be an epistemological fallibilist" (p.26).

This argument is related to the second point in the application where we find a realist commitment, namely in the estimation of parameters. Here, we find essentially the same situation, but in a more general sense. Estimation is a realist concept: roughly speaking, one could say that the idea of estimation is only meaningful if there is something to be estimated. Again, this requires the existence of a true value: In a seriously constructivist view of latent variable analysis, the term 'parameter estimation' should be replaced by the term 'parameter determination'. For it is impossible to be wrong about something if it is not possible to be right about it. And estimation theory is largely concerned with being wrong: it is a theory about the errors one makes in the estimation process. At this point, one may object here that this is only a problem within a frequentist framework, because the idea of a true parameter value is typically associated with frequentism (Fisher, 1925; Hacking, 1965; Neyman & Pearson, 1967). It may further be argued that using Bayesian statistics (Novick & Jackson, 1974; Lee, 1997) could evade the problem. Within a Bayesian framework, however, the realist commitment becomes even more articulated. A Bayesian conception of parameter estimation requires one to specify a prior probability distribution over a set of parameter values. This probability distribution reflects one's degree of belief over that set of parameter values (De Finetti, 1974). Because it is a probability distribution, however, the total probability over the set of parameter values must be equal to one. This means that, in specifying a prior, one explicitly acknowledges that the probability (i.e., one's degree of belief) that the parameter actually has a value in the particular set is equal to one. In other words, one states that one is certain about that. The statement that one is certain that the parameter has a value in the set implies that one can be wrong about that value. And now we are back in the original situation: it is very difficult to be wrong about something if one cannot be right about it. In parameter estimation, this requires the existence of a true value.

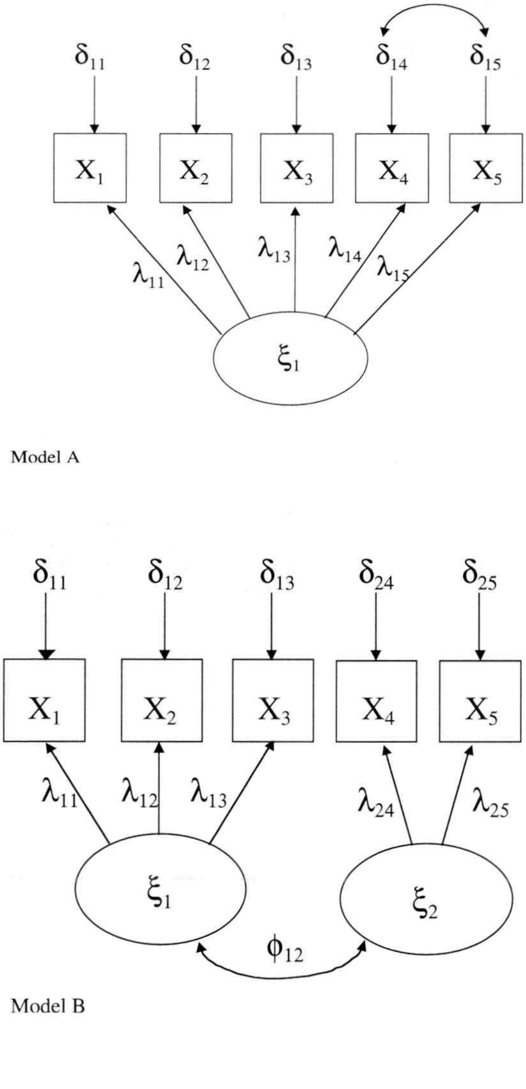
The third point in the application of latent variables analysis where we encounter correspondence truth is in conditionals that are based on the assumption that a model is true. In the evaluation of model fit, statistical formulations use the term 'true model'; for example, the p-value resulting from a likelihood ratio difference test between two nested models with a differing number of parameters is interpreted as the probability of finding this (or a more extreme) value for the corresponding chi-square, assuming that the most restricted model (i.e., the model that uses less parameters) is true. Psychometricians are, of course, aware of the fact that this is a very stringent condition for psychological measurement models to fulfill. So, in discussions on this topic, one often hears that there is no such thing as a true model (Cudeck & Browne, 1983; Browne & Cudeck, 1992). For example, McDonald & Marsh (1990) state that "...it is commonly recognized, although perhaps not explicitly stated, that in real applications no restrictive model fits the population, and all fitted restrictive models are approximations and not hypotheses that are possibly true" (p. 247). It would seem as if such a supposition, which is in itself not unreasonable, expresses a move away from realism. This is not necessarily the case. The supposition that there is no true model actually leaves two options: either all models are false or truth is not relevant at all. The realist, who adheres to a correspondence view of truth, must take the first option. The constructivist will take the second, and replace the requirement of truth with one of empirical adequacy.

If the first option is taken, the natural question to ask is: in what sense is the model false? Is it false, for example, because it assumes that the latent variable follows a normal distribution while this is not the case? So interpreted, we are still realists: there is a true model, but it is a different model from the one we specified, i.e., one in which the latent variable is not normally distributed. The fact that the model is false is, in this sense, contingent upon the state of affairs in reality. The model is false, but not necessarily false (i.e., it might be correct in some cases, but it is false in the present application). One could, upon this view, reformulate the statement that there is no such thing as a true model as the statement that all models are misspecified. That this interpretation of the sentence 'all models are false' is not contrary to, but in fact parasitic on realism, can be seen from the fact that the whole notion of misspecification requires the existence of a true model: For how can we misspecify if there is no true model? Now, we may say that we judge the (misspecified) model close enough to reality to warrant our estimation procedures. We then interpret the model as 'approximately true'. So, upon this interpretation, we are firmly in the realist camp, even though we acknowledge that we have not succeeded in formulating the true model. This is as far as a realist could go in the acknowledgement that our models are usually wrong. Popper (1963) was a realist who held such a view concerning theories.

The constructivist must take the second option and move away from the truth concept. The constructivist will argue that we should not interpret the statement that the model is true literally, but weaken the requirement to one of empirical adequacy. The whole concept of truth is thus judged irrelevant. The assumption that the model is true could then be restated as the assumption that the model fits the observable item response patterns perfectly at the population level. This renders the statistical assumption that a model is true (now interpreted as 'empirically adequate') meaningful, because it allows for disturbances in the observed fit due to random sampling, without assuming a realist view of truth. However, so interpreted, underdetermination rears up its ugly head.

For example, take a simple case of statistically equivalent covariance structure models such as the ones graphically represented in Figure 3.2. (taken from Hershberger, 1994). These models are empirically equivalent. This means that, if one of them fits the data, the other will fit the data equally well. If the assumption that model A is true is restated as the assumption that it is empirically adequate (i.e., it fits the item responses perfectly at the population level), the assumption that model A is true is fully equivalent to the assumption that model B is true.

Figure 3.2. Two equivalent models. The SEM-models in the figure predict the same variance-covariance matrix and are thus empirically equivalent. X's indicate observed variables, ξ's latent variables, λ's are factor loadings, δ's error terms, and ϕ is the correlation between latent variables.



Model A

Model B

Now try to reconstruct the estimation procedure. The estimation of the correlation between the latent variables ξ_1 and ξ_2 takes place under the assumption that model B is true. Under the empirical adequacy interpretation, however, this assumption is equivalent to the assumption that model A is true, for the adjective 'true' as it is used in statistical theory now merely refers to empirical adequacy at the population level. This implies that the assumption that model B is true may be replaced by the assumption that model A is true, for these assumptions are the same. However, this would mean that the correlation between the latent variables ξ_1 and ξ_2 can be estimated under the assumption that model A is true. In model A, however, there is only one latent variable. It follows that, upon the empirical adequacy view, the correlation between two latent variables can be estimated under the assumption that there is only one latent variable underlying the measurements. In my view, this is not particularly enlightening. But it must be said that the situation need not necessarily bother the constructivist, since the constructivist did not entertain a realist interpretation of these latent variables in the first place. However, it would take some ingenious arguments to defend this interpretation.

In sum, the evaluation of the position of a subject on the latent variable, the process of estimating parameters, and the conditional reasoning based on the assumption that a model is true, are characterized by realist commitments. It is difficult to interpret these procedures without an appeal to some sort of correspondence truth. This requires a substantial degree of theory realism. However, what I have shown is only that the natural interpretation of what we are doing in latent variables analysis is a realist one; not that it is the only interpretation. It may be that the constructivist could make sense of these procedures without recourse to truth. For now, however, I leave this task to the constructivist, and contend that theory realism is required to make sense of latent variables analysis.

Causality

The connection between the formal and the operational latent variable requires a realist ontology. The question then becomes what constitutes the relation between the latent variable and its indicators. Note that this question is not pressing for the operationalist, who argues that the latent variable does not signify anything beyond the data, which implies that the relation between the latent variable and its indicators is purely logical. Nor need it bother the constructivist, who argues that we construct this relation ourselves; it is not an actual but a mental relation, revealing the structure of our theories rather than a structure in reality. The realist will have to come up with something different, for the realist cannot maintain either of these interpretations.

The natural candidate, of course, is causality. That a causal interpretation may be formulated for the relation between latent variables and their indicators has been argued by several authors (e.g., Pearl, 1999, 2000; Edwards & Bagozzi, 2000; Glymour, 2001), and I will not repeat these arguments. The structure of the causal relation is known as a common cause relation (the latent variable is the common cause of its indicators) and has been formulated by Reichenbach (1956). Here, I will concentrate on the form of the relation in a standard measurement model. Specifically, I will argue that a causal connection can be defended in a between-subjects sense, but not in a within-subject sense.

For this purpose, we must distinguish between two types of causal statements that one can make about latent variable models. First, one can say that population differences in position on the latent variable cause population differences in the expectation of the item responses. In accordance with the repeated sampling interpretation, this interpretation posits no stochastic aspects within persons: The expectation of the item response is defined purely in terms of repeated sampling from a population of subjects with a particular position on the latent variable. Second, one can say that a particular subject's position on the latent variable causes his or her item response probabilities. This interpretation corresponds to the stochastic subject interpretation and does pose probabilities at the level of the individual. The first of these views can be defended, but the second is very problematic.

Between-subjects causal accounts To start with the least problematic, consider the statement that differences in the latent variable positions (between populations of subjects) causes the difference in expected item responses (between populations of subjects). This posits the causal relation at a between-subjects level. The statement would fit most accounts of causality, for example the three criteria of J.S. Mill (1843). These hold that X can be considered a cause of Y if a) X and Y covary, b) X precedes Y, and c) ceteris paribus, Y does not occur if X does not occur. In the present situation, we have a) covariation between the difference in position on the latent variable and the difference in expected item responses, b) upon the realist viewpoint, the difference in position on the latent variable precedes the difference in expected item responses, and c) if there is no difference in position on the latent variable, there is no difference in expected item responses. The between-subjects causal statement can also be framed in a way consistent with other accounts of causality, for example the counterfactual account of Lewis (1973), or the related graph-theoretical account of Pearl (1999; 2000). I conclude that a causal relation can be maintained in a between-subjects form. Of course, many problems remain. For example, most latent variables cannot be identified independent of their indicators. As a result, the causal account violates the criterion of separate identifiability of effects and causes, so that circularity looms. However, this is a problem for any causal account of measurement (Trout, 1999); and the main point is that the relation between the latent variable and its indicators can at least be formulated as a causal one.

Within-subject causal accounts The individual account of causality is problematic. Consider the statement that subject *i*'s position on the latent variable causes subject *i*'s item response. The main problem here is the following. One of the essential ingredients of causality is covariation. All theories of causality use this concept, be it in a real or in a counterfactual manner. If X is to cause Y, X and Y should covary. If there is no covariation, there cannot be causation (the reverse is of course not the case). One can say, for example, that striking a match caused the house to burn down. One of the reasons that this is possible, is that a change in X (the condition of the match) precedes a change in Y (the condition of the house). One cannot say, however, that subject *i*'s latent variable value caused his item responses, because there is no covariation between his position on the latent variable and his item responses. An individual's position on the latent variable is, in a standard measurement model, conceptualized as a constant, and a constant cannot be a cause. The same point is made in a more general context by Holland (1986) when he says that an attribute cannot be a cause.

Counterfactuals The obvious way out of this issue is to invoke a counterfactual account of causation (see, for example, Lewis, 1973; Sobel, 1994). On this account, one analyzes causality using counterfactual alternatives. This is done by constructing arguments such as 'X caused Y, because if X had not happened, ceteris paribus, Y would not have happened'. This is called a counterfactual account because X did in fact happen. For the previous example, one would have to say that 'the striking of the match caused the house to burn down, because the house would not have burned down if the match had not been struck'. For our problem, however, this account of causality does not really help. Of course, we could construct sentences like 'if subject i had had a different position on the latent variable, subject i would have produced different item responses', but this raises some difficult problems.

Suppose, for example, that one has administered Einstein a number of IQ-items. Consider the counterfactual 'if Einstein had been less intelligent, he would have scored lower on the IQ-items'. This seems like a plausible formulation of the hypothesis tested in a between-subjects model, and it also seems as if it adequately expresses the causal efficacy of Einstein's intelligence, but there are strong reasons for doubting whether this is the case. For example, we may reformulate the above counterfactual as 'if Einstein had had John's level of intelligence, he would have scored lower on the IQ-items'. But does this counterfactual express the causal efficacy of intelligence within Einstein? It seems to me that what we express here is not a within-subject causal statement at all, but a between-subjects conclusion in disguise, namely, the conclusion that Einstein scored higher than John because he is more intelligent than John. Similarly, 'if Einstein had had the intelligence of a fruitfly, he would not have been able to answer the IQ-items correctly' does not express the causal efficacy of Einstein's intelligence, but the difference between the population of humans and the population of fruitflies. We know that fruitflies act rather stupidly, and so are inclined to agree that Einstein would act equally stupidly if he had the intelligence of a fruitfly. And it seems as if this line of reasoning conveys the idea that Einstein's intelligence has some kind of causal efficacy. However, the counterfactual is completely unintelligible except when interpreted as expressing knowledge concerning the difference between human beings (a population) and fruitflies (another population). It does not contain information on the structure of Einstein's intellect, and much less on the alleged causal power of Einstein's intelligence. It only contains the information that Einstein will score higher on an IQ-test than a fruitfly because he is more intelligent than a fruitfly – but this is exactly the between-subjects formulation of the causal account. Clearly, the individual causal account transfers knowledge of between-subjects differences to the individual, and posits a variable that is defined between-subjects as a causal force within-subjects.

In other words, the within-subjects causal interpretation of between-subjects latent variables rests on a logical fallacy (the fallacy of division; Rorer, 1990). Once you think about it, this is not surprising. What between-subjects latent variables models do is to specify sources of between-subjects differences, but because they are silent with respect to the question of how individual scores are produced, they cannot be interpreted as posing intelligence as a causal force within Einstein. Thus, the right counterfactual (which is actually the one implied by the repeated sampling formulation of the measurement model) is between-subjects: the IQ-score we obtained from the i-th subject (who happened to be Einstein) would have been lower, had we drawn another subject with a lower position on the latent variable from the population. Note, however, that the present argument does not establish that it is impossible that some other conceptualization of intelligence may be given a causal within-subject interpretation. It establishes that such an interpretation is not formulated in a between subjects model, and therefore cannot be extracted from such a model: A thousand clean replications of the general intelligence model on between-subjects data would not establish that general intelligence plays a causal role in producing Einstein's item responses.

Exchangeability and local homogeneity But what about variables like, for example, height? Is it so unreasonable to say that 'if Einstein had been taller, he would have been able to reach the upper shelves in the library'? No, this is not unreasonable, but it is unreasonable to assume a priori that intelligence, as a between-subjects latent variable, applies in the same way as height does. The concept of height is not defined in terms of between-subjects differences, but in terms of an empirical concatenation operation (Krantz, Luce, Suppes, & Tversky, 1971; Michell, 1999; see also Chapter 4). Roughly, this means that we know how to move Einstein around in the height dimension (for example by giving him platform shoes), and that the effect of doing this is tractable (namely, wearing platform shoes will enable Einstein to reach the upper shelves). Moreover, it can be assumed that the height dimension applies to within-subject differences in the same way that it applies to between-subject differences. This is to say that the statements 'if Einstein had been taller, he would have been able to reach the upper shelves in the library' and 'if we had replaced Einstein with a taller person, this person would have been able to reach the upper shelves in the library' are equivalent with respect to the dimension under consideration. They are equivalent in this sense, exactly because the dimensions pertaining to within and to between subjects variability are qualitatively the same: If we give Einstein platform shoes which make him taller, he is, in all relevant respects, exchangeable with the taller person in the example. I do not object to introducing height in a causal account of this kind, because variations in height have demonstrably the same effect within and between subjects. But it remains to be shown that the same holds for psychological variables like intelligence.

The analogy does, however, provide an opening: The individual causal account could be defended on the assumption that intelligence is like height, in that the within-subjects and between-subjects dimensions are equivalent. However, the between-subjects model does not contain this equivalence as an assumption. Therefore, such an argument would have to rest on the idea that, by necessity, there has to be a strong relation between models for within-subjects variability and models for between-subjects variability. It turns out that this idea is untenable. The reason for this is that there is a surprising lack of relation between within-subjects models and between-subjects models.

To discuss within-subject models, we now need to extend our discussion to the time domain. This is necessary, because to model within-subjects variability, there has to be variability, and variability requires replications of some kind; and if variability cannot result from sampling across subjects, it has to come from sampling within subjects. In this paradigm, one could, for example, administer Einstein a number of IQ-items repeatedly over time, and analyze the within-subject covariation between item responses. The first technique of this kind was Cattell's so called P-technique (Cattell & Cross, 1952), and the factor analysis of repeated measurements of an individual subject has been refined, for example, by Molenaar (1985). The exact details of such models need not concern us here; what is important is that, in this kind of analysis, systematic covariation over time is explained on the basis of within-subject latent variables. So, instead of between-subjects dimensions that explain between-subjects covariation, we now have within-subject dimensions that explain within-subject covariation. One could imagine that, if the within-subject model for Einstein had the same structure as the between-subjects model, then the individual causal account would make sense despite all the difficulties we encountered above.

In essence, such a situation would imply that the way in which Einstein differs from himself over time is qualitatively the same as the way in which he differs from other subjects at one single time point. This way, the clause 'if Einstein were less intelligent' would refer to a possible state of Einstein at a different time point, however hypothetical. More importantly, this state would, in all relevant respects, be identical to the state of a different subject, say John, who is less intelligent at this time point. In such a state of affairs, Einstein and John would be exchangeable, like a child and a dwarf are exchangeable with respect to the variable height. It would be advantageous, if not truly magnificent, if a between-subjects model would imply or even test such exchangeability. This would mean, for example, that the between-subjects five factor model of personality would imply a five factor model for each individual subject. If this were to be shown, the case against the individual causal account would reduce from a substantial objection to a case of philosophical hairsplitting. However, the required equivalence has not been shown, and the following reasons lead me to expect that it will not, in general, be a tenable assumption.

The link connecting between-subjects variables to characteristics of individuals is similar to the link discussed in the stochastic subject formulation of latent variable models, where the model for the individual is counterfactually defined in terms of repeated measurements with intermediate brainwashing. I have already mentioned that Ellis & Van den Wollenberg (1993) have shown that the assumption that the measurement model holds for each individual subject (local homogeneity) has to be added to and is in no way implied by the model. One may, however, suppose that, while finding a particular structure in between-subjects data may not imply that the model holds for each subject, it would at least render this likely. Even this is not the case. It is known that if a model fits in a given population, this does not entail the fit of the same model for any given element from a population, or even for the majority of elements from that population (Molenaar, 1999; Molenaar, Huizenga, & Nesselroade, in press).

So, the five factors in personality research are between subjects; but if a within-subjects time series analysis would be performed on each of these subjects, we could get a different model for each of the subjects. In fact, Molenaar et al. (in press) have performed simulations in which they had different models for each individual (so, one individual followed a one factor model, another a two factor model, etc.). It turned out that, when a between-subjects model was fitted to between-subjects data at any specific time point, a factor model with low dimensionality (i.e., a model with one or two latent variables) provided an excellent fit to the data, even if the majority of subjects had a different latent variable structure.

With regard to the Five Factor Model in personality, substantial discrepancies between intraindividual and interindividual structures have been empirically demonstrated in Borkenau & Ostendorf (1998). Mischel & Shoda (1995), Feldman (1995), and Cervone (1997) have illustrated similar discrepancies between intraindividual and interindividual structures. This shows that between-subjects models and within-subject models bear no obvious relation to each other, at least not in the simple sense discussed above. This is problematic for the individual causal account of between-subjects models, because it shows that the premise 'if Einstein were less intelligent...' cannot be supplemented with the conclusion '...then his expected item response pattern would be identical to John's expected item response pattern'. It cannot be assumed that Einstein and John (or any other subject, for that matter) are exchangeable in this respect, because, at the individual level, Einstein's intelligence structure may differ from John's in such a way that the premise of the argument cannot be fulfilled without changing essential components of Einstein's intellect. Thus, the data generating mechanisms at the level of the individual are not captured, not implied, and not tested by between-subjects analyses without heavy theoretical background assumptions which, in psychology, are simply not available.

The individual causal account is not merely implausible for philosophical or mathematical reasons; for most psychological variables, there is also no good theoretical reason for supposing that between-subjects variables do causal work at the level of the individual. For example, what causal work could the between-subjects latent variable we call general intelligence do in the process leading to Einstein's answer to an IQ-item? Let us reconstruct the procedure. Einstein enters the testing situation, sits down, and takes a look at the test. He then perceives the item. This means that the bottom-up and top-down processes in his visual system generate a conscious perception of the task to be fulfilled: it happens to be a number series problem. Einstein has to complete the series 1, 1, 2, 3, 5, 8, ..? Now he starts working on the problem; this takes place in working memory, but he also draws information from long-term memory (for example, he probably applies the concept of addition, although he may also be trying to remember the name of a famous Italian mathematician of whom this series reminds him). Einstein goes through some hypotheses concerning the rules that may account for the pattern in the number series. Suddenly he has the insight that each number is the sum of the previous two (and simultaneously remembers that it was Fibonacci!). Now he applies that rule and concludes that the next number must be 13. Einstein then goes through various motorical processes which result in the appearance of the number 13 on the piece of paper, which is coded as '1' by the person hired to do the typing. Einstein now has a 1 in his response pattern, indicating that he gave a correct response to the item. This account has used various psychological concepts, such as working memory, long term memory, perception, consciousness, and insight. But, where, in this account of the processes leading to Einstein's item response, did intelligence enter? The answer is: nowhere. Intelligence is a concept that is intended to account for individual differences; and the model that we apply is to be interpreted as such. Again, this implies that the causal statement drawn from such a measurement model retains this between-subjects form.

Elliptical accounts The last resort for anyone willing to endorse the individual causal account of between-subjects models is to view the causal statement as an elliptical (i.e., a shorthand) explanation. The explanation for which it is a shorthand would, in this case, be one in terms of processes taking place at the individual level. This requires stepping down from the macro-level of repeated testing (as conceptualized in the within subjects modeling approach) to the micro-level of the processes leading up to the item response in this particular situation. I will argue in the next paragraph that there is merit to this approach in several respects, but it does not really help in the individual causal account as discussed in this section. The main reason for this is that the between-subjects latent variable will not indicate the same process in each subject. Therefore, the causal agent (i.e., the position on the latent variable) that is posited within subjects based on a between-subjects model does not refer to the same process in all subjects.

This is a problem for an elliptical account. For instance, one can say that the Titanic has rusted after so many years on the bottom of the sea, because it was made of iron. This explanation is elliptical, because it does not specify all processes that actually lead to the phenomenon we call rust. The reason why the explanation works, however, is that the explanation subsumes the Titanic under the category of iron things, and this category is homogeneous with respect to the processes that will occur when such things are left on the bottom of the ocean. Thus, one may look up the details of the reaction between Fe and H_2O that leads to rust, and unproblematically take these processes to apply to the Titanic. One could say that the category of iron things displays *process homogeneity* with respect to the situation at hand.

In psychological measurement, such process homogeneity is not to be expected in most cases. This is a particularly pressing problem for models that posit continuous latent variables. The reason for this is that an elliptical explanation would probably refer to a qualitatively different process for different positions on the latent variable; probably even to different processes for different people with the same position on the latent variable. Jane, high on the between-subjects dimension general intelligence, will in all likelihood approach many IQ-items using a strategy that is qualitatively different from her brother John's. John and his nephew Peter, equally intelligent, may both fail to answer an item correctly, but for different reasons (e.g., John has difficulties remembering series of patterns in the Raven task, while Peter has difficulties in imagining spatial rotations). It is obvious that this problem is even more serious in personality testing, where we generally do not even have the faintest idea of what happens between item administration and item response. For this reason, it would be difficult to conceive of a meaningful interpretation of such an elliptical causal statement without rendering it completely vacuous, in the sense that the position on the latent variable is a shorthand for whatever process leads to person's response. In such an interpretation, the within-subject causal account would be trivially true, but uninformative.

However, it must be said that in this case latent class models could have an advantage. For instance, in the models used to model children's responses on the balance scale task (Jansen & Van der Maas, 1997), latent class es are considered to be homogeneous with respect to the strategy used to solve the items. In this case, the classes do have process homogeneity, and an elliptical explanation could be defended. The line of reasoning followed in such models could, of course, be extended and could lead to valid elliptical explanations of respondent behavior. Unfortunately, at present such cases of theoretically inspired modeling are rare.

On the basis of this analysis, we must conclude that the within-subject causal statement, that subject i's position on the latent variable causes his item responses, does not sit well with existing accounts of causality. A between-subjects causal relation can be defended, although it is certainly not without problems. Such an interpretation conceives of latent variables as sources of individual differences, but explicitly abstracts away from the processes taking place at the level of the individual. The main reason for the failure of the within-subjects causal account seems to be that it rests on the misinterpretation of a measurement model as a process model, that is, as a mechanism that operates at the level of the individual (see Krueger, 1999, for an explicit example of this fallacy, and Borsboom, 2002, for a criticism).

This fallacy is quite pervasive in the behavioral sciences. For instance, part of the nature-nurture controversy, as well as controversies surrounding the heritability coefficients used in genetics, may also be due to this misconception. The fallacious idea, that a heritability coefficient of .50 for IQ-scores means that 50% of an individual's intelligence is genetically determined, remains one of the more pervasive misunderstandings in the nature-nurture discussion. Ninety percent of variations in height may be due to genetic factors, but this does not imply that my height is for 90% genetically determined. Similarly, a linear model for interindividual variations in height does not imply that individual growth curves are linear; that 30% of the interindividual variation in success in college may be predicted from the grade point average in high school, does not mean that 30% of the exams you passed were predictable from your high school grades; and that there is a sex difference in verbal ability does not mean that your verbal ability will change if you undergo a sex change operation. It will be clear to all that these interpretations are fallacious. Still, for some reason, such misinterpretations are very common in the interpretation of results obtained in latent variables analysis. However, they can all be considered to be specific violations of the general statistical maxim, that between-subjects conclusions should not be interpreted in a within-subjects sense.

3.3 Implications for psychology

It is clear that between-subjects models do not imply, test, or support causal accounts that are valid at the individual level. In turn, the causal accounts that can be formulated and supported in a between-subjects model do not address individuals. However, connecting psychological processes to the latent variables that are so prominent in psychology is of obvious importance. It is essential that such efforts be made, because the between-subjects account in itself does not correspond to the kind of hypotheses that many psychological theories would imply, as these theories are often formulated at the level of individual processes. The relation (or relations) that may exist between latent variables and individual processes should therefore be studied in greater detail, and preferably within a formalized framework, than has so far been done. In this section, I provide an outline of the different ways in which the relation between individual processes and between-subject latent variables can be conceptualized. These different conceptualizations correspond to different kinds of psychological constructs. They also generate different kinds of research questions and require different research strategies in order to substantiate conclusions concerning these constructs.

Locally homogeneous constructs First, theoretical considerations may suggest that a latent variable is at the appropriate level of explanation for both between-subjects and within-subjects differences. Examples of psychological constructs that could be conceptualized in this manner are various types of state-variables such as mood, arousal, or anxiety, and maybe some attitudes. That is, it may be hypothesized, for differences in the state variable 'arousal', that the dimension on which I differ from myself over time, and the dimension on which I differ from other people at a given time point, are the same. If this is the case, the latent variable model that explains within-subjects differences over time must be the same model as the model that explains between-subjects differences. Fitting latent variable models to time series data for a single subject is possible (Molenaar, 1985), and such techniques suggest exploring statistical analyses of case studies in order to see whether the structure of the within-subject latent variable model matches between-subjects latent variables models. If this is the case, there is support for the idea that we are talking about a dimension that pertains both to variability within a subject and between-subjects variability. Possible states of a given individual would then match possible states of different individuals, which means that, in relevant respects, the exchangeability condition discussed in the previous section holds. Thus, in this situation we may say that a latent variable does explanatory work both at the within-subject and the between-subjects level, and a causal account may be set up at both of these. Following the terminology introduced by Ellis & Van den Wollenberg (1993) I propose to call this type of construct *locally homogeneous*, where 'locally' indicates that the latent variable structure pertains to the level of the individual, and 'homogeneous' refers to the fact that this structure is the same for each individual.

Locally heterogeneous constructs Locally homogeneous constructs will not often be encountered in psychology, where myriads of individual differences can be expected to be the rule rather than the exception. I would not be surprised if, for the majority of constructs, time series analyses on individual subjects would indicate that different people exhibit different patterns of change over time, which are governed by different latent variable structures. So, for some people, psychological distress may be unidimensional, while for others it may be multidimensional. If this is the case, it would seem that we cannot lump these people together in between-subjects models to test hypotheses concerning psychological processes, for they would constitute a heterogeneous population in a theoretically important sense. At present, however, we do not know how often and to what degree such a situation occurs, which makes this one of the big unknowns in psychology. This is because there is an almost universal - but surprisingly silent - reliance on what may be called a uniformity of nature assumption in doing between-subjects analyses; the relation between mechanisms that operate at the level of the individual and models that explain variation between individuals is often taken for granted, rather than investigated. For example, in the attitude literature (Cacioppo & Berntson, 1999; Russell & Carroll, 1999) there is currently a debate on whether the affective component of attitudes is produced by a singular mechanism, which would produce a bipolar attitude structure (with positive and negative affect as two ends of a single continuum), or should be conceptualized as consisting of two relatively independent mechanisms (one for positive, and one for negative affect). This debate is characterized by a strong uniformity assumption: It either is a singular dimension (for everyone), or we have two relatively independent subsystems (for everyone). It is, however, not obvious that the affect system should be the same for all individuals; for it may turn out that the affective component in attitudes is unidimensional for some people but not for others. It must be emphasized that such a finding would not render the concept of attitude obsolete; but clearly, a construct governed by different latent variable models within different individuals will have to play a different role in psychological theories than a locally homogeneous construct. I propose to call such constructs *locally heterogeneous*. Locally heterogeneous constructs may have a clear dimensional structure between subjects, but they pertain to different structures at the level of individuals. Thus, we now have a distinction between two types of constructs: locally homogeneous constructs, for which the latent dimension is the same within and between subjects, and locally heterogeneous constructs, for which this is not the case. Locally homogeneous constructs allow for testing hypotheses concerning individual processes, modules, and subsystems, through the analysis of between-subjects variability, while locally heterogeneous constructs do not. In applications, it is imperative that we find out about which of the two we are talking, especially when we are testing hypotheses concerning processes at the individual level with between-subjects models.

Locally irrelevant constructs It will be immediately obvious that constructs which are hypothesized as relatively stable traits, such as the factors in the Big Five, will not exhibit either of these structures. If a trait is stable, covariation of repeated measurements will not obey a latent variable model at all. All variance of the observed variables will be error variance, so that this implies that these observed variables will be independent over time. This hypothesis could, and should, be tested using time series analysis. If it holds, the latent variable in question would be one that produces between-subjects variability, but does no work at the individual level. I propose to call this type of construct a *locally irrelevant* construct. This terminology should not be misread as implying a value judgment, as locally irrelevant constructs have played, and will probably continue to play, an important role in psychology. However, the terminology should be read unambiguously as indicating the enormous degree to which such constructs abstract from the level of the individual. They should, for this reason, not be conceptualized as explaining behavior at the level of the individual. In the personality literature, this has been argued on independent grounds by authors such as Lamiell (1987), Pervin (1994), and Epstein (1994).

It is disturbing, and slightly embarrassing for psychology, that we cannot say with sufficient certainty in which of these classes particular psychological constructs (e.g., personality traits, intelligence, attitudes) fall. This is the result of a century of operating on silent uniformity of nature assumptions by focussing almost exclusively on between-subjects models. It seems that psychological research has adapted to the limitations of common statistical procedures (for example, by abandoning case studies because analysis of variance requires sample sizes larger than one), instead of inventing new procedures that allow for the testing of theories at the proper level, which is often the level of the individual, or at the very least exploiting time series techniques that have been around in other disciplines (e.g., econometrics) for a very long time. Clearly, extending measurements into the time domain is essential, and fortunately the statistical tools for doing this are rapidly becoming available. Models that are suited for this task have seen substantial developments over the last two decades (see, for example, Molenaar, 1985; McArdle, 1987; Wilson, 1989; Fischer & Parzer, 1991), and powerful, user friendly software for estimating and testing them has been developed (Jöreskog & Sörbom, 1993; Muthén & Muthén, 1998; Neale, Boker, Xie, & Maes, 1999). Especially, it would be worthwhile to try latent variable analyses at the level of the individual, which would bring the all but abandoned case study back into scientific psychology - be it, perhaps, from an unexpected angle.

Ontology revisited There remains an open question pertaining to the ontological status of latent variables, and especially those that fall into the class of locally irrelevant constructs. It has been argued here that latent variables, at least those of the reflective kind, imply a realist ontology. How should we conceptualize the existence of such latent variables, if they cannot be found at the level of the individual? It seems that the proper conceptualization of the latent variable (if its reality is maintained) is as an emergent property, in the sense that it is a characteristic of an aggregate (the population) which is absent at the level of the constituents of this aggregate (individuals). Of course, this does not mean that there is no relation between the processes taking place at the level of the individual and between-subjects latent variables. In fact, the between-subjects latent variable must be parasitic on individual processes, because these must be the source of between-subjects variability. If it is shown that a given set of cognitive processes leads to a particular latent variable structure, we could therefore say that this set of processes realizes the latent variables in question. The relevant research question for scientists should then be: which processes generate which latent variable structures? What types of individual processes, for example in intelligence testing, are compatible with the general intelligence model?

Obviously, time series analyses will not provide an answer to this question in the case of constructs that are hypothesized to be temporally stable, such as general intelligence. In this case, we need to connect between subjects models to models of processes taking place at the level of the individual. This may involve a detailed analysis of cognitive processes that are involved in solving IQ-test items, for example. Such inquiries have already been carried out by those at the forefront of quantitative psychology. Embretson (1994), for example, has shown how to build latent variable models based on theories of cognitive processes; and one of the interesting features of such inquiries is that they show clearly how a single latent variable can originate, or emerge, out of a substantial number of distinct cognitive processes. This kind of research is promising and may lead to important results in psychology. I would not be surprised, for example, if it turned out that Sternberg's (1985) triarchic theory of intelligence, which is largely a theory about cognitive processes and modules at the level of the individual, is not necessarily in conflict with the between-subjects conceptualization of general intelligence. Finally, I note that the connection of cognitive processes and between-subjects latent variables requires the use of results from both experimental and correlational psychological research traditions, which Cronbach (1957) has called the two disciplines of scientific psychology. This section may therefore be read as a restatement of his call for integration of these schools.

3.4 Discussion

Latent variable models introduce a hypothetical attribute to account for relations among observable variables. In a measurement context, they assert that a number of items measure the same latent variable. This requires a realist ontology for the latent variable, and a good deal of theory realism for the postulated model. In comparison to classical test theory, latent variable theory is certainly a substantial improvement. It specifies a relation between item responses and the attribute measured, which means that it can be properly considered to give a theory of measurement. Upon closer examination, however, the specific interpretation of the measurement relation is not without problems. Given the realist interpretation of latent variables, causality can be considered a natural candidate, and formulated in terms of subpopulation distributions, a causal account can indeed be defended. The within-subject interpretation of the model, however, is extremely problematic.

Before I discuss some implications of these results, there are two important asides to make concerning what I am not saying. First, it is not suggested here that one cannot use a standard measurement model, and still think of the latent variable as constructed out of the observed variables or as a fiction. But I do insist that this is an inconsistent position, in that it cannot be used to connect the operational latent variable to its formal counterpart in a consistent way. Whether one should or should not allow such an inconsistency in one's reasoning is a different question that is beyond the scope of this chapter. Second, if one succeeds in fitting a latent variable model in a given situation, the present discussion does not imply that one is forced to believe in the reality of the latent variable. In fact, this would require a logical strategy known as 'inference to the best explanation' or 'abduction', which is especially problematic in the light of underdetermination. So I am not saying that, for example, the fit of a factor model with one higher order factor to a set of IQ measurements implies the existence of a general intelligence factor: what I am saying is that the consistent connection between the empirical and formal side of a factor model requires a realist position. Whether realism about specific instances of latent variables, such as general intelligence, can be defended is an epistemological issue that is the topic of heated discussion in the philosophy of science (see, for example Van Fraassen, 1980; Cartwright, 1983; Hacking, 1983; Devitt, 1991). Probably, on the epistemological side of the problem, there are few latent entities in psychology that fulfill the epistemological demands of realists such as Hacking (1983).

It will be felt that there are certain tensions in the application of latent variable models to psychological measurement. I have not tried to cover these up, because I think they are indicative of some fundamental problems in psychological measurement and require a clear articulation. The realist interpretation of latent variable theory leads to conclusions that will seem too strong for many psychologists. Psychology has a strong empiricist tradition and psychologists often do not want to go beyond the observations - at least, no further than strictly necessary. As a result, there is a feeling that realism about latent variables takes us too far into metaphysical speculations. At the same time, we would probably like latent variable models to yield conclusions of a causal nature (the model should at the very least allow for the formulation of such relations). But we cannot defend any sort of causal structure invoking latent variables, if we are not realists about these latent variables, in the sense that they exist independent of our measurements: One cannot claim that A causes B, and at the same time maintain that A is constructed out of B. If we then reluctantly accept realism, invoking perhaps more metaphysics than we would like, it appears that the type of causal conclusions available are not the ones we desired. Namely, the causality in our measurement models is only consistently formulated at the between-subjects level. And although the boxes, circles, and arrows in the graphical representation of the model suggest that the model is dynamic and applies to the individual, upon closer scrutiny no such dynamics are to be found. Indeed, this has been pinpointed as one of the major problems of mathematical psychology by Luce (1997): our theories are formulated in a within-subjects sense, but the models we apply are often based solely on between-subjects comparisons.

What are the consequences of this problem for the conception of psychological measurement that latent variable theory offers? It depends on how you look at it. If one accepts the possibility that a causal account can apply to characteristics of populations, without applying to each element of these populations, the problems are relatively small. Such causal accounts are not uncommon: Variation in the variable 'smoking' causes variation in the variable 'cancer', but it does not do so for each person. Still, I think that causality can be meaningfully applied in this case, be it with the understanding that its validity at the population level does not imply that the causal relation holds for each individual. Upon such a view, one does have to settle for a measurement relation that is solely expressed in terms of variation: Variation on the latent variable causes variation on the observables, but for a single person the latent variable does not have to play a role in this respect. One could argue against this view by saying that, if a causal model is invalid for each individual, then it cannot be valid in the population. Upon this view, a causal account of the measurement process is impossible in the locally heterogeneous and locally irrelevant cases. I think such a view is too restrictive, because it would imply that it is impossible to measure between-subjects differences in attributes, if these attributes are inherently stable within-subjects. This would mean, for instance, that genotypic differences cannot be measured through phenotypic effects. However, if the purpose of a measurement procedure is to measure differences between subjects, then one cannot hold it against the procedure that its results do not apply to differences within subjects. It does seem that these are radically different levels of explanation, and therefore they should not be mixed up.

The same causal account of measurement can be set up within persons, of course, and in the special case that the between-subjects and the within-subjects accounts are both valid, one is in the lucky position to draw within-subject conclusions on the basis of between-subjects data. Whether this assumption applies, how one could gather evidence for it, and which constructs are supposed to be candidates for it in the first place, are important but neglected questions in psychology, as has been argued in this chapter. However, if one takes the position that measurement can apply to sources of variation in a population, without applying directly to the individuals that make up this population, then latent variable theory does not necessarily disqualify as a theory of measurement in the locally heterogeneous and locally irrelevant cases. It may be that the analysis given suggests that we are not measuring the right things, i.e., that we are not investigating what we would want to investigate, but this is not a conceptual problem for latent variable theory. It is a conceptual problem for psychology and for the way it utilizes latent variable models.

For now, I contend that latent variable theory can offer a quite elegant account of the measurement process. The theory has several notable benefits. First, it places the attribute in the measurement model in a way that seems very plausible: Differences in the attribute (either within or between subjects) lead to differences in the observations. It is clear that such a view requires both realism about the attribute and a causal interpretation of the measurement process. Second, although this view introduces some heavy metaphysics, the metaphysics are clearly necessary, serve a clear purpose, and in fact lead to some interesting research questions. This is a substantial improvement over the classical test theory model, which has metaphysics wandering all over the place for no clear purpose except to be able to construct mathematically simple equations. Third, the latent variable view seems to align closely with the way many working researchers think about measurement.

This property cannot be ascribed to the classical test model, and neither to the fundamental measurement model, as will be argued in the next chapter. The latent variable model is, of course, in danger of misinterpretation. However, if the fact that a technique is easily misinterpreted were to be held against it, methodology and statistics would probably be empty within a day. At present, latent variable theory must be considered to formulate a plausible philosophy of measurement.

4. SCALES

It may be that the task of the new psychometrics is impossible; that fundamental measures will never be constructed. If this is the case, then the truth must be faced that perhaps psychology can never be a science...
– Paul Kline, 1998

4.1 Introduction

In the 1930's, the British Association for the Advancement of Science installed a number of its members with a most peculiar task: To decide whether or not there was such a thing as measurement in psychology. The commission, consisting of psychologists and physicists (among the latter was Norman Campell, famous for his philosophical work on measurement), was unable to reach unanimous agreement. However, a majority of its members concluded that measurement in psychology was impossible; Campbell (cited in Narens & Luce, 1986, p. 186), for example, asked "why do not psychologists accept the natural and obvious conclusion that subjective measurements (...) cannot be the basis of measurement". Similarly, Guild (cited in Reese, 1943, p.6) stated that "to insist on calling these other processes [i.e., attempts at psychological measurement] measurement adds nothing to their actual significance, but merely debases the coinage of verbal intercourse. Measurement is not a term with some mysterious inherent meaning, part of which may be overlooked by the physicists and may be in course of discovery by psychologists". For this reason, Guild concluded that using the term 'measurement' to cover quantitative practices in psychology "does not broaden its meaning but destroys it". Reese (1943, p. 6) summarized the ultimate position of the commission: "They [the members of the commission] argue that psychologists must then do one of two things. They must either say that the logical requirements for measurement in physics, as laid down by the logicians and other experts in the field of measurement, do not hold for psychology, and then develop other principles that are logically sound; or they must admit that their attempts at measurement do not meet the criteria and both cease calling these manipulations by the word 'measurement' and stop treating the results obtained as if they were the products of true measurement".

It would seem that the members of the commission anticipated that the alternative 'logically sound' principles for 'true measurement' in psychology would probably never be discovered. But perhaps they did anticipate their report to have the desired impact in the sense that psychologists would finally recognize their errors, and would stop the unauthorized use of terms like measurement and quantity. Interestingly, exactly the opposite has happened: Psychologists *have* developed an alternative, but generally use the term 'measurement' to denote every procedure of assigning numbers *except* the logically 'correct' one. That is, the theory of fundamental measurement (the 'true' measurement theory to which Guild refers) has been extended in such a manner that 'logically sound' principles have become available for psychological measurement situations, primarily through the development of conjoint measurement structures (Luce & Tukey, 1964; Krantz, Suppes, Luce, & Tversky, 1971). Ironically, however, not a soul uses that theory in the practice of psychological measurement: Every year there appears an enormous number of books that have 'psychological measurement' in the title, but few of them even contain a reference to this work. The logical foundation for psychological measurement has thus become available, only to be neglected by its presumed audience – and psychologists have continued to use the term measurement for everything else.

The gist of what has been called the 'axiomatic' approach to measurement (Cliff, 1992), of which the theory of fundamental measurement can be considered a special case, is that measurement is an essentially *representational* activity, i.e., a process of assigning numbers in such a manner as to preserve basic qualitative relations observed in the world (Narens & Luce, 1986). The result of this activity is called a measurement scale. Psychologists are familiar with this concept mainly through Stevens' (1946) famous typology of 'levels of measurement' in terms of nominal, ordinal, interval, and ratio scales. The scale type is often deemed very important for determining what kind of statistics may be used, and in this manner it exerts considerable influence on the practice of data analysis in psychology (or, in any event, on the conscience of psychologists doing the analyses). The prescriptive aspect of scales has been the subject of enduring controversies between measurement theoreticians and statisticians (Lord, 1953; Stevens, 1968; Gaito, 1980; Townshend & Ashby, 1984; Michell, 1986; Velleman & Wilkinson, 1993), mainly because statisticians refuse to be told what is admissible and what not by what they seem to perceive as an utterly impractical theory (Lord, 1953; Gaito, 1980). However, apart from generating such controversies and acting on the psychologist's statistical conscience, scales and the associated theory of measurement have not entered mainstream psychology at all (Cliff, 1992).

This does not mean that nobody works with representationalism in psychology. The original developers of the theory, such as Luce, Suppes, and Narens, continue to work out the mathematical basis of measurement theory, joined by a group of researchers united in the Society for Mathematical Psychology. In a completely different corner of psychology, the advocates of Rasch measurement frequently allude to the fundamental measurement properties of the Rasch model; notable in this context are Wright (1997), Roskam (1984), and Bond & Fox (2001). Finally, at a more conceptual level Michell (1990; 1997) has attacked the common practice in psychology and psychometrics using a line of reasoning based on the axiomatic theory of measurement. His efforts have had impact on at least one psychometrician

(Kline, 1998), and may well influence more. These researchers look to the future, and some of them seem to regard the coming of the "revolution that never happened" (Cliff, 1992) as the only road to a truly scientific psychology (Kline, 1998; Bond & Fox, 2001). Or, like Luce (1996, p. 95), they view such developments as simply "inevitable", so that "the only question is the speed with which they are carried out".

The axiomatic theory of measurement thus has a certain apologetic quality about it. It is also strongly normative, or even prescriptive, as is evidenced by terminology such as 'admissible transformations', and the idea that performing an inadmissible transformation destroys the 'meaningfulness' of conclusions based on the data (see Michell, 1986, for a discussion of this view). Now, methodology is in a sense always normative, but there is no approach in psychological measurement - not even in latent variables analysis - that so pertinently presents itself as the gatekeeper of rationality. Treatises based on the approach also insist on empirical testability of hypotheses in a manner that almost suggests that, if a hypothesis cannot be directly tested, it is meaningless, or at the very least suspect. For example, Michell (2000) has characterized the a priori assumption that psychological attributes are quantitative, which indeed is a strong metaphysical assumption in many latent variable models, as a methodological thought disorder, and this leads him to label the entire field of psychometrics as pathological. The reason for this disqualification seems to be that the hypothesis is not directly testable in commonly used models like the factor model. Those familiar with the philosophy of science may see a parallel with a historical movement that shared both the strong normativity, the desire to demarcate between meaningful and meaningless propositions, and the emphasis on the dangers of metaphysics – namely, the Vienna Circle. In this respect it is interesting to note that fundamental measurement theory originated in roughly the same period as logical positivism – a mere two years separate Campbell's (1920) *Physics: The elements* and Wittgenstein's (1922) *Tractatus Logico-Philosophicus*. There seems to be a certain similarity between, on the one hand, the divide between the empiricist, verificationist orientation of logical positivism and the robust realist, falsificationist philosophy of Popper (1959), and, on the other hand, the schism between representational measurement theory and the latent variables approach.

This chapter develops this intuition by inquiring into the status of the measurement scale, the central concept of representational measurement. This inquiry serves two purposes. First, in view of the critical commentaries of Michell (1990; 1999; 2000), Kline (1998), and the Rasch movement (Bond & Fox, 2001), it is important to scrutinize the axiomatic approach to measurement carefully – not only with respect to its alleged normative force, but also with respect to the philosophical ideas on which it is based. But second, the present chapter will add considerable clarification to the strong conclusions reached in the previous chapter, by showing what a truly empiricist theory of measurement looks like. For representational measurement theory, when compared to the latent variables approach, is almost devoid of metaphysics. It explicitly recognizes that measurement scales are constructions, and in fact builds upon this idea in a way that, it must be said, is consistent, elegant, and powerful. Therefore, the representational measurement approach introduces a sharp contrast, which brings out the realism inherent in latent variable models stronger than any argument could do by itself. However, the present investigation will show that the idea, that measurement is a representational activity, is unsatisfying on a number of counts. In fact, it is argued here that representationalism fails to address some crucial issues in psychological measurement. The relevance of representational measurement for psychological research is therefore concluded to be limited.

4.2 Three perspectives on measurement scales

Representational measurement theory is aimed at specifying the conditions necessary for the construction of an adequate representation of empirical relations in a numerical system. From a formal perspective, this is conceptualized in terms of a mapping of one set of relations into another. The resulting representation is considered adequate if it preserves the observed, empirical relations. Semantically, the interpretation of the measurement process is in terms of a reconstruction of the measurement process. For example, numerical operations are conceptualized as corresponding to empirical operations, even though no scientist ever carried out these operations in the manner described by the theory. From an ontological perspective, scales cannot be considered anything but a construction. It could, of course, be held that these scales have referents in reality, for example objective magnitudes. However, such a realist interpretation, if endorsed, is external to the model, in contrast to the inherent realism in latent variables analysis.

4.2.1 The formal stance

Syntax Representational measurement theory constructs measurement as the mapping of objects and relations between objects from an empirical domain into a numerical domain. Both are characterized in terms of set theory (Scott & Suppes, 1958; Suppes & Zinnes, 1963). We imagine a set of objects, which is is denoted A, and a set of n relations holding between these objects, denoted $R_1, R_2, \ldots, R_n$. A relation between objects may, for example, be one of dominance between objects (e.g., John is larger than Jane), between objects and stimuli (e.g., John 'dominated' an IQ-test item by solving it), or between stimuli (e.g. item 1 is more difficult than item 2). It may also be one of proximity or similarity (e.g., John's political orientation is more similar to Jane's than to Peter's), which may again be considered in terms of similarity between objects, between stimuli, or between objects and stimuli (Coombs, 1964). Still other relations may be based on preference orderings, as is common in subjective expected utility theory. Whatever the precise nature of the relations is taken to be, they are always taken to be purely qualitative (representationalism takes 'larger than' to be a qualitative comparison). Often, there is some operation that can be interpreted as 'combining' two objects to create a new one. This combining operation is denoted $\oplus$. Sometimes this operation is empirical, such as laying two rods end-to-end to create a new one, and in this case we speak of extensive measurement. Such an empirical operation of combining is known as a concatenation operation. Campbell (1920) believed that fundamental measurement must be extensive, that is, there must exist an empirical concatenation operation, and treated all other measurement as 'derived' from these fundamental measures. However, it was later shown that there are cases where representational measurement works without there being an empirical concatenation operation (Luce & Tukey, 1964; Krantz, Luce, Suppes, & Tversky, 1971).

Taken together, the set of objects, the relation between them, and the combining operation form what is called an empirical relational system which we will call $\mathcal{O}$, which may be read as a shorthand for 'observed'. This system is denoted as $\mathcal{O} = \langle A, R, \oplus \rangle$. Now it is the business of representationalism to construct, entirely on the basis of the observed relations between objects in the set and the combinations of these objects, a numerical representation which preserves the information in the empirical system. This basically comes down to assigning to each object in A a number from some numerical domain N, to find a mathematical relation S that represents the empirical relation R, and to find a mathematical operation $\star$ that matches the combining operation $\oplus$. The resulting representational system, call it $\mathcal{R}$, a shorthand for 'representation', is then denoted $\mathcal{R} = \langle N, S, \star \rangle$. Because the representation preserves all the information that was present in the empirical system, the relation between these systems is one of homomorphism (it is not isomorphic because more than one of the elements in the empirical system may map to the same number in the representational system). The combination of $\mathcal{O}$ and $\mathcal{R}$ is called a measurement structure. Measurement, in the representationalist view, is thus essentially a homomorphic representation of objects in a numerical system.

Representational measurement is called axiomatic, because its main strategy is 1) to assume certain axioms to hold with respect to the objects and the relations among them, 2) to prove mathematically that, given these relations, a homomorphic representation is possible (this is done in a *representation* theorem), and 3) to show under which transformations of the scale values this homomorphism is preserved (this is done in a *uniqueness* theorem). The latter proof essentially characterizes the transformations under which the representation stays invariant. It can be interpreted in terms of automorphisms (Narens & Luce, 1986): This means that the uniqueness theorem states the class of transformations which may be used to map the representation into itself, in such a way that no information is lost. Uniqueness results form the basis for the well-known 'levels of measurement' introduced by Stevens (1946). If the structure of the representation is invariant under all one-one transformations, we have a nominal scale; if it is invariant under all monotonic transformations, we have an ordinal scale; if it is invariant under all linear transformations, we have an interval scale; and if it is invariant under all affine transformations, we have a ratio scale. These four scale types do not exhaust the possible scale types (Krantz, Luce, Suppes, & Tversky, 1971), but will do for the present exposition.

Semantics The semantics of representationalism vary somewhat depending on whether one considers extensive measurement, for which a concatenation operation exists, or other forms of measurement. In the extensive case, the semantics can be based on a rather concrete connection of the measurement process and the manipulation of the assigned numbers through the concatenation operation, which is itself mapped into a numerical operation. In cases of measurement that are not characterized by concatenation, the semantics of the theory are limited to representation itself. Here, the discussion will be limited to extensive measurement and one particularly important nonextensive case, namely additive conjoint measurement.

Extensive measurement The semantics of representationalism, and especially of extensive fundamental measurement as envisioned by Campbell (1920), are exquisite. The typical example for which the construction of representational scales is illustrative is the measurement of length. In this case, one may consider a set of objects, say, people, to form the set A. Further, a qualitative relation can be constructed as 'not noticeably longer than', denoted by $\preceq$, where 'Jane $\preceq$ John' means 'Jane is not noticeably longer than John'. Finally, a concatenation operation $\oplus$ is available, namely we can lay Jane and John head-to-feet and compare the resulting combined entity, 'Jane$\oplus$John' to other people, or concatenations of other people, in the set. This gives the empirical relational system $\mathcal{O} = \langle A, \preceq, \oplus \rangle$. Now, we can map the relations in the empirical relational system into a numerical system in such a manner that all relations, holding between the objects in the empirical set, continue to hold between the numbers representing these objects. So, if Jane is is not noticeably longer than John, then the number representing Jane must be smaller than or equal to the number representing John. We can, as is usual among representational measurement theorists as well as carpenters, construct the representation in the set of positive real numbers, Re^{+}, so that each person is represented by a number in this set. A common way to do this is by comparing an object to a unit of measurement, such as a centimeter, by counting the number of units that must be concatenated in order to match the object. This is done through the construction of a so-called standard sequence of equal intervals (Krantz, Luce, Suppes, & Tversky, 1971). A ruler with centimeter marks is an instantiation of a standard sequence. Further we choose the empirical relation $\preceq$ to be represented by the numerical relation $\leq$, and the concatenation operation $\oplus$ by the numerical operation $+$. Suppose that John is assigned the value $\phi(\text{John}) = 1.85$ in the meter scale, and Jane the value $\phi(\text{Jane}) = 1.75$, so that $\phi(\text{Jane}) \leq \phi(\text{John})$. Now a comparison between John and Jane, with the unaided eye, will reveal that Jane is not noticeably longer than John, i.e., Jane$\preceq$John. So, it is indeed the case that $\leq$ does a good job of representing $\preceq$. The representation will hold for all people $a, b, \ldots$ in the set A, and the technical way of expressing this is to say that $a \preceq b$ if and only if $\phi(a) \leq \phi(b)$. Also, we will find that the value assigned to the combined object Jane$\oplus$John will be $\phi(\text{Jane} \oplus \text{John}) = 3.60$, which is equal to $\phi(\text{Jane}) + \phi(\text{John}) = 1.75 + 1.85 = 3.60$. The representation of $\oplus$ by $+$ is therefore also adequate. It can furthermore be shown that the representation preserves all relevant relations in the empirical system, such as transitivity (if Jane$\preceq$John, and John$\preceq$Peter, then Jane$\preceq$Peter).

Thus, the mappings of the objects in A into numbers in Re^{+}, of $\preceq$ into $\leq$, and of $\oplus$ into $+$ have succeeded. Moreover, it can be proven that the scale is invariant up the the choice for a unit of measurement (this is to say that it does not matter whether we express someone's height in centimeters or in meters, as long as we do this consistently). Thus, the scale is insensitive to transformations of the form $\phi'(a) = c\phi(a)$, where $\phi(a)$ is the original scale value, c represents a change in unit of measurement, and $\phi'(a)$ is the resulting transformed value. This means that, if John and Jane are measured in centimeters rather than meters (so that $c = 100$), all relations will continue to hold. For example, $\phi'(\text{Jane}) + \phi'(\text{John}) = 175 + 185 = 360$ will continue to match $\phi'(\text{Jane} \oplus \text{John}) = 360$. However, if we use a centimeter instead of a meter *and* give each measured object a bonus length of 100 centimeters (so that we are in fact performing a linear transformation of the form $\phi''(a) = 100 \times \phi(a) + 100$), the mapping is destroyed. For now we would get, for Jane separately, $\phi''(\text{Jane}) = 100 \times 1.75 + 100 = 275$, and, for John separately, $\phi''(\text{John}) = 100 \times 1.85 + 100 = 285$. So, the sum of their scale values equals 560. But the concatenated object Jane $\oplus$ John, when measured with this bonus centimeter, would receive a scale value of $\phi''(\text{Jane} \oplus \text{John}) = 100 \times \phi(\text{Jane} \oplus \text{John}) + 100 = 360 + 100 = 460$. Thus, the mathematical operation $+$ ceases to be an adequate representation of the empirical operation $\oplus$. The scale values may be multiplied, but not translated, because this destroys the homomorphism between the empirical and numerical systems. This is one way of saying that the measurement of length is on a ratio scale.

Campbell (1920) held that measures that are extensive are the only genuine cases of fundamental measurement. However, Michell (2000; 2001) has noted the interesting fact that the German mathematician Hölder had already shown in 1901 that Campbell was incorrect; he had axiomatically proven that distance was quantitative without invoking a concatenation operation. Campbell and his contemporaries were apparently unaware of Hölder's work (Michell, 2001), and fervently defended the thesis that measurement without concatenation was not really measurement at all. This was the (incorrect) basis of the critique of the commission installed by the British Association for the Advancement of Science; for in psychology, it is generally difficult to identify an empirical concatenation operation. What this would require is something like the following. Suppose that I were to administer an intelligence test to a number of people (objects). Suppose further that John scores 100, and Jane scores 120. Now if I could concatenate (combine) the objects (Jane and John) in a suitable way, and this combination were shown to produce a score of $100 + 120 = 220$, and if this were true not only for John and Jane but for all combinations of people, then I would have shown that an empirical concatenation operation exists and matches the numerical operation of addition. In general, this will not work in psychological measurement. Whether this is important or not is questionable, given the fact that, for centuries, carpenters and tradespeople did quite well in measuring all kinds of things without being aware of the importance of a concatenation operation, and in fact measured many attributes for which no concatenation operation was known (such as temperature). Fortunately, nobody listened to the commission members, for it seems that if we had to wait for empirical concatenation operations to be identified *before* we started measuring, any attempt at constructing psychological measurement instruments would surely be nipped in the bud. Now, the fortunate development of measurement theory has been to reject the restrictive account of Campbell, and to replace it with a more liberal account. The unfortunate development has been that some theorists have elevated the resulting framework to the same normative level that was originally occupied by Campbell's theory, thus stating that there cannot be measurement if there is no homomorphic representation.

Conjoint measurement Although the viewpoints of the commission of the British Association for the Advancement of Science were unreasonable, the discussion of psychological measurement that followed the publication of the commission's report was instrumental in the development of measurement theory. In fact, the mathematical psychologists that took up the challenge ended up with a formalization of measurement that was far more powerful than Campbell's own, and has perhaps even been more important for physics than for psychology. The response of psychologists started with the explicit articulation of representationalism by Stevens (1946). Stevens' representationalism leaned heavily towards operationalism, because he defined measurement as "the assignment of numerals according to rule", where the nature of the rule involved is left unspecified, and Stevens was quite clear that this can be any rule. So, in Stevens' version, measurement occurs more or less by fiat; consequently, it is meaningless to ask whether something is 'really' being measured, because the fact that numerals are assigned according to rule is the sole defining feature of measurement. There is neither a need nor a place for postulating attributes which are prior to the measurement operation, as is explicitly done in latent variable theory. Representationalism, as it developed in the work of Krantz, Luce, Suppes, & Tversky (1971), followed Stevens in dropping the concatenation operation, and also retained the idea that measurement theory is a theory about numerical assignments. However, not any rule of assignment will do, because the assignment rule used must preserve the empirical relations as laid down in the empirical relational system. In essence, this boils down to the fact that representationalism takes off when the empirical relational system is already known, and then views it as its task not to explain how this relational system came into being (which many would consider to be the goal of latent variable models), but to formulate the rules for numerical assignment that preserve the relations in the system.

The broadening of the semantics associated with representationalism, which was a direct result of dropping the demand for empirical concatenation operations, provided an opening for constructing psychological measurement systems. For in this more liberal approach, measurement is no longer seen as necessarily representing empirical operations; any representation that mirrors empirical relations will do, if it complies with the demand that it forms a homomorphic representation. This follows directly from Stevens' move, which for a large part consisted in drawing attention away from the manner in which measurements are obtained (i.e., through concatenation), and towards their relations-preserving character. It also avoids the pitfall of degrading into operationalism, however, because it is possible that the relational system originates from distinct modes of assignment for different parts of the system. This is important, for while it may be possible to concatenate rigid rods of manageable length, it is arguably difficult to concatenate objects to match interstellar distances, or to place Jupiter on a balance scale. Still, my Encyclopedia mentions the fact that the average distance between the earth and the sun is about 149597890 kilometers, and that the mass of Jupiter is approximately 1.967×10^{27}

kilograms; and I strongly suspect that the writers of my Encyclopedia mean these statements to refer to qualitatively the same dimensions as, say, the distance between my cup of coffee and my telephone, and the mass of the computer I am now working on. In the rigid version of measurement theory, which leads directly to Bridgman's (1927) operationalism, these interpretations are not warranted; but in the more liberal representationalist interpretation, they are. Moreover, any imaginable structure that allows for a homomorphic representation can be subsumed under the general category of measurement. This includes structures observed in psychological measurement.

The class of structures most important to the present discussion is the class of additive conjoint structures (Luce & Tukey, 1964; , Luce, Suppes, & Tversky, 1971; Narens & Luce, 1986). Additive conjoint structures pertain to relations between at least three variables. Two of these variables are considered 'independent' variables and one is 'dependent'. The meaning of these terms is similar to that used in analysis of variance. What conjoint measurement does is a little strange from the psychometrician's point of view, because the measurement relation is not defined on any of the three variables, but on all three simultaneously. Call the independent variables A and B, and the dependent variable Y; their levels are denoted a, b, and y, respectively. What is represented in conjoint measurement is the Cartesian product $A \times B$, which consists of all ordered pairs (a, b), and the relation that is mapped in $\geq$ is the effect of these combinations on the dependent variable Y. Denote the levels of the independent variable A by i, j, k and the levels of the independent variable B by l, m, n. The idea is that, if the joint effect of (a_i, b_l) exceeds that of (a_j, b_m), so that $(a_i, b_l) \succeq (a_j, b_m)$, where $\succeq$ again is a qualitative relation and not a quantitative one, then the combination (a_i, b_l) must be assigned a higher number than the combination (a_j, b_m). The process of quantification (i.e., representing qualitative relations in the real numbers) now applies to all three variables simultaneously, but it does not require an empirical concatenation operation. What happens is that the variables A, B, and Y are scaled at once through a quantitative representation of the trade-off between A and B in producing Y (Narens & Luce, 1986). The representation theorem for conjoint measurement axiomatically states the conditions under which this can be done. It turns out that the possibility of constructing a homomorphism hinges on the possibility to find a representation that is additive in the effects of the independent variables on the dependent variable. If this is the case, then mappings f and g of the independent variables A and B into the real numbers can be found so that $(a_i, b_l) \succeq (a_j, b_m)$ if and only if $f(a_i) + g(b_l) \geq f(a_j) + g(b_m)$. The representational structure for the Cartesian product terms (a, b) is for any combination of levels i of A and l of B then given by $\phi(a_i, b_l) = f(a_i) + g(b_l)$. The representation is on an interval scale, because the structure is invariant under linear transformations of the assigned scale values.

Conjoint measurement thus constructs a mapping of an empirical relational system $\mathcal{O} = \langle A \times B, \succeq \rangle$ into $\mathcal{R} = \langle \text{Re}, \geq \rangle$. No empirical concatenation operation is in sight, as is reflected by the omission of $\oplus$ in the notation, although adding scale values is meaningful. It is possible to imagine a kind of concatenation operation that relates the system to extensive measurement structures, but this concatenation is not empirical. In the literature, this is called an 'induced' concatenation operation

(Narens & Luce, 1986). Because this is not an empirical concatenation operation, however, the operation cannot be said to be represented in the way this can be said of extensive measurement.

It is, however, important to consider why conjoint measurement gives an interval scale, i.e., what the meaning of the measurement unit is. This comes down to the question what it is, exactly, that is being measured. Basically, what is represented in the model is a trade-off. The meaning of the measurement unit is in terms of this trade-off. For instance, suppose that we have a given combination (a_i, b_l), and increase the level of A from a_i to a_j, thereby constructing a new combination (a_j, b_l) that is $\succ$ to the original one. The conjoint model then says by how many units the factor B has to be decreased in order to produce a new combination (a_j, b_k) that is not noticeably different from (i.e., that is both $\succeq$ and $\preceq$ to) the original combination (a_i, b_l). Thus, the model states how effects resulting from variations in A can be undone by variations in B, and vice versa. The measurement unit is explicitly defined in terms of this trade-off. The reason for this is that any two distances $a_i - a_j$ and $a_j - a_k$ on the factor A are defined to be equal if they can be matched by the same distance $b_l - b_k$ on the factor B. The measurement unit on the factor A is thus defined as the change in A necessary to match an arbitrary change on the factor B, and the measurement unit on the factor B is defined as the change in B necessary to match an arbitrary change in the factor A. This is the reason why it is crucial to have two factors; one cannot define a unit of measurement on one factor without reference to the other. Because the method does not match levels in A by levels in B, but rather differences between levels of A by differences in levels of B, it can be expected to yield an interval scale. This is formally the case because linear transformations on the scale values assigned to the levels of either of the factors preserve the representation.

4.2.2 The empirical stance

Representational measurement is, as has been stated before, concerned with formulating the conditions that must be fulfilled in order to be able to construct a representation. These conditions, which are formulated as axioms, thus describe the relations that must hold in the data at hand for a representation to be possible. They are of an empirical nature; in Krantz, Luce, Suppes, & Tversky (1971) they are even called empirical laws. For extensive measurement, the axioms involved are rather simple (see Narens & Luce, 1986, for a lucid description). For conjoint measurement, they are more complicated. Basically, if one knew a priori that the effects of the independent variables were additive, there would be no need for the specification of the axioms involved, and an additive representation could be readily constructed. The strategy of representationalism, however, is not to posit variables and relations between them in reality and to look at whether the data structure is more or less consistent with these (i.e., the model fitting approach as used in latent variable modeling). It always starts with the data, never with the metaphysics. So, the axioms of conjoint measurement describe characteristics that the data must exhibit for us to be able to *construct* an additive representation.

As always, what we start with is a set of purely qualitative relations. In this case, however, the elements on which these relations are defined are the combinations (a, b). These combinations are assumed to be ordered. This ordering is in a sense 'induced' by Y. For example, suppose that a subject must judge, for a tone generated by a given combination (a, b) of intensity (A) and frequency (B) whether its loudness (Y) noticeably exceeds $(\succeq)$ that of a tone generated by a different combination. The first axiom of conjoint measurement states that the ordering so produced must be a weak order. A weak order is an order that is transitive and connected. Transitivity means that for each combination of levels i, j, k of A and l, m, n of B, if $(a_i, b_l) \succeq (a_j, b_m)$ and $(a_j, b_m) \succeq (a_k, b_n)$, then $(a_i, b_l) \succeq (a_k, b_n)$. Connectedness means that each comparison is made, and for all comparisons either $(a_i, b_l) \succeq (a_j, b_m)$, or $(a_j, b_m) \succeq (a_i, b_l)$, or both.

The second axiom of conjoint measurement is called independence. It states that the ordering of the levels in A, which is induced by the ordering in Y, must be unaffected by which particular value of B is chosen to assess this ordering; the converse must also hold. So, if we assess the ordering of perceived loudness as produced by varying levels of intensity, we have to do this while holding the frequency of the presented tones constant. The independence condition says that it must not make a difference for the ordering whether we set the frequency at 100Hz or at 1000Hz. Higher intensities must in either case produce either an unnoticeable difference or a higher perceived loudness. This means that, if there is an interaction effect of the independent variables, no additive conjoint measurement representation can be formed. However, the restriction this poses is less serious than it may seem. This is because the original observations on the Y variable are assumed to be merely ordinal. Thus, any monotonic, order-preserving transformation on these observations is permissible. The restriction posed is therefore relatively mild: There must exist a monotonic transformation of the dependent variable that renders the effects of the independent variables additive. It is possible to remove a wide class of interaction effects by transforming the dependent variable. A real problem occurs, however, in the presence of disordinal interactions, i.e., when effects 'cross'. This would be the case, for example, if for tones with a frequency below 1000Hz a higher amplitude would produce a higher perceived loudness, but for tones with a frequency above 1000Hz, a higher amplitude would produce a lower perceived loudness. If this happens, the very ordering on A, as induced by the ordering on Y, depends on the selected level of B, and no additive representation will be possible.

The independence condition allows for the independent ordering of the factors A and B in terms of increasing values of Y. On the basis of this ordering, we can represent the structure in a table like Table 1, which contains three levels for each factor. Factor A is represented as increasing in Y from left to right; factor B is represented as increasing from top to bottom. The entries in the table are the (monotonically transformed) values y as corresponding to each combination (a, b). Because of the independence condition, the entries are increasing both in the rows and in the columns of the table.

Table 4.1. The combinations (a, b) are ascending both in rows (left to right) and columns (top to bottom).

		Factor A		
		1	2	3
Factor B	1	(a_1, b_1)	(a_2, b_1)	(a_3, b_1)
	2	(a_1, b_2)	(a_2, b_2)	(a_3, b_2)
	3	(a_1, b_3)	(a_2, b_3)	(a_3, b_3)

The third axiom of conjoint measurement is called double cancellation and refers to relations between the diagonals of the table. It is basically a consequence of additivity, and invoked because the axiom of monotonicity does not, by itself, guarantee additivity in the two-factor case (Krantz, Luce, Suppes, & Tversky, 1971, p. 250). Additivity requires that any entry (a, b) can be represented by the additive function $f(a) + g(b)$. Therefore, an entry, say, (a_2, b_1), must be $\succeq$ (yield a greater amount of Y) to another entry, say, (a_1, b_2), if and only if $f(a_2) + g(b_1) \geq f(a_1) + g(b_2)$. Suppose that this is the case, and that it is also the case that $(a_3, b_2) \succeq (a_2, b_3)$. Then we have the two inequalities

$$f(a_2) + g(b_1) \geq f(a_1) + g(b_2) \tag{4.1}$$

and

$$f(a_3) + g(b_2) \geq f(a_2) + g(b_3). \tag{4.2}$$

If the effects of the factors are additive, it follows that

$$f(a_2) + g(b_1) + f(a_3) + g(b_2) \geq f(a_1) + g(b_2) + f(a_2) + g(b_3), \tag{4.3}$$

which implies the new inequality

$$f(a_3) + g(b_1) \geq f(a_1) + g(b_3). \tag{4.4}$$

This is the condition of double cancellation ('cancellation', because of the terms cancelling out in the last step of the derivation, and 'double' because there are two antecedent inequalities). The double cancellation axiom must hold for all 3 × 3 submatrices of the larger matrix defined over all levels of A and B.

The final axiom needed is called the Archimedean axiom. This axiom is also commonly used in extensive measurement, where it asserts, for instance, that no object is infinitely larger than any other object. In the present context, the axiom states that no difference in A produces an infinitely larger change in Y than any other difference in A, and that no difference in B produces an infinitely larger change than any other difference in B. This axiom is technical in nature, and I will neglect in the following.

If the data satisfy the above axioms, then an additive representation can be constructed that preserves all of the relevant relations in the data. Conjoint measurement theory shows that fundamental measurement does not require a concatenation operation, and in doing so provides a justification for intensive measurement that is lacking in Campbell's account. It also provides psychology with a system for measurement that is on equal footing with the ones in physics. For the way subjective loudness could be measured and quantified is exactly the same way in which density can be measured and quantified. The representationalists thus showed that the conclusion reached by the commission installed by the British Association for the Advancement of Science was false: Fundamental measurement is, in principle, possible in psychology.

4.2.3 The ontological stance

Representationalism is the only theory of measurement with an explicit ontological status for its central concept. Scales are representations of observed relations and therefore they are constructions. Scales do not 'underlie' the observed relations, and much less are they causally active in producing them. The recipe for scale construction is crisp and clean, and devoid of any metaphysical assumptions whatsoever. Parsimonious and powerful, representationalism is the dream of every empiricist philosopher and scientist alike. We can simply start by observing relations between objects with the unaided eye (Van Fraassen, 1980), and show how theoretical terms like 'length', 'distance', or 'subjective loudness' can be constructed and quantified based on these relations. There is little one can say about this except that it is probably the most comprehensive and adequate empiricist theory of measurement that could possibly be given. Upon closer consideration, however, it appears to be rather unclear what exactly constitutes the relation between the logical structure of representationalism and the actual measurement process. One option is to interpret the axiomatic theory as providing us with a definition of measurement. This would suggest that the theory provides necessary and sufficient, or at the very least necessary, conditions for a procedure to satisfy in order to be covered by the definition of measurement. This interpretation will be argued to be problematic; if the theory gave sufficient conditions, then it would include absurd cases, and if it gave necessary conditions, many recognized instances of measurement would not be covered by the definition because representationalism cannot deal with error. Alternatively, one could interpret the theory as a prescriptive theory that shows how we should go about constructing scales in psychology. This, however, does not work either. Contrary to suggestive wordings like 'not noticeably longer than', representationalism does not describe how fallible human beings such as ourselves could construct scales. It describes, at best, how a Laplacean demon – a rational being with an infinite amount of time, an infinitely large brain, and a capacity for errorless observation – could construct scales on the basis of observable qualitative relations. Because we are not such beings, the representationalist enterprise cannot be seen as a serious proposal for constructing measurements. Therefore, the prescriptive reading of the theory is not justified.

Representation and measurement

The problem with the view that representationalism gives a definition of measurement concerns its central tenet, namely that measurement is essentially about representation. While there is a nontrivial sense in which this is true, namely, we do aim to construct a numerical system that reflects certain systematic relations in the world, there is also a nontrivial sense in which it is false. The sense in which it is false is that measurement is not exclusively representational. In particular, the fact that a representation can be constructed cannot be a sufficient condition in any sensible definition of measurement.

This is evident from the fact that we can construct situations where we have homomorphic representations which are not measurements in any meaningful sense of the word. Consider, for example, the Guttman model, which is a deterministic item response model to which the axiomatic theory applies. The Guttman model is generally seen as a model for ordinal measurement, but its mathematical requirements by themselves do not warrant this interpretation. To see this, consider the following four items:

1. I have parents (yes: 1, no: 0)
2. I have no beard (yes: 1, no: 0)
3. I menstruate (yes: 1, no: 0)
4. I have given birth to children (yes: 1, no: 0)

Suppose that we administered these items to a group of people. Obviously, we would get a triangulated structure that looks as follows:

Item 1	Item 2	Item 3	Item 4	Sumscore
1	0	0	0	1
1	1	0	0	2
1	1	1	0	3
1	1	1	1	4

This triangulated structure is a necessary and sufficient condition for constructing a Guttman scale. The reason that we get this structure, of course, is simply that we have constructed inclusive subclasses of people. People with sumscore 1 are men with a beard; people with sumscore 2 are non-menstruating women and men without a beard, people with sumscore 3 are women without children, and people with sumscore 4 are women with children. Now, if measurement were nothing more than homomorphic representation of empirically observed relations, and the Guttman model produces an ordinal scale, then we would be forced to conclude that we have ordinally measured something here. This does not seem to be the case. However, the example surely provides a case of homomorphic representation. Therefore, representation and measurement are not the same. That a representation can be constructed is not a sufficient condition for obtaining measurements.

This is not surprising because a representation is a purely formal concept, while the question whether measurement has taken place is not a purely formal one, as is evident from the literature on validity (Cronbach & Meehl, 1955; Messick, 1989). Thus, to speak of measurement requires extending the formal framework with a substantive interpretation. And this interpretation cannot, in principle, be given by the formal model itself. It seems to me perfectly in order to say that a representation may be constructed in cases where nothing is measured. Now this is not an argument against the representational view in general; it merely says that there is more to measurement that representation alone, and that the ability to construct a representation cannot be a sufficient condition for measurement.

The problem of error

While it is quite clear that representationalism cannot give sufficient conditions for measurement, we could at least imagine that the theory gives necessary conditions for measurement, as is suggested, for example, by Michell (1990; 1999). This, however, is also difficult because the theory has a hard time dealing with the problem of error. If the possibility to construct a homomorphic representation were to be a necessary condition for measurement, this entails that we should be able to gather data that fit the measurement model perfectly. This is because, strictly speaking, models like the conjoint model are refuted by a single violation of the axioms. For example, if there is a single triple of observations where transitivity is violated, or a single 3×3 submatrix that violates double cancellation, the model is falsified, because no homomorphic representation will be possible. Since we can safely assume that we will not succeed in getting error-free data – certainly not in psychology – we must choose between two conclusions: Either measurement is impossible, or it is not necessary to construct a perfect homomorphic representation. If we accept the former, we may just as well stop the discussion right now. If we accept the latter, then we have to invent a way to deal with error.

The return of Mr. Brown The natural means of introducing a theory of error would be to construct a statistical formulation of representational measurement theory. In such a theory, one would have to introduce parameters to represent the true values of the objects. One way to do this would be by replacing sentences like '$a \preceq b$ if and only if $\phi(a) \leq \phi(b)$' with sentences of the form '$\tau_a \preceq \tau_b$ if and only if $\phi(a) \leq \phi(b)$'. Here, the τ variable could serve the function of denoting the true value of the objects on some instrument used to make the comparison between a and b. This instrument could be a particular meter stick, but it could also be an item in a psychological test. Scheiblechner (1999) who follows this line of reasoning, calls the indirect comparison of objects, through their true scores on an instrument, an instrumental comparison (p.299). The so constructed model allows for error because it may be the case that a particular observer judges that $a \preceq b$ while it is actually the case that $\tau_a \succ \tau_b$.

The problem, of course, is that the very introduction of error requires an account of what the true values are. The common approach to this problem in statistics is by introducing the idea that the observed values are realizations of a random variable. Conceiving of the measurement apparatus as yielding a value x for each object, we could implement this idea by interpreting x as a realization of the random variable X. We may then introduce the assumption that $\mathcal{E}(X_a) = \tau_a$, analogous to the way this is done in classical test theory. The interpretation of the so constructed sentence in terms of length would be 'the expected centimeter reading of a is not noticeably larger than the expected centimeter reading of b if and only if the number assigned to a is smaller than the number assigned to b'. Because nobody can observe the expected values, we should delete the word 'noticeably'. This implies that we should also replace the symbol $\preceq$, which stands for 'not noticeably longer than' by the symbol $\leq$ which means 'has a lower expected centimeter reading than'[1]. That is, the instrumental comparison can only be made by examining relations between expected values, which are by necessity numerical. So, an interesting shift takes place here: While the fundamental measurement model aims to construct quantitative metrics from qualitative observations, the instrumental comparison introduces a kind of quantitative metric directly at the level of the comparisons made.

Expected values are not observable, and the fact that we are introducing relations between unobservables at such a fundamental level in the construction of the model has far-reaching consequences. In effect, we are now already working with a true score model. And if we aim to construct a measurement instrument that measures a single attribute with a number of observed variables, we will build a structure that strongly resembles a latent variable model. Considered in terms of a psychological test consisting of a number of items, this would work as follows. Interpreting the numerical assignment ϕ as a latent variable (now interpreted as a rescaling of the true score), which represents an item $\times$ subject Cartesian product with an ordering induced (in both items and subjects) by the τ variable, we can construct an additive conjoint representation if the item and subject effects are independent, additive, and satisfy the double cancellation axiom with respect to the values of τ (Scheiblechner, 1999). An example of a model that has these properties is the Rasch model (Rasch, 1960). Thus, this statistical decomposition of observed values, in true and error components, leads directly to the class of additive Item Response Theory models. I will have more to say about this relation in Chapter 5.

This approach to the problem of error is useful because it shows that the divide between representationalism and latent variable theory is formally speaking a fine line. From a philosophical viewpoint, however, crossing this line has serious consequences; in effect, the main tenets of representationalism are lost in the present approach. The first problem is that we have assumed the existence of an instrument that gives the measurements to apply the expectation operator to. The present approach merely allows for the construction of a ruler with equal intervals on the basis of comparisons made by using a ruler with unequal intervals. It can be used to show how a scale can be linearized, analogous to the way that Rasch models linearize sumscores by appropriately stretching the far ends of the scale. However, representationalism is not served by assuming, a priori, that a ruler exists. For the theory is aimed at showing how a ruler with equal intervals could be constructed on the basis of direct qualitative comparisons with respect to the variable in question

[1] While probably not intended in this manner, the fact that Scheiblechner (1999) retains the $\preceq$ relation in the introduction of his ADISOP models is slightly misleading. The relation is interpreted as a 'stochastic dominance' relation (p. 299), where $a \preceq b$ means, for example, that subject a has a lower expected value on a given item than subject b. The notation is adequate in that the denoted relation is only taken to establish an ordering, and in this sense is qualitative, but it does certainly not have the connotation of noticeability or observability, a connotation that such relations generally do have in the representational approach.

– whether it is length, density, or subjective loudness – and not at showing how length can be measured given the fact that a ruler already exists. More importantly, however, the very construction of a ruler is frustrated in the present approach. The reason for this is that the construction process would have to be carried out though the evaluation of stochastic dominance relations of the above type. These relations are clearly unobservable. Moreover, expectations cannot be empirically concatenated in principle. As a result, even the possibility of extensive measurement now vanishes. The third and most troubling consequence of this move is that in most cases of measurement, but certainly in psychology, we will encounter serious problems in the interpretation of the expected values involved. In fact, we are likely to be forced to interpret the expected values in a propensity sense. So we can now hear Mr. Brown knocking on the back door; and the representationalist certainly would not want to let him in. It thus seems that, in this approach, we are quickly losing the gist of the representationalist theory. For we are not building homomorphic representations of observable objects and qualitative observable relations between them; we are building isomorphic representations of unobservable true scores and equally unobservable relations between them.

Introducing the Laplacean demon A second way to introduce a concept of error would be to introduce true relations between objects, rather than to assume true scores for the objects. This could be done by replacing sentences like '$a \preceq b$ if and only if $\phi(a) \leq \phi(b)$' with sentences of the form '$a \preceq_{true} b$ if and only if $\phi(a) \leq \phi(b)$'. That this will not work is obvious from the fact that the values ϕ are, in representationalism, constructed from the data and not hypothesized a priori. Because we cannot observe the true relations, we cannot construct these values and the above formulation is nonsensical. It would be an idea, however, to take the idealization one step further and to introduce true values for the ϕ involved. These values are not to be interpreted as existing independently of the relations they figure in, as in the introduction of expected values above. Rather, they should be seen as the values that would be constructed if we could observe the true relations between objects. Their status as true values is thus derived from positing true relations, rather than the other way around. Also, the relation $\preceq$ does not have to be interpreted as a relation between propensities. It can be taken to be a completely deterministic relation between objects. So now we could get '$a \preceq_{true} b$ if and only if $\phi_{true}(a) \leq \phi_{true}(b)$'. Interpreted in terms of length, this sentence says that a is truly not noticeably longer than b if and only if the number, that would be assigned to a if we could observe the true relations, is smaller than or equal to the number, that would be assigned to b, if we could observe the true relations. We thus retain the construction of quantitative scales out of qualitative relations, and refrain from introducing relations between unobservables in the definitions. The only problem is that the relation $\preceq$ has no natural interpretation anymore. For what does 'truly not noticeably longer than' mean? Does it mean that nobody could, in principle, notice that a is longer than b if a is actually longer than b? No, because if this were the case, we could just use fundamental measurement theory as it stands; for there would be no error, and consequently there would be no need for the present exercise. Does it then mean that no perfect observer could notice that a is longer than b if a is truly longer than b? Possibly, but who is the perfect observer? A Laplacean demon?

The problem we are facing here is clearly caused by the word 'noticeably'. The use of this term suggests that somebody is actually noticing relations between objects, and the judgments of this anonymous somebody would produce transitive and ordered data when measuring attributes that sustain measurement. Upon closer inspection, the identity of this anonymous observer is mysterious. The interpretation of the word 'noticeable' is unproblematic for an empiricist reading of the theory as long as we interpret it as 'noticeable with the unaided eye', that is, noticeable in practice. Because in this interpretation the theory is unable to deal with error, we have to move beyond the practically feasible observational powers of human beings and construct the relations as noticeable for somebody with observational powers that markedly exceed our own. This is necessary because the introduction of error means that we need to be able to say that we are wrong, and being wrong is always relative to being right. That is, error is a deviation, and the natural interpretation of the concept is that it is a deviation from a standard. In a theory that works its way up from qualitative, noticeable relations, we need somebody to notice the correct relations, which could function as such a standard. And if it cannot be us, then it must be a demon with infallible observational powers. Hence the need to introduce a Laplacean demon.

Now, if we want to pursue this line of reasoning without introducing propensities, expected values, and latent variables into reality, it is obvious that we must limit the relation $\preceq$ to be a relation between objects, and not between true scores. If we do not do this, then we must again introduce expected values and relations between them for the demon to notice. This requires that such values and relations exist in reality, so that we would again be introducing the metaphysics we sought to evade; in effect, we would arrive at the same conception of measurement as in the previous attempt to deal with error. Dismissing relations between propensities, however, has a very important consequence: It excludes any model that posits relations between expected values. Thus, in this interpretation, additive models like the Rasch model (1960) and the ADISOP models (Scheiblechner, 1999) are not representational models because they posit relations between propensities.

Perhaps the representationalist would not object to the exclusion of additive IRT models. One rarely encounters a reference to these models in the representationalist literature, and I would indeed suspect that representationalists reject such models because of the fact that they introduce too much metaphysics. The advocates of additive IRT models tend to flirt with fundamental measurement theory (e.g., Wright, 1997; Bond & Fox, 2001), but the reverse is definitely not the case. However, even the pet examples of representationalism would have difficulty surviving the demands posited in the approach we are presently considering. Consider the measurement of subjective loudness. What would we have to posit in order to be able to say that, while subject i did not notice the combination (a_i, b_l) to be $\preceq$ to the combination (a_j, b_k), he erred in this response? Or to say that, while i said he preferred stimulus j to stimulus k, he was misjudging his own preferences? The problem here is, of course, that the word 'noticeable' is, in these cases, intended as 'noticeable for subject i' and not as 'noticeable for a Laplacean demon'. The very subjective nature of the comparisons on which fundamental measurement operates in these cases precludes the introduction of error. For this requires us to say that$i$ is objectively wrong concerning his subjective state of mind. This does not seem to go in the right direction, whether we consider the situation from the representational point of view or otherwise. Thus, in this approach few of the accomplishments of representational theory are preserved: The additive IRT models are excluded from consideration, and subjective scales for loudness, preference, etc., are deprived of their intended interpretation. In fact, the only examples of measurement that would sit well with this approach are examples from physics. The measurement of length, mass, and density do sustain the idea that they represent deterministic relations between objects as they could be observed by a Laplacean demon. But the measurement of psychological variables is not satisfactorily incorporated in this approach.

Reconceptualizing error A final possibility to deal with imperfect observations is not to view them as error at all. Whatever the ontological status of error may be, in the final analysis the only epistemological criterion to detect error is as a deviation from a theoretical model. Instead of laying the blame on the observations, so to speak, one may attribute the deviations to a failure of the model. In such a view, the model is not interpreted as aspiring truth, but as an approximation. One may then choose to minimize the distance between, for instance, the conjoint representation and the data matrix. This can be done by constructing a stress measure for this distance, and then minimizing the stress of the model with respect to the data. Interpreted in this manner, representational measurement theory would be a (possibly multidimensional) scaling technique, because error is not conceptualized as inherent to the variables observed, but as the distance between the data matrix and the representation. (In multidimensional scaling it is nonsensical to ask what the 'true' representation is, in contrast to latent variable models, where the quest for the true model is often deemed very important.) Representationalism does have a structure that is similar to scaling techniques (Coombs, 1964), so that this approach would seem a natural way for representationalism to deal with error. However, in this approach the main idea of representational measurement theory is also lost, because whatever the relation between the data and the representation may be, it will not be a homomorphic mapping.

So, it seems that representational theory is stuck between a rock and a hard place: It must either say that no psychological set of data satisfies the axioms, thereby forcing the conclusion that psychological measurement is impossible after all, or it must introduce a concept of error. The three ways of doing this, as discussed above, are not satisfactory. In the first attempt, we were forced to introduce expected values for the objects. This not only requires the existence of an instrument yielding values to apply the expectation operator to, but must also posit probabilities that can only be interpreted as propensities. In effect, the structure we end up with strongly resembles a latent variable model, and the homomorphism constructed involves unobservable relations between unobservable true scores. This can hardly be considered to maintain the spirit of representationalism. The second attempt introduced true qualitative relations between the objects, and derived true values only in virtue of these relations. However, in this interpretation we must hypothesize a supernatural being to observe the true relations. Although this conception is perhaps closest in spirit to the formulation of representational measurement, it cannot be considered a case of progress in terms of keeping a low metaphysical profile. Finally, if we choose a more pragmatic approach, and simply minimize the distance between the data and the representation, we refrain from introducing metaphysics, but at the same time lose another central idea in representationalism, which is that we are constructing a homomorphic mapping of objects into the real numbers. Thus, the inability to deal with error seems to be deeply entrenched in the structure of representationalism. Attempts to incorporate an error structure seem to invariably destroy one or another of the tenets of the theory. This does not, of course, imply that the formal structure of representational theory could not be applied to stochastic systems. It merely means that to do so requires giving up the empiricist connotation of the theory.

Representationalism as rational reconstruction

Representationalism does not state sufficient conditions for a definition of measurement, because there are representations that are not measurements. Neither does it provide necessary conditions, because the conditions as stated will be false in the presence of error, and it is hard or impossible to modify the theory to bypass this problem. We cannot hold, therefore, that representationalism defines measurement in general. We may, at best, hold that it states necessary conditions for perfectly reliable (i.e., deterministic, errorless) measurement, which would be a definition without much practical use – particularly in psychology. If it does not offer a definition of measurement, however, what is the status of representationalism? What is its relation to actual measurement?

First, we must note that representational theory certainly elucidates the structure of measurement. Especially in physics, the theory has greatly clarified the nature of various measurement techniques by concentrating on the relation between a empirical relational system and a numerical relational system. Second, it is clear that the theory seeks to offer logical underpinnings of scale construction. In the extensive case, the logical requirements are connected to the process of measurement by relating the concatenation operation to the numerical operation of addition. This is a particularly interesting case because it seems to show what it 'actually is' that carpenters and tradespeople have been doing all along. The theory gives a formal structure, and it would seem that this formal structure is in some way 'instantiated' in the behavior of cashiers and scientists alike. This suggests that the theory gives a *reconstruction* of the measurement process.

Of course, such a reconstruction should definitely not be taken to be an actual reconstruction of the historical process that led to measurement as we now know it. Such an interpretation of the approach would be vulnerable to the same objection that, in the long run, proved fatal to the idea that logical positivism described the actual structure of theory development: It simply does not work that way. For instance, in his historical overview of social science measurement, Wright (1997) quotes a passage from the Magna Carta, dating from 1215, in which King John of England declared:

> "There shall be one measure of wine throughout our kingdom, and one of ale, and one measure of corn, to wit, the London quarter, and one breadth of cloth, to wit, two ells within the selvages. As with measures so shall it be with weights."

The quotation illustrates the obvious but interesting fact that King John, who clearly understood the basic principles of measurement and their importance quite well, is nevertheless unlikely to have thought of measurement as the homomorphic mapping of an empirical relational system into a numerical one. Likewise, carpenters, tradespeople, and scientists use the principles of measurement without apparent awareness of the higher mathematics they are involved in. So, unless we assume that Freud's unconscious not only exists, but actually has a serious expertise in set theory, it would surely be outrageous to say that representational measurement theory is a theory of how measurement is carried out in practice. King John knew nothing about set theory, and still he believed that measurement was important enough to include a passage about it in the Magna Carta. The Egyptians certainly knew how to measure the bricks they used in constructing pyramids, but they did not even have the number zero, let alone the real number system. Scientists use measurement procedures all the time, but they cannot explain to you how representational measurement theory works unless they have studied it. Clearly, the reconstruction given is not a reconstruction of the historical development of measurement, and neither can it be interpreted as a psychological description of the scientist carrying out the measurement procedure.

The question that forces itself upon us then becomes: If representationalism offers a reconstruction, then what is it reconstructing? Exactly the same problem was faced by the logical positivists, who described the structure of scientific theories in a way that was quickly realized to be inadequate as an actual description. Reichenbach (1938) circumvented this problem by stating that he was giving a *rational reconstruction* of scientific theories. Such a conceptualization seems to fit representationalism quite nicely. In accordance with the findings in the previous section, we may then interpret the theory not as showing how a scale *can* be constructed, but how a scale *could* be constructed by a Laplacean demon, i.e., a being equipped with powers that enable him to make errorless observations of qualitative relations between objects. I think this is the best possible interpretation of the theory. Representationalism offers a theory of homomorphisms that certainly has intuitive appeal in its mathematical elegance and parsimony. The interpretation in terms of rational reconstruction does nothing to devaluate the theory at this level. However, the interpretation also makes clear that representationalism is not a theory about how the concept of measurement developed, and it is not a theory of what scientists do, for the simple reason that they are not Laplacean demons. While this does not diminish the importance of the analysis for understanding measurement structures, it does raise doubts with respect to the prescriptive force of representationalism.

Reconstruction does not entail prescription

One of the most serious mistakes one can make when thinking about any research topic is the following: in a certain population, we have observed that elements with property x tend to have property y, so if we equip elements lacking property x with that property, then they will also develop property y. Philosophers of science who view it as their mission to equip scientists with prescriptive criteria of theory development, model selection, or scale construction, commonly commit this fallacy. For example, they think that, because successful scientific theories have property x, we can construct successful theories by forming theories that have property x.

So, because we observe that Copernican astronomy is more parsimonious and yet equally adequate in prediction as the Ptolemean system, we should select models in psychology on the basis of parsimony and predictive adequacy – regardless of whether such a strategy has been shown to yield better theories in psychology (it has not). Because we observe that highly successful theories in physics, like Newtonian mechanics and the theory of relativity, are falsifiable, we should construct falsifiable theories in psychology, and psychology will automatically become a serious scientific discipline – regardless of whether it makes any sense to construct falsifiable theories at the present stage of theory formation in psychology. And so it is with the advocates of fundamental measurement: Measurement has been very successful in physics, where it allegedly obeys the structure of fundamental measurement theory, so if we construct psychological measures based on this theory, then psychology will finally become the long sought quantitative science we have all been dreaming of. This kind of science fiction is continually being propagated by the advocates of Rasch models (Wright, 1997; Bond & Fox, 2001), fed by the otherwise admirable theoretical work of Michell (1990; 1999, 2000), and has led Kline (1998) to adopt the mysterious view that psychology cannot be scientific without fundamental measurement.

This puts the horse behind the cart. We have seen that fundamental measurement theory cannot be interpreted as more than a rational reconstruction of the measurement process. This does not devaluate the theoretical insights it provides, but it is important to keep in mind that reconstruction does not entail prescription. One should remember that fundamental measurement theory is entirely post-hoc. It has not helped to construct measurements of length or mass; it has not even helped to construct measurements of conjoint concepts like force or density. It has elucidated the structure of such cases of measurement. But it is highly questionable, given its deterministic nature, whether the theory could have been of any use in constructing these measures in the first place. Probably, the clean logic of the theory would have been an obstacle rather than an aid to the development of measurement procedures – a process that is messy, inexact, full of spurious results, packed with error variance, and that more often than not requires decisions to be made on the basis of intuition than on the basis of logic. It is a fallacy to think that, because established forms of measurement allow for a philosophical reconstruction in terms of model x, all measurements should be constructed to obey the prescriptions of model x - regardless of whether model x is a fundamental measurement model, a latent variable model, a generalizability theory model, or some other technique.

The reason why fundamental measurement theory cannot be a prescriptive methodological framework has to do with its particular deterministic structure, but much more with the fact that it is just a model. One should be weary of models that are propagated as prescriptive frameworks in a universal sense, because whether or not a model is adequate in a given situation strongly depends on the substantive considerations that are relevant in that particular situation. Additivity, for example, is desirable because it is simple, but it is only desirable if substantive theory suggests that additivity should hold. Now, if we had different kinds of meter sticks that produced crossing interactions, we would certainly be surprised: It would be strange if on some meter sticks longer things had a higher expected value than shorter things, and on other meter sticks longer things had a lower expected value than shorter things. Yet, this is what the presence of crossing interactions (as present in factor models with unequal factor loadings as well as in Birnbaum models) signifies. On the other hand, if we asked people a number of questions to measure their height, we might certainly encounter such situations. For instance, the item "I can touch the top of a doorpost with my hands" can reasonably be considered to measure bodily height, be it indirectly. It will show a quite steep curve as a function of height, jumping from "no" to "yes" at about 1.75 meters. Coding the item as "yes":1 and "no":0, we might imagine this item to have an expected value of .80 for people 1.80 meters tall, and an expected value of .20 for people 1.70 meters tall. The item "I am pretty tall" is less direct, but may nevertheless be considered to validly measure the trait at hand. Because it is less direct, the item characteristic curve will not jump from 0 to 1 as suddenly and steeply as the previous item. This yields the possibility that people who are 1.70 meters tall will have an expected value of .30, while people who are 1.80 meters tall may have an expected value of .70. Thus, for people who are 1.80 meters tall, the first item is 'easier' than the second, but for people who are 1.70 meters tall, the second item is easier than the first. Technically, this means that there is a crossing interaction between the subject and item factors, which implies that additivity is violated and no conjoint additive representation can be found. Does this mean we cannot use the two items to construct a valid measure for height? And what about items used to measure cognitive abilities or personality characteristics? Should we always demand additivity in such cases? One should be weary to draw this conclusion because it depends on a dogmatic view that leans towards essentialism about the term 'measurement'. In the absence of a rationale based on substantive, rather than philosophical, considerations that sustain various formal properties like additivity (or, for that matter, unidimensionality, measurement invariance, and the like) one should be very careful in propagating the universal demand for such properties. It amounts to pure speculation to say that constructing measures on the basis of these formal criteria will lead to better measurement.

It is certainly true that many measurement practices in psychology, as well as models that employ continuous latent variables, assume that psychological attributes are quantitative. This is a serious assumption, as has been clarified by Michell (1990; 1999). It is also true that finding an additive conjoint measurement representation, even if it is probabilistic like in the Rasch model or the ADISOP models discussed by Scheiblechner (1999), yields support for this assumption. But it is not true that the existence of a continuous, quantitative psychological attribute guarantees that one can construct a psychological test for it that allows for an additive representation. It has never been proven, and will never be proven, that constructing tests on the basis of this requirement will yield better tests, regardless of the dimension one is trying to measure or the substantive field one is working in.

In the substantive context of psychology, we should seriously consider the possibility that psychological measurement is so much more complicated than measurement in physics, that it is a different ball game altogether. That is, fundamental measurement may be fundamentally inappropriate. Consider the simple fact that some questions are better asked in a dichotomous format, while some perform better in a polytomous format, whereas still other questions are more appropriately put in open-ended form. Assume for a moment that the trait indicated by the term 'general intelligence' exists, and is quantitative. It may well be that the best possible test for intelligence would be composed of different item formats (in fact, such tests commonly are). Now how are we going to implement the constraints of fundamental measurement in this case? Clearly, we are not going to do this at all, at least not in the presently available forms of the model.

Now one may either give up and conclude that psychology "will never be a science" (Kline, 1998), or one may try to accommodate for the problem, for example by assuming that a latent variable exists and underlies the observations, and try to build a model that can handle the situation. A serious attempt at doing this has been undertaken by Moustaki & Knott (2000), who have formulated a general latent variable model that allows for the use of different item formats. Probably, this model would be rejected by fundamental measurement theorists: No representation or uniqueness theorems are available for it, and I would not be surprised if it were proven that no appropriate representations exist at all. Therefore, the model cannot be said to be a measurement model in the fundamental measurement theory sense. But is this the constructive contribution of representationalism to the problem of psychological measurement? That it is impossible to measure something if it does not allow for a formal reconstruction in terms of representation and uniqueness theorems? I consider that a very unsatisfying option. It is undoubtedly the case that representationalism has done much to help the case of psychological measurement, and there may indeed be applications of the theory, in psychophysics and other basic areas of psychology, where it actually works. In the case of higher order constructs like extraversion, intelligence, and attitudes, however, the theory has not been very useful so far. And considerations like the above (and there are many more) suggest that it will not do very much in these areas. The idea that the representationalist strategy is required in every case of measurement may be taken to be a truism, if one limits the meaning of measurement to homomorphic representation. I consider such a limitation to be unduly restrictive. Moreover, because representationalism is no more than a reconstruction, the general demand for fundamental measurement is based on an overinterpretation of the theory.

It seems that the advocates of a prescriptive reading of fundamental measurement theory are making a serious mistake. They think that, because physical measurement can be reconstructed in terms of axiomatic theory, it follows that psychology should construct psychological tests through the application of axiomatic theory. The justification for this conviction is usually cast in terms of logical, philosophical, and mathematical arguments, but as it stands it is no more than a belief. Moreover, the arguments adduced to support it are based on the fallacy that something that works in one scientific field (physics) will also work in a completely different field (psychology). Arguments for or against a psychological method, however, should be based on considerations that bear on psychology, and not on considerations that bear on physics. I congratulate the physicists with the lucky situation that the structure of the world they study is simple enough for them to reconstruct measurement procedures in terms of the lucid and powerful formal framework of fundamental measurement. However, in psychology we do not study stones, atoms, and quarks, but human beings. The human being is one of the most complex systems in the universe. Still, with respect to some psychological characteristics, human beings seem to vary from one another in a systematic fashion, and psychological measurement procedures attempt to capture this variation. Everybody knows that these practices are packed with assumptions, both substantial and auxiliary, and that the measurement outcomes they return should be interpreted with care. These measurement outcomes will more often than not violate the requirements of fundamental measurement. But to say that this implies that they are not measurements at all, and that the only way to construct measurement is to follow the axiomatic theory, is preliminary and hinges on an essentialist philosophy concerning the meaning of the term 'measurement'.

It is clear that, if one chooses to define measurement as homomorphic representation, then many assessment techniques that are commonly viewed as instances of measurement are not covered by this definition. So interpreted, these procedures do not yield measurements. If it makes the fundamental measurement theorists feel better, we may decide to call them 'fleasurements' instead – although I doubt whether that will make much of a difference. However, it is not a fact, either empirical, logical, mathematical, or otherwise, that measurement *is* homomorphic representation. At most, it is a convention. If we decide to designate by the term 'cow' every animal with eight eyes, then what we commonly recognize as being a cow no longer is, while spiders are suddenly in the possession of cowhood. While this will change the meaning of the word 'cattle' as well as the size of T-bone steaks, this juggling around of terms will not add much to the science of biology. The same holds for measurement in psychology. Nothing forces a specific definition of the term upon us, and nothing forces us to follow a specific approach towards psychological testing; certainly not when we consider the observation by Cliff (1992) that the axiomatic approach has not been able to produce a single striking psychological example to illustrate its benefits. Thus, it would perhaps be an idea for the advocates of the prescriptive reading of fundamental measurement theory to start showing the superiority of the approach, rather than to talk about it. For the prescriptive reading of the theory is not founded upon a serious consideration of the problems inherent in psychological measurement, but rather on a mindless mimicking of physics. In this context, it is ironic that the theory that originated as a reply to the overstated conclusions of Campbell and his associates has, in the hands of some of the more vigorous proponents of fundamental measurement, come to occupy the very same position it once sought to oppose.

4.3 Discussion

Representationalism offers a powerful conceptual framework for thinking about measurement. The focus on mapping empirical relations into a number system has liberated measurement theory from the demand that a concatenation operation must always be available, and as such it has provided a justification for moving away from operationalism. The logic of the theory allows one to see very clearly what one is assuming in different situations, and this is indeed an invaluable theoretical aid in model construction and evaluation. One should, however, not overinterpret the theory. As a formal framework, the theory can be considered adequate for its purpose, but as a philosophical or even prescriptive framework, it is too simplistic. In a philosophical interpretation, the clean logic that is the theoretical strength of the model quickly becomes its weakness. Representational theory has no natural means of incorporating error, and must abandon its central tenets when it is equipped with a method to do so. In face of this problem, a choice has to be made between two conclusions: Either measurement is impossible in the presence of error, or representational measurement theory is not a theory of how measurement is, can be, or should be carried out in practice. The first conclusion is absurd, but rejecting it leads immediately to the second, and this raises the question what the status of representationalism is. It has been argued here that representationalism offers a rational reconstruction of the measurement process. That is, it elucidates measurement procedures by recasting them in idealized logical terms, and it does this very well.

Whatever the conceptual status of rational reconstruction may be, however, it does not have prescriptive force. Therefore, the advocates of a prescriptive reading of the theory are not justified in their position. In effect, they are trying to sell a conceptual theory of measurement as a method for test construction and analysis. A method, however, can only be used if applicable, and because the inability to deal with imperfect observations is so deeply entrenched in the structure of representational theory, its applicability in the social sciences must be considered limited. It seems safe to assert that, in psychology, the clean observations that representationalism requires will not be realized in our time, if at all. Therefore, it is unclear what theorists like Michell (1990; 1999; 2000) and Kline (1998) are advocating. At best, they must be interpreted as proposing the use of unidimensional, additive IRT models, because these are the only models that allow for error *and* bear at least a superficial resemblance to additive conjoint structures. So interpreted, however, their claims do not seem all that radical. Moreover, the step from additive to nonadditive latent variable models, while philosophically important, is a small one from statistical, practical, and substantive points of view. Substantive considerations do not generally support additivity, as is evidenced by the fact that it is easy to give examples of test items that violate it but still measure the same latent variable. Statistical considerations suggest that the odds of finding a model with perfectly parallel item response functions are vanishingly small, so that the demand should not be taken overly seriously. Practical considerations lead to the conclusion that it will be virtually impossible to construct a situation, where a model that satisfies the usual IRT assumptions (monotonically increasing item response functions, unidimensionality, and local independence) fits the data, but the correlation between the simple sumscore and the latent variable drops below .90, which would seem more than enough for the average researcher. We must therefore conclude that positing additivity as a universal demand is, at best, preliminary.

Moreover, from a philosophical viewpoint, the difference between additive and nonadditive latent variable models seems much smaller than the difference between latent variable models and strict representations. A researcher working within representationalism is not making claims with respect to the question where the data came from - he is merely representing the data. From this viewpoint, additivity is central, because a violation of additivity precludes the possibility of homomorphic representation using a quantitative metric. The researcher who uses latent variable theory is engaged in a different activity. He knows (or should know) that homomorphic representation is strictly taken impossible because he is modeling stochastic relations that are not directly observable. In order to model such relations, he posits the existence of a latent variable. Because he is now introducing metaphysics, he needs a justification for these metaphysics, and this justification will not come from the data or from methodology. For instance, what type of latent variable model to use (e.g., a class model or a trait model), or how to conceptualize the relation between the latent variable and the observed variables (e.g., as additive or nonadditive, parametric or nonparametric), are typical examples of questions that cannot be answered by the data or by methodological considerations. Exactly because the researcher is now positing a data generating mechanism, rather than constructing a representation, these questions must be answered by substantive theory. For instance, developmental theory suggests that conservation data (Dolan, Jansen, & Van der Maas, *submitted*) should not be modeled as originating from a continuous latent variable model, but from a class model where the classes correspond to different developmental stages. Nothing in the data themselves forces this choice; it may well be possible to model the data using a continuous latent space. Neither do there exist methodological considerations that say developmental data should be modeled in this particular way. Clearly, the burden of proof shifts from the area of logic and mathematics to the area of substantive theory, which must give a justification for the metaphysics introduced.

Thus, a researcher, who conceptualizes a psychological construct as a measurement scale, is ascribing a completely different theoretical status to that construct as compared to a researcher who conceptualizes a construct as a latent variable. A measurement scale is a representation of observed relations, whereas a latent variable model is a guess about the structure of the data generating mechanism, i.e., a posited probabilistic explanation of such relations. Mathematically, the representationalist approach is as useful for studying latent variable models as for studying deterministic measurement structures. Its focus on mappings yields interesting insights into the relations between various levels of representation, and is a good theoretical tool in the study of different models and the relations between them. This holds true for the latent variable model as it does elsewhere. These insights, however, have more to do with the formal logic of the theory than with its philosophical account of what measurement is. The philosophical account of representationalism involves a highly restrictive empiricist point of view. It requires the direct observation of relations among objects in every case – the conjoint structures included. Representationalism could be considered to play an important role in empiricism, because it gives an account of how we get from qualitative observations to quantitative theoretical terms. This avoids metaphysical speculation, because the observations can be conceptualized as being made with the unaided eye. However, if we could judge the relation 'more intelligent than' directly, there would be no need for intelligence tests. From this point of view, the problem in psychological measurement is simply that the unaided eye does not work very well. It has to be supplemented by statistical assumptions concerning the behavior of aggregates, substantive hypotheses on the nature of data generating processes, and metaphysical postulates concerning the existence of propensities and latent variables. All this is required in order to get the endeavor off the ground in the first place. Representationalism does not pay attention to these problems but ignores them. This is fine as long as the concerns of the theory are limited to formal structures, but when interpreted as a conceptual framework for tackling the problem of psychological measurement in general, or even as a prescriptive framework for scale development, the theory can hardly be considered adequate. Thus, although representationalism is very important in elucidating some of the problems in psychological measurement, the mathematical structure of models, and the differences between measurement in the natural sciences and in psychology, its importance is limited and should not be overstressed.

5. RELATIONS BETWEEN THE MODELS

Three umpires are discussing their mode of operation and defending their integrity as umpires. "I call 'em as I see 'em," said the first. The second replied, "I call 'em as they are." The third said, "What I call 'em makes 'em what they are."
- R. L. Ebel, 1956

5.1 Introduction

The choice between different mathematical models for psychological measurement, of which this book has discussed three types, involves both an ontological commitment and a position concerning what one regards as measurement. The true score model is operationalist: It views any observed test score as a measure of a true score, where the true score is exhaustively defined in terms of the test score. The representationalist model is empiricist, but not operationalist. It views scales as constructed representations of the data, but it is highly restrictive in the kind of representation that counts as a measurement scale. The meaning of scales does not explicitly derive from a realist ontology regarding attributes, but neither is it defined in terms of a specific measurement procedure in the way the true score is. Latent variable models introduce an a priori hypothesis concerning the existence of theoretical entities. The latent variable model does not work its way up from the data, like representationalism, but posits an explanatory account of where the relations in the data came from. Thus, classical test theory is basically about the test scores themselves, representationalism is about the conditions that should hold among test and person characteristics in order to admit a representation in the number system, and latent variable theory is about the question where the test scores came from.

However, in spite of the fact that such philosophical differences between the approaches exist, they are also related in important ways. At one level, the relations between the models are clear. This is the level of syntax. Mathematically, it has been known for quite some time that strong relations exist between true scores and latent variables (Lord & Novick, 1968; Jöreskog, 1971; Hambleton & Swaminathan, 1985). It has also been observed that special cases of latent variable models bear a strong relation to specific versions of the representationalist model (Brogden, 1977; Fischer, 1995; Perline, Wright, & Wainer, 1978; Roskam, 1984; Scheiblechner, 1999). Such relations also exist between classical test theory and representationalism, if the classical test model is extended with the appropriate assumptions, as was already suggested by Lord & Novick (1968, Ch. 1) and is illustrated below.

Thus, mathematically speaking, the models are strongly related, and sometimes a special case of one model is also a special case of another model. A question that has, however, been largely neglected is what kind of interpretation has to be given to the concepts in these models in order to maintain their interrelatedness at a semantic level. And an even more interesting question that has, to the best of my knowledge, never been addressed is the question whether these relations could also be conceptualized to hold at the ontological level. That is, does there exist an ontological viewpoint upon which the models are not in contradiction, but supplement each other? It will be argued in this chapter that such a viewpoint exists under one condition. The condition is that the probability semantics in the true score and latent variable models are interpreted at the level of the individual, that is, if the probabilities in the models are interpreted as propensities. If this is the case, then the models are syntactically, semantically, and ontologically related, and merely address different levels of the measurement process. However, as soon as the existence of propensities is denied, the models are decoupled in all these senses. In that case, the true score model is necessarily false, the latent variable model is exclusively about relations between characteristics of subpopulations, and the representationalist model is solely about deterministic relations.

5.2 Levels of connection

We can address the individual theoretical terms in the measurement models at different levels, and therefore we can also discuss the relations between these terms at different levels. I will concentrate here on the levels of syntax, semantics, and ontology. It will be shown that, while the syntactical connections are easily established and straightforward, the semantical and ontological connections leave much freedom of interpretation. An integrated theoretical framework for discussing the models will be presented, but it will also be shown that this framework collapses as soon as the propensity interpretation of the probabilities in the models is denied.

5.2.1 Syntax

Latent variables and true scores Syntactically, the true score model and the latent variable model are closely connected. In fact, they are so closely connected that the distinction between true scores and latent variables may get blurred in certain situations. It is suggested by Schmidt & Hunter (1999, p. 185), for example, that the relation between true scores and latent variables is 'usually close enough to linear' so that the latent variables approach has no conceptual or practical advantage. This is not the case, because whether there is any relation in the first place depends on the dimensionality of the latent variable model, which is not tested in the classical test model. The mistake made by Schmidt & Hunter (1999) is understandable, however, because *if* a unidimensional model holds *then* it will often be possible to construct a simple sumscore that can reasonably be used as a proxy for the latent variable in question.

Consider the Item Response Theory model for dichotomous items. It is well known (e.g., Lord & Novick, 1968; Hambleton & Swaminathan, 1985) that in this case the expectation of the sumscore is a function of the latent variable. Suppose subject i's sumscore X is defined as the sum of his item responses on N items, 1, $\ldots$, j, $\ldots$, N. Let U_{ij} denote i's response to the j^{th} item. Thus, $X_i = \sum_{j=1}^{N} U_{ij}$ and i's true testscore is $t_i = \mathcal{E}(X_i)$. For a fixed test consisting of dichotomous items, there exists a monotonic relation between t and the latent variable θ. The true score is the sum of the individual item response probabilities under the IRT model:

$$t_i = \mathcal{E}(X_i \mid \theta_i) = \sum_{j=1}^{N} P(U_{ij} = 1 \mid \theta_i). \tag{5.1}$$

If the IRT model is parametric, then the function relating t to θ is also parametric and can be used to linearize the sumscore so that equal distances in the latent variable match equal distances in the transformed sumscore. For some models, like the Rasch model, the function that does this is so simple (the natural logarithm of $(X_i/N)/[1 - (X_i/N)]$) that it can be implemented on a pocket calculator. For nonparametric IRT models, no parametric function for the relation exists, but under relatively mild assumptions the latent variable still is stochastically ordered by the sumscore (Hemker, Sijtsma, Molenaar, & Junker, 1997). Thus, conditional on the assumption that a unidimensional model holds, the true score will often be strongly related to the latent variable. This can also be seen from the fact that Jöreskog (1971) actually derived the congeneric model for continuous responses by introducing the requirement that the true scores be perfectly correlated. In this case, each true score is a linear function of every other true score, which means that all true scores can be conceptualized to be a linear function of a single factor score. Although the true score model is usually seen as weaker than the latent variable model, Jöreskog in fact introduced the congeneric model by replacing the classical test theory assumption of essential tau-equivalence with the weaker assumption that the tests are congeneric. The true score model for continuous test scores that satisfy essential tau-equivalence is thus nested under the common factor model; it can be derived by introducing the restriction that the factor loadings are equal.

These results are easily misinterpreted and overgeneralized to the conclusion that there is basically no difference between the latent variable and true score models. This conclusion is erroneous because the relation does not hold in general. For instance, in the case of polytomous IRT models, the latent variable is generally not even stochastically ordered by the sumscore. In latent variable models with correlated errors, which are not uncommon in SEM, the relations will also be more complicated, and in case of multidimensional latent variable models the relations break down quickly. Finally, if no latent variable model holds at all, we may still conceptualize a true score, because the only assumption that is necessary for the definition of a true score is that the propensity distribution on which it is defined is nondegenerate and has finite variance (Novick, 1966). However, it is obvious that, under the proper conditions, the true score bears a functional rather than stochastic relation to the sumscore. Thus, the relation between the true score model and the latent variable model is mathematically explicit in some cases, and indeed is a strong one.

Latent variables and scales There are also strong connections between the latent variable model and the additive conjoint measurement model. Specifically, special cases of latent variable models, in particular additive versions of such models, can be considered to be mathematically covered by the additive conjoint model. The class of models for which this connection can be set up is quite general (Scheiblechner, 1999), but for clarity of exposition attention is limited here to the Rasch (1960) model. The Rasch model hypothesizes the expected item responses (true item scores) to be a logistic function of the latent variable. Thus, subject i's response to item j is assumed to follow the function

$$P(U_{ij}) = \frac{e^{\theta_i+\beta_j}}{1+e^{\theta_i+\beta_j}}, \tag{5.2}$$

where $P(U_{ij})$ is the probability of a correct or affirmative answer and β_j is the location of item j, conceptualized as the point on the θ scale where $P(U_{ij}) = 0.5$. Now, a monotonic transformation of the item response probabilities will yield a simple additive representation. Specifically, the model can be rewritten as

$$\ln\left[\frac{P(U_{ij})}{1-P(U_{ij})}\right] = \theta_i + \beta_j, \tag{5.3}$$

where ln denotes the natural logarithm. The axioms of additive conjoint measurement hold for the model in stochastic form.

First, the $P(U_{ij})$ form a weak order by definition: Transitivity (if $P(U_{ij}) \succeq P(U_{kl})$, and $P(U_{kl}) \succeq P(U_{mn})$, then $P(U_{ij}) \succeq P(U_{mn})$) and connectedness (either $P(U_{ij}) \succeq P(U_{kl})$, or $P(U_{kl}) \succeq P(U_{ij})$, or both) must hold because probabilities are numerical, and numbers are ordered. This interesting fact seems to result from the imposition of the Kolmogorov axioms on the probabilities, which, as a result, are ordered by assumption.

Second, the independence condition holds. That is, item difficulty and person ability are seen as the two independent variables, and items and subjects are independently ordered on ability and difficulty, respectively, by the dependent variable $P(U_{ij})$. Rasch (1960) actually derived the model from the requirement of parameter separation, i.e., it should be possible to estimate the ordering of items and subjects independently, which basically comes down to the same type of requirement as posed by the independence axiom in the additive conjoint model. Rasch called this property specific objectivity. Statistically, this implies that the item and person parameters can be estimated independently, because the sumscore is a minimally sufficient statistic for the person parameter, which enables parameter estimation by Conditional Maximum Likelihood (Andersen, 1973).

Third, if the Rasch model is true, then the double cancellation condition is satisfied. If, for any three levels of ability and any three levels of item difficulty, if it is true that

$$\theta_2 + \beta_1 \geq \theta_1 + \beta_2 \tag{5.4}$$

and it is also true that

$$\theta_3 + \beta_2 \geq \theta_2 + \beta_3 \tag{5.5}$$

then

$$\theta_2 + \beta_1 + \theta_3 + \beta_2 \geq \theta_1 + \beta_2 + \theta_2 + \beta_3, \tag{5.6}$$

so that

$$\theta_3 + \beta_1 \geq \theta_1 + \beta_3 \tag{5.7}$$

and double cancellation holds. Thus, the structure of the Rasch model sustains representational measurement theory. As soon as the model is extended with a discrimination parameter, as in Birnbaum's (1968) model, this resemblance vanishes because the independence condition will no longer hold.

Scales and true scores The fact that the latent variable model can be constructed from the imposition of restrictions on the relations between true scores, and the fact that additive latent variable models are special cases of representational measurement theory, suggests that appropriately constructed versions of the classical model can be written in representational form too. For instance, the true score model for tau-equivalent tests assumes that for any two true scores of person i on tests j and k, denoted t_{ij} and t_{ik}, it is true that $t_{ij} = c + t_{ik}$, where c is constant over persons. The structure of the model can be written in terms of a common factor model (Jöreskog, 1971):

$$\mathcal{E}(X_{ij}) = \nu_j + \lambda\theta_i \tag{5.8}$$

where the ν_j parameter is a test-specific intercept term that absorbs the effect of the constant c in the definition of tau-equivalence, λ is the factor loading, and θ_i is subject i's position on the latent variable. Because, by the definition of tau-equivalence, λ is constant over tests, it has no test subscript as in the congeneric model. We may set it to unity without loss of generality. This gives the additive representation

$$\mathcal{E}(X_{ij}) = \nu_j + \theta_i. \tag{5.9}$$

The axioms of conjoint measurement then hold for the so constructed model. The instrumental comparison is made through the true scores on the tests, as it is made through the item response probabilities in the Rasch model. The true scores induce an ordering because, like probabilities, true scores are numbers and numbers are ordered. The condition of independence holds because the item and person effects do not interact (this would occur if the factor loadings differed across items): Persons can be stochastically ordered by true scores, regardless of which test is used for this purpose, and tests can be stochastically ordered by true scores, regardless of which person is used for this purpose. That the double cancellation axiom holds is obvious, because the additive decomposition of the observed scores into a test and person specific part guarantees this to be the case; one may follow the line of reasoning as discussed above for the Rasch model and substitute ν_j for β_j.

Because the fundamental measurement model works its way up from relations between objects, and the presently formulated relations are indistinguishable from the relations assumed to hold in the true score model with essential tau-equivalence, the classical test theory model allows for an additive conjoint representation under the restriction of essential tau-equivalence. It is interesting to note that such a representation cannot be constructed under the stronger conditions of tau-equivalence and parallelism. Both tau-equivalence and parallelism assume equal true scores across tests, which means that the intercept terms ν_j are equal across tests. This implies that the true scores cannot induce an ordering in these tests, so that the additive conjoint model cannot be formulated.

So, the true score, latent variable, and additive conjoint models are strongly related syntactically. Imposing appropriate restrictions on the models allows one to juggle the terms around so as to move back and forth between the mathematical structures. The true score model with the essential tau-equivalence restriction seems to serve as a bridge between the latent variable model and the additive conjoint model: It is a special case of the latent variable model, and the restrictions it poses on the true scores guarantee that an additive representation is possible. On the other hand, there are syntactical differences between the models that should not be forgotten; one can formulate latent variable models that are nonadditive and therefore do not generate the possibility to construct an additive conjoint representation; the true score model can be formulated without invoking a latent variable, and latent variable models can be constructed where the true score bears no direct functional relation to the latent variable (i.e., multidimensional models, models with correlated errors, or models for polytomous items); and the additive conjoint model can generate deterministic structures that render the true score undefined (i.e., the propensity distribution is non-existent or degenerate, depending on one's point of view) and the latent variable model obsolete (i.e., trivial or unnecessary, depending on one's point of view). Nevertheless, under the right conditions, there is a strong correspondence between the models. The question now becomes: What kind of semantics do we need to relate the models not only in terms of mathematics, but to keep a consistent interpretation of these relations, and what kind of overall conceptualization of the measurement process would this give?

5.2.2 Semantics and ontology

The semantics of true score theory, latent variable models, and representational measurement are markedly different, as should be clear from the preceding chapters. The reason that the models can nevertheless be related syntactically is that, in the above discussion, the models were uncritically defined on probabilities and relations among them. However, we have seen in the preceding chapters that the interpretation of the probability calculus is not straightforward in the case of psychological testing. In the true score model, probabilities must be interpreted as propensities which are defined at the level of the individual; in the latent variable model, they may either be interpreted as such propensities, or as characteristics of subpopulations; in the additive conjoint measurement model, the observations are assumed to be free of measurement error, so that no interpretation of probability is necessary at all. In order to set up the above connections, we have required the representational model to take a step back from its empiricist foundation, and to grant the existence of probabilities of some kind, but we have not yet interpreted these probabilities. Neither have we made a choice with regard to the conceptualization of the item response probabilities in latent variable models. If we are going to interpret the connections between the models, we will have to make such a choice.

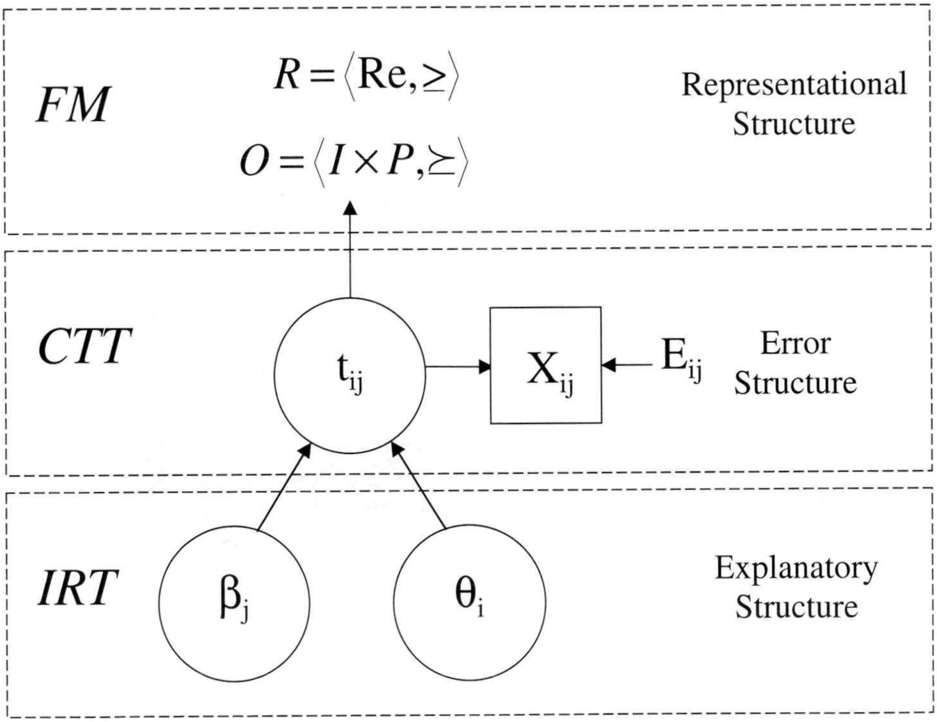
Admitting propensities As is so often the case, the most elegant situation occurs if we introduce the strongest metaphysics. This, of course, comes down to a propensity interpretation of the probabilities in the model. In this case, we conceptualize the probabilities as propensities that are uniquely defined for a particular person at a particular time point. Interpretation of these probabilities will in general require a thought experiment like Mr. Brown's infamous brainwash.

In this interpretation, the true score, latent variable, and representationalist models are strongly related. Semantically, true score theory discusses the relation between propensities and observables; latent variable theory posits a hypothesis to explain the relations between propensities; and representationalism shows the conditions necessary to construct a representation that preserves the relations between subjects, where these relations are defined indirectly via the propensities. Thus, true score theory describes, latent variable theory explains, and fundamental measurement represents. Moreover, under appropriate conditions the models are not at odds with each other; they simply focus on different levels of the measurement process. This is graphically represented in Figure 5.1.

As the figure illustrates, we have a division of labour between the different theories. Classical test theory provides a theory of the error structure. It does so by defining the true score as the expected value of the propensity distribution for subject i on item or test j. Latent variable models, such as the item response theory model, provide a hypothesis concerning the data generating process. The hypothesis is that there exists variation on an attribute (the latent variable) which produces variation in the true scores. The item difficulty (which could be the intercept term in a continuous model) also produces such variation. In the figure, these person and item effects are represented as independent.

The true scores can be used for the instrumental comparison $\succeq$ of the Cartesian product terms (i, j), which are defined on the Items × Persons matrix, denoted $I \times P$ in Figure 5.1. The true scores will form a weak order because they are already numerical. Because the effects of item difficulty and latent variable are independent, the instrumental comparison will allow for the independent ordering of items and subjects. This gives the empirical relational system $\mathcal{O} = \langle I \times P, \succeq \rangle$. Perhaps, it should be called a quasi-empirical relational system, because it is defined on unobservable propensities. The fact that the effects of person ability and item difficulty are independent guarantees that, if the model is true, a transformation of the true scores can be found that yields an additive representation, as is the case in the Rasch model. The so constructed representation is the numerical relational system $\mathcal{R} = \langle \text{Re}, \geq \rangle$. Together $\mathcal{O}$ and $\mathcal{R}$ form an additive conjoint measurement structure. What is represented is the trade-off between item and person effects in producing true scores. The representation of this trade-off is invariant up to a linear transformation, so it is measured on an interval scale.

Figure 5.1. The relation between Item Response Theory, Classical Test Theory, and Fundamental Measurement Theory.


FM
$R = \langle \text{Re}, \geq \rangle$
$O = \langle I \times P, \succeq \rangle$
Representational Structure
CTT
t_{ij}
X_{ij}
E_{ij}
Error Structure
IRT
β_j
θ_i
Explanatory Structure

The division of labour highlights the different functions of the theories. For instance, in the present conceptualization one would not say that the Rasch model is a fundamental measurement model, but one would say that the Rasch model describes (one of the) hypothetical data generating mechanisms that would produce data that allow for an additive representation in the fundamental measurement theory sense. This is a large conceptual difference, that lies primarily in the different ontological status of the numerical representation, which is a construction even if based on relations between propensities, and the latent variable, which is a hypothetical attribute that underlies relations between propensities. A related difference is that the latent variable model is a hypothesis on the data generating process, and therefore claims more than the relations it implies in the data. The representation does not have this property, because it is not a posited explanation of the relations in the data, but a representation of these relations. That is, one can say that the latent variable model is true or false, but one cannot say that a homomorphic representation is true or false; one can only say that it can or cannot be constructed.

Note also that the representation is purely hypothetical: Because unsystematic variance is introduced in the error structure, there is no representation of observed relations as in typical fundamental measurement theory models. So, strictly taken, it is impossible to actually construct the desired representation on the basis of observed data. It is, however, the case that, if the model were true, then a representation could be constructed if the true scores were observed. True scores cannot be observed, so that the representational account must then be viewed as inherently based on a counterfactual line of reasoning. So, even if the latent variable model were true, the representation would stay counterfactual as long as we cannot observe true scores. It think that this is why, in latent variable models, it is more usual to say that one estimates a person's position on the latent variable, than to say that one measures that position. This difference in terminology also seems to reflect the ontological difference between a latent variable and a measurement scale. Thus, in the present scheme, the models are about as closely connected as possible, but the difference in ontological tenets remains: latent variables are entities that figure in an explanation of how relations in the data arise, while measurement scales are constructed representations of the relations in the data.

The truth of a latent variable model must be considered conceptually independent of the possibility to construct a fundamental measurement theory representation. In principle, the latent variable model may be true, while it is impossible to construct a homomorphic representation, and it may be possible to construct such a representation, while the latent variable model is not true. An example of the former situation would occur in case a common factor model with unequal factor loadings, or a Birnbaum model, were true. An example of the latter situation would occur when the relations in the data would admit a fundamental measurement representation, although no latent variable were responsible for these relations, as in the case of spurious Guttman scaling discussed in section 4.2.3., and in the coin-tossing example discussed by Wood (1978). This does not mean that either theory is in some sense inadequate, but that latent variable and fundamental measurement theory are concerned with distinct problems.

Dismissing propensities Assuming the existence of propensities allows for connecting the latent variable, true score, and fundamental measurement models. However, if one dismisses the existence of propensities, the unified picture discussed above falls like a house of cards.

First, if propensities do not exist, then the true score model in the Lord & Novick (1968) formulation is necessarily false. This is because in this interpretation, the observed score can no longer be viewed as a realization of a random variable at the level of the individual, which means that the true score model cannot be constructed. If no randomness is associated with the item response process, then the probability distribution on which the true score should be defined is degenerate, and the core assumption of the true score model (Novick, 1966) is therefore violated. Sentences like 'the reliability of this test is .88 in population X' cannot, in principle, be true in this interpretation. At most, one could rephrase such sentences in the counterfactual form, and state that 'if the observed scores had been generated by a random process, etc., then the reliability of the test scores would have been .88 in population X'. Such counterfactuals may be useful and informative, but the place and conceptual status of counterfactual information about test scores would require some serious rethinking of the use of classical test theory in test analysis.

The latent variable model could be true without a propensity interpretation if a repeated sampling perspective is adopted. The validity of latent variable models would then be relevant only at the level of aggregate statistics; because there is no randomness associated with an individual's item responses, the models would be necessarily false at the individual level. In this interpretation, the connection between true score theory and latent variable models breaks down. Since there is no randomness at the individual level, there is no true score, and a statement to the effect that the true score is monotonically related to the latent variable cannot be made. In a repeated sampling interpretation, the latent variable model states that differences between subpopulations at the latent level lead to differences between subpopulations at the observed level, and nothing more. The model uses probability semantics and the expectation operator, but only to deal with sampling variation; the expectation is conceptualized as a population mean, as it is in standard population-sampling schemes in statistics, and not as a true score. To interpret such models as process models that apply at the level of the individual amounts to a logical fallacy. Nevertheless, the basic idea of latent variable models, which is that variation in the latent variable produces variation in the observed scores, may be maintained, elaborated upon, and endowed with a substantive theoretical interpretation.

For fundamental measurement theory, denying the validity of probability assignments at the individual level has no theoretical consequences. Since the model is most naturally stated in deterministic terms in the first place, the theory does not have to be modified or reinterpreted when responses are stripped of randomness at the level of the individual. Such a view does lead to the conclusion that the axioms of fundamental measurement are usually not satisfied by empirical data, either in psychology or elsewhere. This observation, of course, is hardly surprising, given the strict demands of the theory. What is interesting, however, is that the connection between probabilistic latent variable models and fundamental measurement breaks down if the propensity interpretation is denied. For instance, the stochastic dominance relations as discussed by Scheiblechner (1999) no longer apply, because they are defined on true scores, which are no longer admitted in the present interpretation. Thus, the only item response model that is properly admitted as a case of fundamental measurement in this interpretation, is the deterministic Guttman model. It is thus clear that the popular view, which holds that the Rasch model 'is' a fundamental measurement model (Perline, Wright, & Wainer, 1978; Scheiblechner, 1999; Bond & Fox, 2001), is parasitic on the stochastic subject interpretation of the item response model. Once that interpretation is denied, the Rasch model has little to do with fundamental measurement. In fact, the only thing that conjoint measurement and Rasch models have in common, in this interpretation, is additivity.

Thus, there are at least two ways of looking at the relations between the different theories of measurement discussed in this book. The similarities and dissimilarities between these models depend on a rather high level philosophical assumption, namely on whether one wants to admit propensities into psychological measurement or not. Admitting propensities gives a consistent and unified picture, in which the different approaches focus on different parts of the measurement process, but are not necessarily at odds. Denying the existence of propensities immediately destroys the classical test model, and leaves one with two models that have relatively little to do with each other.

5.3 Discussion

This chapter has aimed to clarify the relations between the models discussed in this book. It has been shown that the models are syntactically related in quite a strong sense. However, when viewed from a semantic perspective, whether these connections continue to hold depends on the interpretation of probability: Probability must be interpreted in a propensity sense, otherwise the models are unrelated. In spite of this, the difference in ontological tenets with respect to the central concepts in the models (i.e., true scores, latent variables, and scales) remains, regardless of the interpretation of probability. These conclusions will be examined in somewhat greater detail in the next sections. First, the general problem concerning the theoretical status of measurement concepts will be discussed. Second, I will shortly review arguments for and against the propensity and repeated sampling interpretations of probability. Third, further differences between the latent variable model, on the one hand, and the representational model, on the other, will be discussed in terms of the degree of experimental control that they presuppose, an issue that proves to be closely connected to the local homogeneity condition discussed in Chapter 3. Finally, the models will be evaluated in terms of the semantics they yield for validity; in this section, it will become apparent that, when the models are required to specify a relation between the observed scores and a theoretical attribute, both the classical test theory model and the representationalist model converge to a latent variable formulation; classical test theory because it has to be strengthened, and representationalism because it has to be weakened. The formulation of validity so reached has important consequences for validity theory in general, which will be discussed in the next chapter.

5.3.1 Theoretical status

It is instructive to review the conclusions reached in this book with respect to the theoretical status of the central concepts in the measurement models discussed. We have seen that classical test theory defines the true score in terms of the expectation of a series of replications of the same item or test. It has been argued in Chapter 2 that it does not make sense to say that two tests x and y 'measure' the same true score, as is suggested in the definitions of parallelism and tau-equivalence. It does make sense to say that the true scores on test x and test y have the same numerical value, but this is a statement of an entirely different character. The fact that the true score is explicitly defined in terms of a particular test, implies that the meaning of the true score is exhausted by reference to the operations that lead to it. That the operations require brainwashing and cannot be carried out is peculiar, but does not refute this conclusion. Thus, the psychologist who defines intelligence as a true score takes an operationalist position with respect to the construct. He cannot do otherwise.

Latent variable theory supplements classical test theory precisely by broadening the meaning of the theoretical terms in the model. Latent variables are not exhaustively defined by a series of operations, otherwise two distinct tests could not measure the same latent variable. That latent variable theory allows for the statement that different tests can measure the same latent variable is obvious; if this were not possible, common applications like test equating and adaptive testing would lack a theoretical basis. That they do not lack such a basis means that the latent variable has surplus meaning over the observation statements and the operations that lead to them. It is not an operationalist concept. Upon this conclusion, the question occurs whether the theoretical term 'latent variable' must be taken to refer to reality or not. It seems to me that it should be taken to do so. Several arguments for this conclusion have been adduced in Chapter 3. I would like to discuss one other argument because it brings out clearly the difference with representationalism.

It has been observed several times that the syntactical equivalence between probabilistic versions of additive conjoint measurement and latent variable theory breaks down if we allow the slope of item response functions to differ across items, as is the case in the congeneric model and in the Birnbaum model. Mathematically speaking, the reason for this is very simple, because it means that no additive representation is possible if additivity is violated, which comes down to the trivial observation that additivity is violated if additivity is violated. Conceptually, however, there are more interesting things going on.

What the existence of nonadditive latent variable models illustrates, is that latent variable theory not only allows for the possibility that different items measure the same latent variable, but that it also allows for the even stronger claim that a given set of items can measure the same latent variable differently in different subpopulations. This is clear from the fact that nonadditive latent variable models imply that items have different difficulty orderings in subpopulations high and low on the trait.

Similar considerations play a role in the definition of bias with respect to group membership. The concept of bias means that the expected value of an item response differs across groups, conditional on the latent variable, for at least one position on that latent variable. Such a situation occurs, for instance, when females have a lower expected item response on an IQ-item than males, where the comparison is between subpopulations of males and females that have the same level of intelligence. This method of conditioning on the latent variable is very common in latent variable models. It is highly interesting.

The reason for this is the following. What do we assert when we say that an item has different expected values across groups, conditional on the latent variable? It seems to me that we are in effect asserting that the item has different expected values across groups, conditional on the *same* latent variable. What we have to assume, then, is that the item *does* measure the same latent variable across groups. Otherwise it would be meaningless to condition on this latent variable. The problem formulated in item bias is not, therefore, that the item in question measures a different latent variable in each group, but that it measures the same latent variable differently in each group. Thus, not only is it the case that latent variables are not exhaustively defined by the items that measure them; they are not even exhaustively defined by the item response functions. For if the latter were the case, this would preclude the formulation of item bias. And nothing precludes the formulation of item bias.

The common practice of conditioning on the latent variable across groups with different response functions presupposes a kind of meaning invariance of the latent variable concept. Now, this invariance cannot be reduced to the fact that a particular set of items is used, as in operationalism, for this would preclude the possibility of unidimensionality and adaptive testing. It cannot be reduced to the ordering of the items, for in a nonadditive model this ordering is not invariant across trait levels. It cannot be reduced to the invariance of item response functions, for these may be different across groups. And it cannot be reduced to the invariance of theoretical relations in which the latent variable enters, for these will also be different across groups (for instance, a latent variable may be correlated to some other variable in one group but not in another, while we are still talking about the same latent variable). Where, then, does this meaning invariance come from? What would allow us to say that we are measuring the same latent variable in all these cases? It seems that this meaning invariance can only be upheld if the latent variable is granted an existential status that is essentially independent of the measurement procedure or the theory in which it figures. Thus, the psychologist who views a theoretical concept like intelligence as a latent variable must subscribe to a realist position.

It has been argued in Chapter 4 that representationalism is based on an empiricist philosophy of science. Its central concept, the measurement scale, is a constructed representation of relations between the objects measured. Can a representationalist formulate a concept such as item bias? It seems to me that this will be fairly difficult. Suppose that we have two populations A and B, and that in each population the responses on a three item scale, consisting of items j, k, and l, conform to a Rasch model. Further suppose that item l is, in latent variable terms, biased, and that it is biased to such a degree that it is more difficult than item k in population A, but less difficult than item k in population B. So, in each population an additive conjoint representation is possible, but in the union of these populations it is not. Now, the latent variable theorist could, in principle, allow for the different item orderings in each population and still estimate the position on the latent variable. He could even compare the populations with respect to the latent variable distributions. This may, in many cases, be objectionable from a substantive point of view, but it is logically and technically possible (see Borsboom, Mellenbergh & Van Heerden, 2002-b[1], for some examples where this procedure may also be plausible from a substantive point of view). However, the important point is not

[1] This paper is included in this dissertation as Appendix B.

whether this would be generally appropriate, but that nothing in the formulation of the latent variable model precludes it. The representationalist does not seem to be in a position to take such a course of action. The qualitative relations mapped into the numerical domain in population A are different from those in population B. Because measurement is representation, it seems to me that the representationalist must say that something different is being measured in each population, not that the same attribute is being measured differently. The representationalist cannot therefore assume the kind of meaning invariance that the latent variable theorist can.

The reason for this lies in the different ontological tenets of the models. If the representationalist cannot construct a representation, nothing is measured; he cannot reify a measurement scale without contradicting himself. The latent variable theorist can imagine the wildest situations because he takes the ontological freedom to postulate a latent variable, and take it from there; the representationalist cannot imagine any measurement situation where he could not construct a homomorphic representation on the basis of empirical relations, for such a situation would not allow for use of the term measurement. Thus, the representationalist model does not have the metaphysical richness to allow one to posit the existence, in reality, of more than the relations in the data to be represented. Where the latent variable theorist cannot keep a consistent position *without* a realist interpretation of latent variables, the representationalist cannot keep a consistent position *with* a realist interpretation of measurement scales.

The researcher who views intelligence as a measurement scale thus takes a constructivist position with respect to the attribute in question. Because the existence of a measurement scale depends on the possibility to construct it, such a researcher must moreover conclude that general intelligence does not exist at the present time, because nobody has constructed a general intelligence test that allows for a homomorphic representation. However, all is not lost, because it also follows from the identification of intelligence with a measurement scale, that general intelligence may come to exist tomorrow at 2.14 PM, if someone were to construct a homomorphic mapping of general intelligence test items at that particular time. This kind of relativism with respect to theoretical entities is strongly reminiscent of positivism.

These observations are relevant with respect to the theoretical status of psychological constructs in general. Of course, positions of all kinds can be defended for a construct like intelligence. The reason for this is that the theory of intelligence is not formulated in sufficient detail to imply a realist, constructivist, or operationalist position. So, one may hold the view that intelligence is a causally efficient entity, or that it is just a heuristic concept, useful to organize our observations, or that it is a dispositional characteristic, or that it is a social construction, and so forth. But when a construct like intelligence is related to the observations, some kind of measurement model must come into play. And it is at this point that the researcher must commit to an ontology for the construct. If he is an operationalist or constructivist, he should not let himself be drawn into latent variable models; for then he will have to posit an ontological position that is too strong. If he is a realist, then research conducted within the framework of classical test theory cannot be considered to put the proposed ontology to the test. If he does not want to commit to realism, but neither to operationalism, he may opt for representational measurement theory.

If I am correct in my analysis, psychology suffers from a substantial conceptual confusion in the interpretation of its theoretical terms. For instance, some researchers in personality give the impression that executing a principal components analysis tests the hypothesis that the Five Factors of personality are real and causally efficient entities. A principal component analysis, however, is a special case of the formative model discussed in Chapter 3, so as far as I am concerned this specific ontological tenet (which is the subject of heated discussions; Pervin, 1994) has not been tested in such research. Similarly, many people working in latent variable theory seem to regard latent variables as nothing more than economic representations of the data. However, commonly used latent variable model are usually not representations of the data in a rigorous fundamental measurement theory sense, and it is unclear why one would need latent variables analysis for economic representations in a less rigorous sense; principal components seem good enough, and are much easier to obtain. Others think that a factor in a factor analysis is the 'common content' of the items; but this is also inconsistent, for common content is a characteristic of items, while a latent variable is a characteristic of subjects. Finally, I suspect that the majority of researchers in psychology, who hold a realist position with respect to their constructs, will not hesitate to equate these constructs with true scores; a position that is, in general, inconsistent.

Is this important? That depends on the situation. I personally feel that the most serious mistake consists in asserting realism about constructs on the basis of the wrong model. Somebody who thinks that he has proven the existence of general intelligence because one principal component had an Eigenvalue larger than one, or because Cronbach's α was over .80, has never tested the ontological claim involved. Such cases abound in psychology. Of course, someone who has successfully fitted a unidimensional latent variable model has not proven the existence of a latent variable either, but at least that hypothesis has been tested, however indirectly. Mistaken reification seems to me the most serious fallacy that can be made with respect to the problems discussed here. The other mistake, i.e., claiming that no theoretical concept in the models discussed could ever exist, does not seem so grave. I see no particular problem with an intelligence researcher who neither believes that intelligence exists, nor that such a hypothesis is tested in a model of any kind. One could say such a person is perhaps being overly skeptic, but the skeptic has a philosophical problem, not necessarily a scientific one. Moreover, skeptics usually play a healthy role in the scientific discussion, while communities of believers seem to be able to propagate mistaken conclusions indefinitely. This is especially true of psychology, where ontological realists about attitudes, personality traits, and general intelligence, are hardly ever pressed to use the right model for testing their claims.

5.3.2 The interpretation of probability

The interpretation of the theoretical status of the discussed models, the theoretical terms figuring therein, and the relations between these models, were seen to depend crucially on the interpretation of probability. Obviously, neither the stochastic subject nor the repeated sampling interpretation of probability is logically imposed upon us. Can we nevertheless force a choice between these interpretations? For example, could such a choice be defended on more general metatheoretical principles?

From this point of view one may, for instance, argue that the stochastic subject interpretation is flawed, because Mr. Brown's brainwash is simply a ridiculous and inadmissible thought experiment. However, the interpretation of probability in models like the ones discussed here always requires a thought experiment of one variety or another. Mr. Brown's brainwash is the variant that goes with the stochastic subject interpretation. The repeated sampling interpretation, however, no less requires a thought experiment. Usually, we are not sampling at random from well defined populations, as the statistician would like us to do. In fact, generally nothing that resembles the statistician's idea of sampling has occurred in the first place; in psychology, 'sampling' often merely means that not all six billion people on this earth have been tested. Thus, the random sampling view must also take recourse to a thought experiment – this time in terms of hypothetical repeated sampling from a subpopulation of people with the same position on the latent variable – if an interpretation of its terms is asked for. Moreover, the population in question will often be idealized. For instance, the population may be assumed to be normally distributed over a continuous latent variable, which is unrealistic if only because there are not enough people to realize that assumption. Thus, the introduction of a thought experiment seems unavoidable in both interpretations, and it may well be unavoidable in applied statistics in general (Borsboom, Mellenbergh, & Van Heerden, 2002-a). One cannot argue that the propensity interpretation must be discarded because it invokes a thought experiment, for the repeated sampling interpretation does so too. At best, one could argue that one of the interpretations should be favored because it introduces a 'better' thought experiment, but I do not see what the grounds for such an argument could be.

One could also claim that propensities should be cut away by Occam's razor, because they are superfluous: The model can be formulated without mentioning propensities. Ellis (1994, p. 5) quotes a personal communication with Paul Holland, in which the latter is reported to have said that "...the stochastic subject hypothesis is a bad hypothesis. Like God, it is not needed". Such an argument may seem attractive, but I think it it oversimplifies the problem. First, it is most certainly not the case that propensities do no theoretical work at all: We have seen in this chapter that, at the very least, they yield a unified and consistent picture of psychometric theory. And unification could be seen as a metatheoretical principle with about equal force as the parsimony principle. Moreover, the psychologist who maintains that his theory is about propensities is justified in using these propensities to derive predictions with respect to between-subjects data. That his predictions could also be derived from a theory which does not mention individual level propensities means that the theory is underdetermined by empirical data; but this cannot be taken to be a decisive argument against his use of propensities, because every theory is underdetermined by empirical data. And that there is usually an alternative explanation of the between-subjects data, which does not use propensities, does not imply that such an alternative explanation is plausible; in fact, it may well be that no substantive interpretation is available for that explanation, so that it remains a purely statistical oddity. Thus, although the introduction of propensity undoubtedly introduces a metaphysical element in a psychological theory, one cannot say that it should therefore be considered inadmissible, unless one holds an unduly narrow view of what is admissible in scientific research.

Perhaps, many more philosophical arguments for one or another interpretation could be given. However, I think that none will be decisive. Methodological principles or philosophical arguments do not have enough force to clinch this problem. This may have to do with the fact that the interpretation of probability is an intricate problem in general, and not just in psychometric models (e.g., Nagel, 1939; Fine, 1973; DeFinetti, 1974; Popper, 1963; Hacking, 1965). No decisive argument has, to my knowledge, ever been presented for or against a specific interpretation of probability, and there seems no reason to expect that such an argument would be available in the present situation. If this is correct, i.e., if the choice between these interpretations cannot be motivated on general principles, then it must be motivated on other grounds. It would seem that the problem should then be passed on to substantive psychological theory. And this brings us back to a problem that was already discussed in Chapter 3: Namely, what is the range of application of theoretical constructs? That is, do they apply to individuals, or solely to interindividual comparisons, or to both? I am aware of the fact that I am passing on a highly difficult problem to psychologists. On the other hand, it would be strange if the interpretation of a term so crucial as probability would be given by methodological considerations. If psychology constructs probabilistic laws, as has often been said in the philosophy of science (Hempel, 1962; Nagel, 1961), then it is up to psychology to decide in which sense they are probabilistic.

5.3.3 Experimental control and local homogeneity

An important point of difference between representational measurement on the one hand, and latent variable theory on the other, concerns the importance of experimentation. It is seems that the examples, that representationalism considers to be genuine instances of measurement, require quite a large degree of experimental control. Latent variable theory does not require such control; in fact, it does not even require that the latent variable position can be manipulated in principle.

Consider, for instance, the fundamental measurement account of length. The adequacy of this account hinges on the possibility to concatenate objects. It is paradigmatic for fundamental measurement that, if one has two objects a and b of unequal length, say $a \preceq b$, then it must always be possible to find a third object c so that the concatenation of a and c is not noticeably different (i.e., both $\preceq$ and $\succeq$ to) from b. This is a prediction of what would happen upon executing a special kind of experiment. Concatenation can thus be considered an experimental manipulation of the variable length, in which the additivity of length is tested. Of course, nobody would ever carry the experiment out, because it is obvious from the outset that length supports such experimental manipulations, and most people will have an overwhelmingly strong intuition that this experimental 'hypothesis' is true (although one could, strictly speaking, doubt it; Batitsky, 1998; Rozeboom, 1966-b).

Similar considerations are invoked in conjoint measurement. The idea of conjoint measurement is that one can experimentally vary the levels of both independent variables and assess their effect on the dependent variable. For a trade-off to be represented, it must not only be theoretically, but experimentally possible to find a change of levels in the first factor that can undo the effect (on the dependent variable) of a change levels level in the second factor. This means it is essential for the possibility of conjoint measurement that one has experimental control over the independent variables.

In contrast, even in the cases of latent variable theory that admit for an additive representation to be based on the true scores, such experimental control will not often be possible. For this would require not only the ability to induce changes in item difficulty (something that one could imagine to be relatively manageable), but also the ability to induce changes in the position on the latent variable. It is interesting to note that, if we were able to change a person's position on a latent variable in order to test the axioms of conjoint measurement, the resulting changes in true scores should comply with the model in order to sustain additive conjoint measurement. Thus, additive conjoint measurement presupposes that local homogeneity, as discussed in Chapter 3, holds.

It would seem, then, that the additive conjoint measurement model requires the validity of the very same within-subjects causal accounts that were argued, in Chapter 3, to be untenable for many situations where latent variable theory is applied. Moreover, not only does the additive conjoint model require that such accounts are true; it requires that we are actually able to induce changes in the latent variable to show that it is true. And only if we are able to induce these changes, as well as changes in the item difficulty, and are able to show that the so constructed trade-off is additive, could we say that we have a latent variable model that is truly a measurement model in the representationalist sense.

It cannot be doubted that, if one has the degree of experimental control that additive conjoint measurement requires, one has the strongest possible evidence for the validity of the testing procedure. For it would mean that one knows exactly how to manipulate the latent variable to bring about effects of any required size. It would mean that, if I brought to you Mr. Brown, and asked you to to change his position on the latent variable 'attitude towards the United Nations' by just the amount necessary to change his true score on the item 'Is your attitude towards the United Nations favorable' from .90 to .92, you would actually know what to do. This is, of course, very far removed from fitting a Rasch model. It also means that, if a representationalist would allow latent variable models into his conception of measurement, he would need to adhere to the same form of realism that has, in Chapter 3, been argued to be indispensable for a consistent interpretation of latent variable theory.

Thus, representationalism could be viewed as making experimental control a more or less defining feature of measurement. However, there seem to be many situations in which latent variable theory can be applied, where the required degree of experimental control is not only practically infeasible, but prohibited by the construct definitions. For instance, some conceptualizations of general intelligence hold that this is a stable, or even immutable, attribute, so that experimental control is structurally impossible. In such cases, the construct definitions resist a treatment of the construct in terms of conjoint measurement. In fact, one would have to say that, if this theory of intelligence were true, then it would be impossible to measure intelligence in the additive conjoint sense. This, however, would seem to me an argument against the generality of conjoint measurement, rather than an argument against the hypothesis that individual differences in general intelligence – if general intelligence exists – could be measured.

Now, in Chapter 3, the suspicion has been raised that most psychological constructs will not be of the locally homogeneous kind. If this is correct, then the additive conjoint account, when supplemented with the demand that experimental control be possible, would have to say that we cannot use testing procedures to measure interindividual differences on these constructs. Latent variable theory would, of course, still be applicable, because its statistical formulation does not include the local homogeneity hypothesis. Thus, the difference between the latent variable and representational models remains substantial, even though their formalizations may, in some cases, be very similar.

I do not know exactly where representationalism stands on this issue. It seems to me, however, unreasonable to bring experimental control into a general definition of measurement. In fact, this seems to put the horse behind the cart. In the context of length measurement, for instance, we can concatenate some – not all – objects because length is quantitative and supports additivity. I find it strange to turn this argument around, and to say that length is quantitative and supports additivity because we have enough experimental control to execute concatenation operations. Moreover, concatenation operations are possible not just because length is quantitative, but also because the type of objects which we would choose to concatenate have a large number of other physical properties that allow for such experiments. One such property is that we can imagine objects to be stable with respect to length ('rigid' as representationalism would say), and also pretty stupid so that they do not change their manifest behavior (which is doing nothing) as soon as they notice we concatenate them (something that people tend to do when they know they are being measured). This, however, is not just because length is quantitative and can be measured; it has to do with the physical structure of objects of a certain manageable length, and with the fact that we are not inclined to disagree with the multitude of silent assumptions made, like the assumption that rods are rigid. Thus, while the possibility to execute experiments like concatenation supports the claim that measurement is taking place, it cannot be taken to be a defining characteristic of measurement. If a psychologist is measuring individual differences on a locally irrelevant construct, he is not claiming that he can experimentally manipulate a person's position on this construct; in fact, he may be claiming the opposite. This does not invalidate the claim that he is measuring individual differences, for he has never claimed to have experimental control of the required type, nor has he formulated an assumption of this sort in the model. It is certainly the case that he dimension he is working with applies only to individual differences and to nothing more. But this does not imply that he cannot measure individual differences.

5.3.4 Validity and the relation of measurement

Because the theories discussed in this book entertain a radically different conception of what it means to measure something, one may expect them to give different accounts of what it means for a measurement procedure to be valid. In this respect, it is remarkable that influential treatises on validity, a concept deemed central to measurement, only superficially address theories of measurement, if at all. It seems to be tacitly assumed that it does not really matter whether one conceives of measurement from a true score perspective, a latent variables perspective, or a fundamental measurement theory perspective. As these theories conceive of the measurement process differently, however, it is likely that the semantics of validity that they give will differ. To investigate this matter, consider a simple sentence like 'IQ-tests measure intelligence'. Let us inquire what would make this sentence true in each of the theories discussed.

First, consider the measurement process from a classical test theory perspective. We have seen in Chapter 2, that classical test theory conceives of measurement in a statistical fashion. As Lord & Novick (1968, p. 20) put it, a test score is a measure of a theoretical construct if its expected value increases monotonically with that construct. At first sight, the theoretical construct could be taken to be the true score. Oddly enough, however, the true score is itself defined as the expected test score. Because true scores are identical to expected scores, and because any variable increases monotonically with itself, every test must measure its own true score perfectly. Therefore, if the true score on an IQ-test is considered to be identical to intelligence, the proposition 'IQ scores measure intelligence' is true by definition. This is because the proposition 'IQ-scores measure intelligence' is transformed to 'the expected IQ-scores are monotonically related to the true scores on the IQ-test' which is vacuously true since the true scores are identical to the expected scores. Because the line of reasoning succeeds for every conceivable test, in this interpretation every psychological test is valid. However, it is only valid for its own true score. This is the price of operationalism: If the construct is equated with the true score, each distinct test defines a distinct construct, because it defines a distinct true score.

An alternative interpretation of classical test theory is that the observed scores do not measure the true scores (after all, it is rather odd to say that an expected value measures itself), but that the true scores measure something else, in the sense that they are themselves monotonically related to the theoretical construct in question. Viewing the issue in this way, the sentence 'IQ-scores measure intelligence' is true if the true scores on the test are monotonically related to intelligence. From a classical test theory perspective, this means that the theoretical construct cannot be conceived of as represented in the measurement model for the test in question, but must be viewed as an external variable. This prompts the conceptualization of validity as correlation with a criterion variable, which yields the concept of criterion validity.

Criterion validity has been extremely important to the theoretical development of the validity concept, for the following reason. Originally, the criterion was considered to be an observed variable, such as grades in college. Because the validity question refers to measurement and not to prediction, and because IQ-scores do not attempt to measure college grades (which are, after all, observable) but intelligence, the criterion validity view was never an adequate conceptualization of test validity. One possible response to this is to sweep the criterion variable under the carpet of unobservability, and to grant it the status of a hypothetical entity. In such a view, the definition of validity in terms of a statistical relation (i.e., the true score increases monotonically with the theoretical construct) is typically retained. The measurability of the intended construct (intelligence) is thereby hypothesized a priori, and the validity of the measurements (IQ-scores) is conceptualized as a monotone relation of the true scores on the IQ-test with this hypothetically measurable attribute.

In this view, validity is external to the measurement model, because in classical test theory a theoretical construct such as intelligence cannot be non-vacuously represented inside the measurement model. The proposition 'IQ-scores measure intelligence' thus becomes 'the true IQ-scores increase monotonically with a hypothetical criterion variable called intelligence'. Attempts to find 'perfect' measurements of intelligence that could function as a standard, analogous to the standard meter in Paris, have, of course, proven fruitless. The type of thinking introduced by looking at intelligence as a criterion variable outside the measurement model is, however, still a very common way of thinking about test validity. That is, there is 'something out there', and the question of validity is how high the correlation between our test scores and that something is. This renders the semantics of validity dependent on two assumptions: 1) there really is something out there (intelligence), and 2) the test scores have a monotonically increasing relation with that something. If this is the case, then the proposition 'IQ-scores measure intelligence' is true. An interesting aspect of this view is that, because expected test scores will have monotonic relations with many attributes, any given test measures an indeterminate number of attributes. Thus, measures are not uniquely tied to a construct. If measurement is further reduced to correlation, everything measures everything else to a certain extent, and all tests must be valid. However, the requirement that true scores be monotonically related to the attribute to be measured is highly similar to the latent variable model; in fact, latent variable theory can be viewed as an elaboration of this idea.

The reason that classical test theory must consider theoretical constructs as external to the measurement model is that the syntactical machinery of the theory is not rich enough to represent constructs inside the model. As we have seen, the true score cannot perform this function without rendering a completely trivial account of measurement. Latent variable models do possess the required terminology. As has been discussed in Chapter 3, such models can be viewed as relating the true scores on a number of items or tests to a latent variable, or as relating subpopulation parameters to a latent variable. In either case, the latent variable must be considered to function as a representative for the theoretical construct (to be distinguished from the function of fundamental measurement scales, which are representations of observed relations). The relation of measurement in latent variable models is rather similar to the statistical formulation of classical test theory; namely, it is conceived of in terms of a stochastic relation that the observed scores have with the latent variable. However, these models do have the power to dispose of the problem that tests are valid for any attribute they are monotonically related to, because the dimensionality of the latent space can be specified in the model.

For example, in the unidimensional case, a latent variable model specifies that the true scores on each of a number of indicators are monotonically related to the same latent variable. Moreover, within such unidimensional models it is assumed that the indicators measure only this latent variable and nothing else. This implies that the indicators are independent, conditional on the latent variable. If, conditional on the latent variable, the indicators are still related to another variable (for example, group membership), the indicators are considered biased. Thus, if unidimensionality is posited, measurement can be seen as a monotonic relation of the expected scores with a latent variable, and only with this latent variable (in the sense that they do not systematically relate to another variable, given the latent variable). The proposition 'IQ-scores measure intelligence' then becomes 'the expected IQ-scores increase monotonically with the latent variable intelligence, and, given the latent variable, with nothing else'. It follows that the semantics of unidimensional latent variable models do not allow indicators to be valid for more than one latent variable, in contrast to the classical test model. Of course, this only holds for unidimensional models, and not for latent variable models in general.

In representationalism, measurement is a process of representing observed relations between subjects and items in a number system, which results in a measurement scale. This scale is a product of human activity: it is therefore not necessary to assume, a priori, that scales exist independently of the act of measurement, and that they are somehow responsible for the observed relations. This is in sharp contrast to latent variable models. Scales represent relations, they do not cause relations. Now, if observed relations can be represented in the number system (that is, if a homomorphism can be constructed), the resulting scale is an adequate representation by definition, and therefore measurement has succeeded. If the procedure fails, measurement has not taken place.

Let us consider our paradigm example, and interpret the proposition 'IQ-scores measure intelligence' from this perspective. In a strict interpretation, representationalism demands direct observability and experimental control with respect to the attribute in question. In this interpretation, IQ-tests cannot be considered valid for measuring intelligence; for neither the required relations, nor the experimental control over the attribute, have been shown to hold. Thus, from a representationalist perspective, IQ-tests in psychology cannot possibly measure intelligence, for they cannot be said to measure anything at all. The proposition 'IQ-scores measure intelligence' is thus false. Moreover, from a fundamental measurement perspective, measurement is extremely rare in psychology (if it occurs at all), because very few psychological tests produce the type of consistencies required for representational theory to operate. Thus, according to this definition of measurement, most or all psychological tests are invalid.

Still, this does not answer the question where representationalism would put the relation of validity; it merely says that psychological tests are invalid. I think that, if representationalists took the theoretical presuppositions of psychologists seriously, they would end up with a relation that is highly similar to, or in fact even the same as, the one posited in latent variable theory. The representationalist would first need to accommodate for the problem of error, that is, he would need to incorporate probabilistic relations. It has been argued in Chapter 4, and in the present chapter, that this will almost unavoidably lead to a latent variable model formulation. Second, he would need to step back from the requirement of experimental control. For it is ridiculous to demand such control if psychological theory itself holds that such control is not possible; therefore, representationalism would have to admit the possibility that constructs, which are irrelevant or heterogeneous at the level of the individual, may still be invoked in the measurement of interindividual differences – as long as the measurement relation is not misinterpreted as applying to within-subject dimensions. Of course, in the locally homogenous case, there is no problem at all, because experimental manipulations of the latent variable – if possible – would lead to changes that are in accordance with the hypothesized model, as the representationalist would require. Dropping the requirement of experimental control does not prohibit a causal interpretation of the relation between the attribute and its indicators; in fact, it would seem plausible for the representationalist to demand that such an interpretation holds. This requires the representationalist to abandon the empiricist position completely; for now he will have to hold that the attribute exists and has causal relevance for the observed variables. It thus seems that, if the representationalist gave up the empiricist foundation of the theory, incorporated a probabilistic relation between the attribute and the observed variables, and weakened the requirement of experimental control to the requirement that a causal relation should hold, he could occupy the same philosophical position with respect to the validity concept, as the latent variable theorist.

So, with respect to the relation of validity, we must conclude the following. Classical test theory does not formulate a serious account of measurement, and therefore is inadequate to deal with the question of validity. In fact, if it begins to formulate such an account, it invokes a kind of embryonic latent variable model. Latent variable theory is able, by its very conceptualization, to hold that measurement is a causal relation between the latent variable and its indicators. In fact, this is a natural interpretation of the theory, because it is virtually equivalent to a common cause model (Glymour, 2001). Representationalism works on observable relations between objects, and therefore has no place for the relation of validity: the very fact that we are supposed to be able to judge relations like 'not noticeably longer than' with the unaided eye, means that validity is presupposed in the model. However, upon closer inspection, representational measurement is strongly related to the requirement of experimental control; and this requirement cannot be considered to demand anything less than the possibility to intervene in a causal system. If the representationalist now drops the condition that relations between objects be 'noticeable', which is unrealistic in the first place, he turns out to have been hiding a latent variable model under the cloak of noticeability all this time. And if he reduces the demand for experimental control to the weaker demand that a causal relation between the attribute and its indicators hold, then he turns out to formulate virtually the same semantics of measurement as the latent variable theorist.

So, when we look upon these models in the context of validity, they converge to a surprising extent. As a prelude to the following chapter, I will now abstract what I think are good ideas from the different models. In my opinion, true score theory is wholly inadequate insofar as we are talking about measurement. It is a purely statistical theory on the behavior of (composite) random variables and, in the case of psychological testing, not a very plausible one. I think that latent variable theory has a reasonable philosophy of measurement. However, it places too much emphasis on technical requirements, such as unidimensionality. Unidimensionality is a good idea in itself because it has clear statistical implications, but I think that, in latent variable theory, it has come to occupy an unreasonably strong position. Strictly taken, unidimensionality is not a very realistic assumption to make when dealing with psychological test scores. The assumption appears moreover to be motivated by an, in itself understandable, desire to measure one thing at a time, rather than from a psychological theory that says why we should expect unidimensionality to hold in a particular testing situation. But the psychometrician's desires would not seem to be sufficient as a motivation for an assumption as strong as unidimensionality. The unidimensionality assumption could be motivated, however, by invoking a causal relation between variation on the latent variable and variation on its indicators. In this case, one says that unidimensionality will hold if the causal relation between the latent variable and its indicators holds, if this relation is correctly specified, and if the latent variable is the only attribute that causes variation on the indicators. Unidimensionality can then be considered a specific instantiation of the common cause idea, and local independence is one of its testable consequences. The kind of causal relation I am envisioning does not require local homogeneity, for I am taking the position that one can reasonably say that variation on an attribute causes variation on the observed scores, without the attribute being a causally efficient entity at the individual level.

Representational theory makes some very strong points, but, being deterministic, it is too restrictive. Moreover, requiring that we have full experimental control over the independent factors in additive conjoint measurement is too strict, because the possibility of experimental control depends on much more than a measurement relation. However, one may view the 'experiments' in representationalism as interventions in a causal system. Such interventions are not always possible, but if they are impossible this does not imply that the causal relation is false. Thus, one may reasonably weaken the requirement that experimental control be possible to the requirement that a causal relation must hold. Now if one does this, one is unavoidably drawn to a realist position with respect to the attribute in question. That is, if one is to say that the attribute does causal work in producing variation on the measurement outcomes, one cannot hold that it is constructed out of these very same measurement outcomes.

In conclusion, the two theoretical requirements that seem essential for validity are realism about the attribute in question, and a causal relation between variation on the attribute and variation on the measurement outcomes. This observation has serious consequences for the theory of validity. These consequences are the topic of the next chapter.

6. THE PROBLEM OF VALIDITY

6.1 Introduction

That the conceptual problems inherent in measurement in general, and psychological measurement in particular, are poorly understood is obvious from the lack of agreement on the meaning of the term 'measurement', the multitude of conceptually different models for implementing it (e.g., Lord & Novick, 1968; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Hambleton & Swaminathan, 1985; Krantz, Suppes, Luce, & Tversky, 1971), and the fact that no psychologist can point to a field where psychological measurement has succeeded without eliciting an immediate claim to the contrary from another psychologist. Given that virtually all aspects of the measurement problem are the subject of ongoing debates (Borsboom & Mellenbergh, 2002; Borsboom, Mellenbergh, & Van Heerden, *in press*; Lamiell, 1987; Lumsden, 1976; Maraun, 1999; Michell, 1986, 1999, 2000; Schmidt & Hunter, 1999), one would expect these debates to culminate in fierce discussions on the most central question one can ask about psychological measurement, which is the question of validity. It is therefore an extraordinary experience to find that, after proceeding up through the turmoil at every fundamental level of the measurement problem, one reaches this conceptually highest and presumably most difficult level only to find a tranquil surface of relatively widespread consensus (Kane, 2001; Shepard, 1993). In fact, this is not only surprising but slightly worrying, because validity is a largely philosophical topic. Consensus on philosophical problems is rare, and for good reasons: Philosophy is the art of critical thinking and critical thinking generally does not lead to consensus but to debate.

A second remarkable aspect of current validity theory is that the concept validity theorists are concerned with seems strangely divorced from the concept that most researchers have in mind when posing the question of validity. That is, most validity theorists have come to see the validity concept as embracing virtually every test-related problem that may be raised (Cronbach, 1988; Messick, 1989; Shepard, 1993), while many researchers are under the impression that the problem of validity simply concerns the question whether a test measures what it should measure. Moreover, if one regards this simple question as legitimate and crucial, as I do, then one has a very hard time understanding most recent papers on validity. To give but one example that I find puzzling: One can find in Messick (1989; p. 30) the idea that some psychological attributes are real, while others are not, and that in both cases the concept of validity applies. I have great difficulty in understanding the supposition that one can ask what may be called the 'simple' question of validity, which I construe as the question whether, for example, IQ-tests really measure the attribute we call 'intelligence', in a situation where there is nothing in reality that corresponds to intelligence. That is, if the realist position cannot be taken, I do not understand why the question of validity should apply at all.

Now, when encountering philosophical positions that seem elusive and difficult to understand, one always faces a problem of attribution. In general, there are three possible sources of confusion, and it is often hard to decide which is at play. The first and least attractive possibility is that the confusion arises from one's own limited cognitive resources. The second is that there is a flaw in the position itself. And the third is that the authors in question are analyzing the wrong problem. I think that, in the present case, the third explanation applies. It is my intent to convince the reader that most of the validity literature either fails to articulate the validity problem clearly, or misses the point entirely. I will argue that it is an unfortunate historical accident that the validity concept has been divided into different kinds, torn from its rightful place in science, and reunified by constructing it as an umbrella term intended to cover virtually every thinkable aspect of inference – be it scientific, philosophical, political, or ethical. Validity is not complex, faceted, or dependent on nomological networks. It is a very basic concept and was correctly formulated, for instance, by Kelley (1927, p. 14), when he stated that a test is valid if it measures what it purports to measure.

The argument to be presented is exceedingly simple; so simple, in fact, that it articulates an account of validity that may seem almost trivial. It is this. If something does not exist, then one cannot measure it. If it exists, but does not causally produce variations in the outcomes of the measurement procedure, then one is either measuring nothing at all or something different altogether. In these two cases, a test does not possess validity. In all other cases, it is valid. Thus, a test is valid for measuring an attribute if and only if a) the attribute exists, and b) variations in the attribute causally produce variations in the outcomes of the measurement procedure. Now, one may find this unsurprising. In fact, there is a good chance that many readers are inclined to respond that they tell their students this all the time. However, in the validity literature of the past two decades, it is difficult to find a explicit formulation resembling the above. In the writings of leading theorists (i.e., Cronbach, 1988; Kane, 2001; Messick, 1981, 1989, 1998; Shepard, 1993), one will not find much that sustains it; rather, one is likely to find this type of idea in a discussion of historical conceptions of validity (Kane, 2001, p. 319-323). The ontological part of Messick's (1989; 1998) unificationist conception of validity leans towards it, but it is not clearly articulated and it is questionable whether it is consistent with the other, epistemological, part of his synthesis (Markus, 1998). Moreover, Messick (1989, p.13) views validity as a judgment, and thus conceptualizes the term as applying primarily to the evaluation of evidence bearing on that judgment. Similar views are put forward in Kane (1992), Shepard (1993), and Moss (1992). In keeping with this idea, the current literature conceptualizes validity as applicable to test score interpretations only (Cronbach, 1988; Kane, 2001; Messick, 1989), while the conception stated here is consonant with the older conception that it is a property of tests. Indeed, finding similar conceptualizations requires browsing some of the older archives (e.g., Kelley, 1927; Cattell, 1946; Loevinger, 1957).

However, these treatises are not based on a causal, but on a correlational conception of validity, which I will argue is crucially mistaken. In latent variable theory, one may find causality based lines of reasoning (Bollen & Lennox, 1991; Bollen & Ting, 2000; Edwards & Bagozzi, 2000), but they will not be explicitly linked to validity theory. It thus seems that the validity concept, as formulated above, has not been explicitly proposed in the literature, although it certainly has been hinted at (Cattell, 1946; Loevinger, 1957; Campbell, 1960). An observation that underscores the apparent novelty of the stated conception is that it contradicts most of the conventional wisdom in validity theory, which means that this theory either does not sustain it, or is inconsistent, or both. For example, the above conception implies that a) validity is not a matter of degree, b) the square root of the reliability coefficient is not the upper limit of validity, and c) unreliability, item bias, and other supposedly undesirable characteristics of tests bear no direct relation to validity. This is in contradiction with every paper on validity I know, but I think it is correct.

The aim of the present chapter is to elaborate the implications of this view, and to discuss the ways in which it diverges from, or converges with, both historical and current conceptions. The argument will focus on four points where validity theory seems to have taken the wrong turn. First, it has confused ontological and epistemological claims; second, it has mistaken questions about reference for questions about meaning; third, it has been plagued by a correlational account where there should have been a causal account; and fourth, the idea that validity applies to test score interpretations, rather than to tests, is inadequate. Finally, it will be argued that current validity theory deals with too many issues at the same time, so that it collapses under its own weight. It is proposed that the question of validity must be taken to apply only to the question whether one is measuring the right attribute; not to the question how well one is measuring that attribute. This latter question is left to the technically oriented psychometric literature, which deals with it in a more sophisticated way than the validity literature.

6.2 Ontology versus epistemology

If the crucial issue in validity concerns the existence of an attribute that causally influences the outcome of the measurement procedure, then the central claim is ontological, and not epistemological. This is to say that one is claiming something about which things inhabit reality, and what they are doing there. Such claims are about ontology, and as such they are conceptually distinct from the ability to find out about reality, which is the central issue in epistemology. Measurement, of course, is the prototypical epistemological activity in science, and it is therefore easy to make the mistake that we are primarily claiming something on this front. This is because *if* the ontological claim holds, *then* the measurement procedure can be used to find out about the attributes to which it refers. Put more simply: If differences in intelligence cause differences in IQ-scores, then the IQ-score differences can be used to find out about the intelligence differences. Thus, in this very special case, the truth of the ontological claim guarantees the epistemological access.

It would seem, then, that to talk about the ontology is to talk about the epistemology, and there surely is a sense in which this is correct. Now it is a small step to conclude that, instead of laying down our ontological claims, which make so abundantly clear what kind of radical assumptions we are making (Borsboom, Mellenbergh, & Van Heerden, *in press*; Michell, 1999), we could just as well limit our discussion to the epistemological side of the endeavor, which is respectable and familiar. It is another small step to conclude that the question of validity is about particular aspects of this epistemological process we call measurement. The final step leading to some very dark philosophical dungeons from which escape is impossible, is to start talking about some presumed universal characteristics of this epistemological process (usually derived from a few paradigm cases like length or temperature measurement) that, if present, would allow one to somehow be rationally justified in concluding that the ontological claims are true.

This, of course, will not work. The family of procedures, that scientists – as opposed to philosophers – regard as instances of measurement, is diverse and incoherent and does not have universal characteristics. Length and temperature, blood pressure and brain size, pathology and intelligence all could be said to involve measurement, but the associated measurement practices are based on vastly different lines of reasoning and employ vastly different methodologies. So now one gets into trouble. What on earth could it be that this heterogeneous set of successful measurement procedures has in common? Is it the way the test looks? Representative sampling from a universe of behaviors? The line of reasoning on which it is constructed? The correlation between a test and some external variable called the 'criterion'? The (presumed) fact that the test figures in a 'nomological network' of 'constructs' ? Or is it just that we can do something 'useful' with regard to some 'purpose' which is presumably *different* from measuring the hypothesized attribute? Or are we on the wrong track here, because what is important is not a characteristic of tests or test scores, but of test score interpretations – which are, again, presumably *different* from the obvious ones like 'IQ-scores measure intelligence'?

This line of reasoning quickly gets us nowhere. The reason is that there *are* no universal characteristics of measurement, *except* the ontological claim involved. The only thing that all measurement procedures have in common is the either implicit or explicit assumption that there is an attribute out there that, somewhere in the long and complicated chain of events leading up to the measurement outcome, is playing a causal role in determining what values the measurements will take. This is not some complicated and obscure conception but a very, very simple idea. If we, however, fail to take it into account, we will end up with an exceedingly complex construction of superficial epistemological characteristics that are completely irrelevant to the validity issue. And because the measurement processes and models are diverse and complicated, we are likely to buy into the mistaken idea that the concept of validity must also be complicated. So now we get a multiplication of terms. For the human condition is such that someone will inevitably distinguish between 'kinds of validity' and 'degrees of validity' and so we are bound to come up with a hundred or so 'validities', which all come in 'degrees', until someone stands up because this is clearly ridiculous, and claims that 'all validation is one' (Cronbach, 1980, p.99) so that all kinds of validity can be integrated and subsumed under one giant umbrella (Messick, 1989). And since we are now thoroughly convinced that we are concerned with characteristics of an epistemological process rather than with an ontological claim, we are going to reach the conclusion that all this time we were really just talking about the one grand epistemological process – scientific research (Cronbach & Meehl, 1955; Loevinger, 1957; Messick, 1989). However, given that every attempt at drawing a line between 'scientific' and 'unscientific' research either fails or duplicates the distinction between good and bad research, we have now discovered the exciting fact that validation research is research. In other words, we have discovered nothing at all. And the reason for this is that there was nothing to be discovered in the first place.

When claiming that a test is valid, one is taking the ontological position that the attribute being measured exists and affects the outcome of the measurement procedure. This is probably one of the more serious scientific claims one can make, and it is difficult to prove or refute it. This, however, does not mean that the validity concept itself is complicated. Every test constructor in every scientific discipline has the stated line of reasoning in mind when she is constructing, administering, or interpreting a test. It is the only aspect that measurement procedures have in common. If one is going to search for homogeneity in the superficial characteristics of these procedures one is not going to find any, and one is likely to build ever more complicated systems covering different 'aspects' of validity. These systems, however, do not cover different aspects of validity but describe different research procedures for validation. So 'asking people what they think about the test' becomes 'face validity'; 'checking whether we can predict some interesting things with it' becomes 'predictive validity'; 'investigating whether the data fit our theory about the attribute' becomes 'construct validity'; and so on.

Turning verbs into nouns often leads to understandable classificatory systems but when doing philosophy one does well to stay clear of it. For the union of all possible test related activities of this kind is not validity, but validation. These terms are sometimes used interchangeably in the literature, but they are not the same. This is clear from the fact that validity is a property, while validation is an activity. In particular, validation is the kind of activity we undertake to find out whether a test has the property of validity. Validity is a concept like truth; it represents an ideal or desirable situation. Validation is more like theory testing; the muddling around in the data to find out which way to go. Validity is about ontology; validation is about epistemology. The two should not be confused. Now, I think that most of the validity literature has not dealt with the problem of validity, but with the problem of validation. While there is nothing wrong with describing, classifying, and evaluating validation strategies, such activities are not likely to elucidate the concept of validity itself. In fact, if one concentrates on the epistemological problems long enough, one will move away from the validity concept rather than towards it. Consider, for example, Messick's (1989) widely cited definition of validity: 'validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment' (p. 13; italics in the original). No view could be farther apart from the one being advanced here. Validity, in the present conception, is not a judgment at all. It is the property being judged.

6.3 Reference versus meaning

That the position taken here is so at variance with the existing conception in the literature is largely due to the fact that I have reversed the order of reasoning. Instead of focussing on the epistemological processes and trying to fit in existing test practices, I have started with ontological claims, and I derive the adequacy of epistemological practices only in virtue of their truth. This means that the central point in validity is one of *reference*: The attribute to which the psychologist refers must exist in reality, otherwise the test cannot possibly be valid. The position here is thus a strongly realist one, in that I construct measurement as involving realism about the measured attribute. This is because I cannot see how the sentences 'Test X measures the attitude towards nuclear energy' and 'Attitudes do not exist' can both be true. If you agree on this point, then you are in disagreement with some very powerful philosophical movements which have shaped validity theory to a large extent.

One particularly strong variant of these movements once proudly went by the name of logical positivism. Philosophers and scientists endorsing this theory saw it as their mission to exorcise all reference of theoretical terms (like 'attitude'), because such reference introduces metaphysics, which the logical positivists thought was bad. They therefore constructed theoretical terms as nonreferential. This lead them to focus on the *meaning* of theoretical terms. Meaning and reference are easily confused, but are very different concepts. To give a classic example (Frege, 1892), 'the morning star' and 'the evening star' have different meanings (namely 'the last star still to be seen at morning' and 'the first star to be seen at evening'), but refer to the same thing (namely the planet Venus). Because the positivists had a slightly phobic attitude towards metaphysics, they wanted to explain the use of theoretical terms like 'attitude' without letting these terms refer to reality.

This was an interesting endeavor but it failed (see Suppe, 1977, for a good overview). However, one of the relics of the approach has plagued validity theory to this day. This is the nomological network. A nomological network is a kind of system of laws relating the theoretical terms to each other and to the observations. For the positivists, this network served to create meaning without reference for the theoretical terms. The idea is that the meaning of a theoretical term is solely determined by the place of that term in the nomological network: the meaning of the term 'energy' is fixed by the network and by nothing else – certainly not by a reference to actual energy. Thus, in this view we can have meaning without reference, and can invoke theoretical terms without automatically engaging in ontological claims, which always introduce a lot of metaphysics.

This idea was used by Cronbach & Meehl in 1955 to put forward their idea of 'construct validity'. Many people think that construct validity is the same as the kind of validity being proposed here, but this is not the case. The construct validity position does not invoke reference (it does not say that the attribute to be measured should exist), and it does not talk about causality (it is not necessary for the attribute to have a causal role in determining the measurement outcomes). The classic position, as articulated by Cronbach & Meehl (1955), holds that a test can be considered valid for a construct, if the empirical relations, in which the test stands to other tests, match the theoretical relations, in which the construct stands to other constructs. One can imagine this as two path models, one hovering over the other. One model stands for theoretical relations, the other for empirical relations. If the models match, then there is 'construct validity' for test score interpretations in terms of the nomological network. For instance, suppose the nomological network says that the construct 'intelligence' is positively related to the construct 'general knowledge' and negatively to the construct 'criminal behavior'. Further suppose that one observes a correlation of .5 between an IQ-test and a test for general knowledge, and a correlation of -.4 between the IQ-test and the number of months spent in prison. There is thus a match between empirical and theoretical relations. In construct validity theory, it is this match that constitutes and defines the validity concept.

To define construct validity, no reference to the existence of theoretical entities is necessary, and their causal impact on the measurement outcomes is not even a topic of discussion. Read Cronbach & Meehl (1955) to see how carefully they avoid this issue. As an illustration of the ambiguity of Cronbach & Meehl's (1955) paper, one may confer Bechtold (1959) and Loevinger (1957), who both discuss construct validity, but are talking about two completely different interpretations of the concept – one positivist, the other realist. In principle, however, within the construct validity perspective there is no friction between 'Test X measures the attitude towards nuclear energy' and 'Attitudes do not exist'. As long as the empirically observed relations, between test X and other tests, match the theoretical relations in the nomological network, all is fine. So, this view has a little bit in it for everyone.

The problem, of course, is that we have few if any nomological networks in psychology that are sufficiently detailed to do the job of fixing the meaning of theoretical terms. To fix this meaning requires a very restrictive nomological network. The reason is that the theory that has to be invoked for construct validity to work is an account similar to the descriptive theory of meaning (Kripke, 1972). This theory does not say 'intelligence is a real attribute with causal impact on our measurements', but 'intelligence is whatever has the relations to other constructs as specified in the nomological network'. Cronbach & Meehl (1955) do not mention the descriptive theory of meaning, but that they rely upon it is evident from statements like 'a construct is defined implicitly by a network of associations or propositions in which it occurs' (p. 299-300). It is crucial for the ideas formulated in Cronbach & Meehl (1955) that a descriptive account of meaning is possible, because otherwise one is forced to invoke a reference for intelligence, which brings in the very metaphysics to be avoided through the back door.

In some highly developed theories, like the ones in physics, one could at least begin to consider this account, because they are restrictive enough to single out one particular theoretical term, which is the only one that has all the right relations. In psychology, such an account does not work because we do not have the required theories. That this is not just an academic point, but a decisive argument against using a descriptive theory of meaning can be immediately seen by considering the intelligence example discussed before. One does not get anywhere by saying that 'intelligence is whatever is positively related to general knowledge and negatively to criminal behavior', because there are too many theoretical terms that will satisfy this description, and many of them will evidently not be the same as intelligence. No theoretical term in psychology can be unambiguously identified in this way. Thus, this theory will not be able to single out theoretical terms by merely describing where they stand in a nomological network. Cronbach & Meehl (1955) do discuss the problem that nomological networks are incomplete and vague in psychology, but they do not mention the most important implication of that problem: It is fatal to any positivist reading of their account, because it shows that reference, and the accompanying realist metaphysics of measurement, cannot be avoided. Instead, they conclude that it leads to vagueness in the construct definitions. This is, of course, true, but not the primary problem. The primary problem is that too many theoretical terms will satisfy construct definitions of the kind Cronbach & Meehl (1955) are discussing (see also Rozeboom, 1960), and that therefore the theory of meaning they use fails to work.

Now, this should not be regarded as a grave problem for psychology in general, because the descriptive theory of meaning is not a very good one anyway (Kripke, 1972). Neither is there any particular problem about not having nomological networks, because one has to start somewhere – and the tight, lawlike relations that make up nomological networks are more likely to be the result of research than a prerequisite for doing it. However, one would expect psychologists to dismiss any account of validity, that requires the existence of nomological networks, as inadequate from the outset because when one thinks the matter through, such a theory would have very undesirable consequences. For example, some psychological tests certainly appear to be measuring something important, and one should be able to say that one thinks, suspects, or hypothesizes that IQ-tests validly measure intelligence even if one momentarily has no nomological network available to fix the meaning of the term 'intelligence'. In a theory of validity that requires the availability of nomological networks this is, strictly taken, impossible. Every psychologist should object to this; not only because the view is unduly restrictive but because it is completely inadequate. Validity does not depend, and has never depended, on the availability of nomological networks because it is not about meaning but about reference.

In this context, it has been noted by validity theorists (Shepard, 1997; Kane, 2001), that requiring the existence of a nomological network is unrealistic in psychology. However, if one removes the nomological network from construct validity theory, one is left with very little indeed. In fact, dropping the nomological network leaves one without the heavily needed theory of meaning, and one is likely to be forced to introduce reference again, that is, to interpret the theoretical terms as referring to things out there in the world. I think that this is a plausible move, as will be evident, but the consequence is that the main idea of construct validity, as put forward by Cronbach & Meehl (1955), loses its bite. That is, if one reintroduces reference, then it is difficult to maintain that what constitutes validity is a match between empirical relations and theoretical relations. For this match is now rendered a helpful epistemological criterion, which may be given a signalling function, but not much more. Thus, if there is a grave discrepancy between the theoretical and empirical relations, one knows that something is wrong somewhere; but this can hardly be considered news. If the theoretical and empirical relations match, this match does nothing more than corroborate the theory, to use a Popperian term. The match is no longer constitutive of validity, however, because the reintroduction of the realist metaphysics forces one to shift back to reference as the primary defining feature of validity.

The emphasis that is placed on the importance of ruling out alternative rival hypotheses for corroborating data (Cronbach & Meehl, 1955; Messick, 1989) partly acknowledges this. One can readily see this by introducing the question to what hypothesis the alternative one should be considered a rival. Obviously, to the hypothesis that there is an attribute in reality that produces variation in the measurement outcomes. What, then, is to be seen as the defining feature of validity if not exactly the truth of that hypothesis? And if this is correct, then where does this leave the instrumentalist, positivist, and empiricist? Consider, for example, instrumentalism. This view does not invoke truth, but usefulness as the primary criterion for the adequacy of scientific theories and measurements. However, we are surely not seriously considering the idea that we have to rule out rivals to the hypothesis that intelligence tests are useful. The Wechsler Adult Intelligence Scale comes in a big heavy box, which is very useful to hit people on the head with, but the hypothesis that the WAIS is valid for inflicting physical injury is certainly not the kind of hypothesis we are interested in. Clearly, from the viewpoint of ruling out alternative hypotheses, the hypothesis that the test is useful is neither intended nor relevant, except for the very special hypothesis that it can be used to measure intelligence *because intelligence produces variations in IQ-scores.*

In conclusion, a positivist or instrumentalist reading of construct validity requires a descriptive theory of meaning which must invoke nomological networks. Cronbach & Meehl (1955) tried to construct an account of validity on this basis. However, the nomological network interpretation of construct validity is inadequate, as has been recognized in the literature. Dropping the nomological network from consideration simply means that one has to go back to a realist interpretation of psychological attributes. In a realist interpretation, however, the crucial issue is reference and not meaning. Therefore, a question like 'are IQ-tests valid for intelligence?' can only be posed under the prior assumption that there does exist, in reality, an attribute that we designate when we use the term 'intelligence'; and the question of validity concerns the question whether we have succeeded in constructing a test that is sensitive to variations in that attribute.

6.4 Causality versus correlation

Although construct validity theory is, in its original form, inadequate, it does represent a serious attempt to forge a validity concept that has an account of meaning, a function for theory, and that stresses the fact that there is no essential difference between validation research and research in general. Moreover, if one removes the nomological network from consideration, replaces meaning with reference, and reintroduces the realist perspective, much of what is said in construct validity theory remains consistent and plausible. Also, the idea of construct validity was introduced to get rid of the atheoretical, empiricist idea of criterion validity, which is a respectable undertaking because criterion validity was truly one of the most serious mistakes ever made in the theory of psychological measurement. The idea, that validity consists in the correlation between a test and a criterion, has obstructed a great deal of understanding and continues to do so. The concept continues to exert such a pervasive influence on the thinking of psychologists, because many are under the impression that construct validity is really criterion validity with the criterion replaced by the construct (this fallacy cannot be attributed to construct validity theorists, as is evident from the writings of Cronbach & Meehl, 1955; Kane, 2001; and Messick, 1981, 1989). However, the inadequacy of this view does not depend on whether one views the criterion as a variable to be predicted from test scores, or as an 'infallible' measure of the theoretical construct to be measured, or as the theoretical construct itself. The crucial mistake is the view that validity is about correlation. Validity concerns measurement, and measurement has a clear direction. The direction goes from the world to our instruments. It is very difficult not to construct this relation as causal. Criterion validity employs correlation and similarity, where it should employ direction and causality.

Of course, causality is a laden term, and many researchers seem afraid to use it. The platitude 'correlation is not causation' is deeply inscribed in the conscience of every researcher in psychology, and in the literature the word 'causes' is often replaced by euphemisms like 'determines', or 'affects', or 'influences'; in measurement, we see traits 'manifesting' or 'expressing' themselves. What is meant is that traits cause observed scores. It is perfectly all right to say this because hypothesizing a causal account does not mean that one interprets every correlation as a causal relation. This, again, is the epistemological side of the issue which remains as problematic as ever – although progress has been made in this respect, as is evidenced in the work of writers like Pearl (2000) as well as in the development of latent variable models. The primary power of causality lies in the theoretical opportunity to think directionally rather than in terms of similarity or correlation (see, for some good examples, Pearl, 2000; Glymour, 2001). Now, I insist that measurement is a causal concept, not a correlational one, and that validity is so too. To clarify this, it is useful to point out some absurdities to which any theory based on a correlational account of validity leads. The criticisms must be explicitly understood as applying not just to the criterion validity view, but to any view that does not invoke a causal arrow pointing from the attribute to the measurement outcomes.

First, it has been observed by Guilford (1946) that the idea of criterion validity leads to the conclusion that a test is valid for measuring many things, as epitomized in his famous statement that a test is valid for anything with which it correlates. However, it can be shown that the set of zero correlations is a null set, which means that the likelihood of encountering a zero correlation in real life is exceedingly small (Meehl, 1978), and it has also been observed that in the social sciences everything tends to correlate with everything. Therefore, the upshot of any line of thinking that sees correlation as a defining feature of validity is that everything is, to some degree, valid for everything else. This absurdity does not arise in a causal theory because it is not the case that everything causes everything else.

Second, the idea has the unfortunate consequence of introducing degrees of validity: The higher the correlation, the higher the validity. The limiting case is the case where two variables correlate perfectly, which would imply perfect validity. That is, if one views validity as correlational, one is bound to say that if two constructs have a perfect correlation, then 'they are really the same construct under two different labels' (Schmidt & Hunter, 1999, p.190). This is very problematic. For instance, suppose one is measuring the loudness of thunder. The readings will probably show a perfect correlation with the simultaneously measured intensity of lightning. The reason, of course, is that both are the result of the distance between one's position and the location of the electrical discharge in the clouds, and of the severity of the discharge. However, the loudness of thunder and the intensity of lightning are not the same thing under a different label. They are strongly related quantities, one can be used to find out about the other, and there is a good basis for prediction, but they are not the same thing. When one is validly measuring the loudness of thunder, one is not validly measuring the intensity of lightning for the simple reason that one is not measuring the intensity of lightning at all. The limiting case of the correlational view implies that perfect correlation is perfect validity, and this leads to the idea that deterministically related quantities are the same thing. This absurdity does not arise in a causal theory because variations in the intensity of lightning do not play a causal role in producing variations in the loudness of thunder.

Third, the correlation is a population dependent statistic, that is, it is sensitive to the amount of variability in the attribute to be measured across populations. A well known instance is the attenuating effect of restriction of range in the presence of imperfect relationships between variables. Any correlational view must therefore hold that validity itself is by necessity variable over populations. Corrections for unreliability and restriction of range (Lord & Novick, 1968) are going to solve some of the trouble here but not all of it. In particular, there is one important, well-established case of valid measurement where the population dependence of correlations raises serious problems. This is the case of extensive measurement, as discussed in Chapter 4 (Campbell, 1920; Krantz, Luce, Suppes, & Tversky, 1971). This is very troubling because extensive measurement is more or less the paradigm example of measurement in general (Narens & Luce, 1986). In extensive measurement, attributes are not defined solely with respect to individual differences between objects (as is the case in almost all instances of psychological measurement), but with respect to an empirical concatenation operation. In this case, it can be meaningful to say that one is measuring one individual object (which is meaningless with interindividual difference variables). Now suppose we are measuring the length of rods, and that the measurement apparatus used is a meter stick. Further suppose that we are measuring without error. The correlation between the measurement outcome and the real length will be unity in most populations, as it should be, but there is an important class of populations where it will be zero. This is the population of rods of equal length. Therefore, we must conclude that, in this population, the centimeter is not valid for measuring length. This is a strange result. In extensive measurement, it is quite meaningful to say that all objects in such a subpopulation are, say, 4.2 feet long, and that this measurement is valid. In the causal account, this absurdity does not arise. This because causality is directional and conditional: The causal account says that, *if* there are differences in the attribute, *then* these will produce differences in the measurement outcome. However, if there are no differences in the attribute, no differences in the measurement outcomes are expected. This in no way precludes the validity of the measurement outcomes themselves, which is exactly as it should be.

Correlations are epistemologically relevant because they are sometimes indicative of causality, but they are not, and cannot be, constitutive of validity. I have dealt with the refutation of this view in somewhat greater detail than is perhaps necessary, as criterion validity has been considered inadequate at least since Cronbach & Meehl's (1955) introduction of construct validity (Messick, 1989; Kane, 2001). A thorough refutation seemed important, however, because I am under the impression that many people, who do not subscribe to the criterion validity perspective, still have a correlational conception of validity – the only difference is that they have replaced the criterion with the construct itself. I propose that if attribute differences do not play a causal role in producing differences in measurement outcomes, then the measurement procedure is invalid for the attribute in question. Correlations are not enough, no matter what their size. Height and weight correlate about .80 in the general population, but this does not mean that the process of letting people stand on a scale and reading off their weight gives you valid measurements of their height. To state otherwise is to abuse both the concepts of measurement and of validity. In fact, I consider the very fact that a correlational view of measurement allows for this kind of language abuse as a fundamental weakness; and I suggest that any theory of validity that sustains such absurdities should immediately be dropped from consideration. I hope I have convinced the reader that not just criterion validity, but *any* correlational conception of validity is hopeless.

The causal view of validity is clearly very powerful in comparison to the correlational one. I have not been able to find any implications of it that are remotely near the aberrant behavior of the correlational conception. However, conceptual power always comes at a price, and this price is usually paid in metaphysical currency. That is, I have put causality to work, but this comes at the cost of introducing a heavy assumption into the proposed conception of measurement. Is it plausible that this workhorse will ride in psychology? In the present context, the main danger is that the causal account may seem to be just too strict for psychological measurement. There are measurement experts as well as psychologists who are under the impression that any causal account of psychological measurement is untenable. In particular, I anticipate the following argument.

It may seem that I am proposing that, for instance, John's intelligence causes his IQ-score. This would require me to introduce an (at best) dispositional attribute as a cause. It is important to make clear that this argument does not apply. For if it did, it would not just be problematic but fatal to my position. However, the argument confuses the two distinct kinds of causal statements that have been discussed at length in Chapter 3. Specifically, the confusion arises from the singular use of the term 'intelligence'. This usage seems to imply that, for a particular subject, intelligence plays a causal role in producing test scores. This would indeed be a flawed account because intelligence is not the kind of variable that can be unproblematically introduced as a process variable. Intelligence is an interindividual difference variable, and as such it does not apply at the level of the individual. However, we do not have to assume that intelligence works at this level, because IQ-scores are not intended to measure intelligence in this way. Rather, *differences* in IQ-scores are intended to measure the effect of *differences* in intelligence. And in this sense, the causal account surely can be set up.

Thus, we do not have to suppose that intelligence causes IQ-scores in order to claim validity; we merely have to suppose that differences in intelligence cause differences in IQ-scores, which is a much weaker claim. This claim is not refuted by an argument against the use of dispositions as causes. We may remain silent on what happens at the individual level; in fact, we do better to refrain from introducing intelligence as a cause there. Interpreted strictly in terms of differences, I do not think the causal link proposed here is untenable. In fact, it does not seem to be all that controversial. The bold statement 'intelligence exists' will give rise to extended discussions among measurement experts, intelligence researchers, and at birthday parties. However, the statement 'differences in intelligence exist' is unlikely to elicit more than a faint smile. Similarly, to claim that your intelligence causes your IQ-scores will elicit your denial, and rightly so. But is it really so extraordinary to suppose that differences in IQ-scores are causally determined by differences in intelligence, especially when one considers that such a proposal does not presuppose that intelligence differences are the only cause at work in producing the IQ-score differences?

I think that such claims are not extraordinary at all, and that most measurement practices proceed along just this line of reasoning. In fact, as I have said before, the introduction of a causal line of reasoning is probably one of the few universals in measurement. Of course, latent variable models (Hambleton & Swaminathan, 1985; Bollen, 1989, 2002) explicitly incorporate this idea, because they can be viewed as common cause models (Reichenbach, 1956; Glymour, 2001). It may be less obvious that other approaches also take the causal stance, be it in a more indirect manner. For example, generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) is certainly amenable to this analysis. While the idea of tests as samples from a universe of behaviors is not an explicitly causal one, generalizability theory surely assumes that differences in universe scores lead to differences in domain scores; at that level, it is not at all difficult to introduce a causal relation between the two concepts. Similar accounts can be set up for most measurement practices and models.

The only model that truly does not seem amenable to this analysis is the formative model discussed in Chapter 3 (Bollen & Lennox, 1991; Edwards & Bagozzi, 2000), because, in that model, the values of the attribute are determined by the indicators rather than the other way around. However, I do not see this as a problem because the question, whether the weighted sum of the indicators 'salary', 'quality of neighbourhood', and 'educational level' yields a 'valid' measurement of SES, seems rather contrived in the first place. It may therefore be a good idea to call the formative model an instance of indexing rather than of measurement. Whether one has the right indicators for an index variable would seem a matter of convention and usefulness rather than of validity. This does not preclude that there may be sound arguments for including some indicators and not others; but I think that these considerations do not bear upon the question of validity. Of course, the question of validity may be raised at the level of the indicators (when one asks, for example, whether the question 'what is your annual income?' is a valid measure of annual income); but it does not apply at the level of SES because SES is not, properly speaking, measured but constructed.

In conclusion, the causal conception of validity avoids the absurdities of a correlational perspective because it is directional and conditional. So is measurement. And although it is true that the statement, that an attribute like intelligence causes the IQ-scores for an individual subject, is either meaningless or false, this argument does not pose a serious threat to the causal interpretation at hand. Like perception, measurement is about detecting variations, contrasts, and differences; and the only statement we need to make to claim validity is that the measurement instrument will detect the relevant variations, i.e., that variations in the attribute will cause variations in the test scores. Some reflection shows that this line of reasoning is not all that extraordinary, but underlies most or all measurement procedures; and it is certainly the case that most measurement models allow for this interpretation. The only model that does not sit well with this interpretation is the formative model. However, that this model is excluded from consideration in a validity context does not seem to be a problem for, but rather a virtue of, the present conception. This is because the formative model is, in my view, not a model for measurement but for indexing. It seems to me that a causal interpretation of validity is reasonable, and I propose it be considered in the literature.

6.5 Tests versus interpretations

Test theory abounds with unlucky terminology; some of the more infamous examples are 'true scores' (Lord & Novick, 1968; Borsboom & Mellenbergh, 2002), 'admissible transformations' (Stevens, 1946; Lord, 1953), 'meaningfulness' (Suppes & Zinnes, 1963; Michell, 1986), and 'reliability' (Lord & Novick, 1968; Lumsden, 1976; Mellenbergh, 1996). All of these terms indicate important concepts, but in every case the label is awkward, because it suggests an unintended meaning. It almost seems as if the one thing that measurement theorists have in common is the curious ability to flawlessly pick the one name for a concept that will guarantee its misinterpretation. And as is the case for most concepts in test theory, the term 'validity' is not very well chosen.

In particular, 'valid' is an adjective that may naturally be applied to arguments, statements, theories, and judgments, but to apply it to nonlinguistic entities like tests is to stretch the grammatical limits of natural language. Tests are instruments, they serve a purpose and may be useful in this respect, but to say that they are valid seems just as absurd as to say that they are true. Tests do not purport to measure anything, and neither do scores. It is us, the investigators, who desire to measure attributes; it is us who interpret the scores; and it must be us who present the validity argument. Therefore, it can only be the test score interpretation (Cronbach & Meehl, 1955; Messick, 1989), or else the argument that justifies the interpretation (Kane, 1992; 2001), to which the adjective 'valid' may apply – but not the tests themselves. So the argument of modern validity theory goes.

Although most of the steps taken in the above argument are based on plausible ideas, I do not find the argument very convincing on the whole. It is certainly true that to apply the term 'validity' to tests is not to use the term in a natural manner, but terminology is just a matter of convention and I cannot help it that this particular terminology has been introduced. It would be a good idea to change it, but history teaches that any attempt to change a term so deeply entrenched as 'validity' is guaranteed to fail (Ebel, 1956, is an example in this context). It is, however, my conviction that the old fashioned way of using the adjective 'valid', which is to apply it to tests and not to test score interpretations, singles out a very important property of tests. And it is exactly this property that provides such a useful vehicle for saying that we have succeeded in measuring what we set out to measure. To ascribe to the test this property is meaningful and I am willing to defend this practice. The validity literature has taken the other option and has deserted this usage (and, in doing so, left almost the entire research community behind). It is now standard for a validity theorist to say that validity applies only to test score interpretations, with the possible extension to actions based on test scores (Messick, 1989).

There are several reasons why this conception is not optimal. First, one can construct cases where a test score interpretation is valid but the test evidently is not. I have, for example, developed a new test. It is called the number test and contains one question. The question is 'write down a number between 70 and 130'. I have the following interpretation for the scores: 'the test scores resulting from administering the number test do not measure any psychological attribute whatsoever'. I have done a number of studies that provide strong evidence for the validity of this interpretation. For example, it turns out that the test scores do not correlate with height, IQ, and extraversion. It seems to me that nobody can reasonably dispute that this interpretation of the scores on the number test is valid, and that the evidence strongly supports this conclusion. It also seems to me that nobody can reasonably dispute that the *test* is invalid; for it does not measure any interesting attribute whatsoever. It is thus sensible to say that the interpretation 'the number test is invalid' is itself valid. This establishes that the validity of tests and the validity of test score interpretations are quite distinct topics. I think that the validity of tests is the topic of interest in psychological measurement.

One may reply that this is an unfair example, because Cronbach & Meehl (1955), Messick (1989), and others meant their definitions to apply to a limited set of test score interpretations, and not to all such interpretations. But how are we to determine which interpretations are eligible for consideration? I think that some reflection will show that there is basically just one class of interpretations in which we are interested. These are interpretations of the form 'test X measures attribute Y'. Maybe I am missing something here, but it seems to me that this gets us back to square one with a vengeance. For is not the only condition, that unambiguously sustains the validity of the interpretation 'IQ-tests measure intelligence', the condition that the proposition expressing this interpretation is true? But if this is the case, then the term 'validity', as applied to test score interpretations, turns out to do nothing more than the concept of truth was already doing (Borsboom, Van Heerden, & Mellenbergh, *in press*). That is, what prohibits us from saying that IQ-tests are valid for measuring intelligence if and only if the proposition 'IQ-tests measure intelligence' is true? And if I am correct on this score, then where does the concept of validity come in at the level of test score interpretations?

The answer is that validity, as applied to interpretations, must be introduced to deal with the epistemological side of the question, and in fact cannot be introduced at any other level. We cannot know whether a test score interpretation is true any more than we can know whether quantumtheory is true, for we have no conclusive method of verification. However, what we can do is to evaluate the evidence and theory supporting the interpretation at hand. This is why Messick (1989), Kane (2001), and other theorists see validity as an evaluative judgment, and not as a property of tests. I have no quarrel with the importance of the evidential, theoretical, and consequential issues involved here; but I seriously doubt whether we need to address epistemological issues when defining and delineating the validity concept. In this context, it is important to realize that epistemology is in a quite hopeless state at the present time: The stage is crowded with philosophers subscribing to relativism (Meiland, 1977), scientific realism (Devitt, 1991), constructive empiricism (Van Fraassen, 1980), falsificationism (Popper, 1959; Lakatos, 1978), social constructivism (Latour, 1987), postmodernism (Foucault, 1970), and to smaller movements based on bootstrapping methodology (Glymour, 1980), logical reliability (Kelly, 1996), problem solving (Laudan, 1977), computational approaches (Thagard, 1988), and game theory (Hintikka, 2001). No epistemological criteria for truth or validity are accepted by more than a handful of philosophers of science, and legitimate doubts can be raised as to whether such criteria can be found at all. Therefore, a definition of validity in terms of the quality of the argument put forward, or in terms of the evidence adduced, is unlikely to provide a firm foundation for the concept. Moreover, once we turn to the epistemological side of the problem, we are in no position to claim that we are specifically discussing it with respect to psychological measurement. We are discussing the validity of interpretations in general. So, this is rapidly becoming a very ambitious project. For now we will have to find our way through the epistemological labyrinth that generations of philosophers have so carefully crafted – and given the nature of philosophers, the exit is likely to be missing. Do we really want to go there?

I think that no such expedition is called for. In fact, I submit that the focus on test score interpretations, as opposed to tests, is yet another instance of the emphasis that validity theory has come to place on epistemology, where it should be concerned with ontology. An ontological framework that unambiguously defines what it means for a test to be valid need not be so complicated. I think that I am putting forward an adequate, and yet simple, proposal in this very chapter. The property of tests that we indicate with the admittedly infelicitous term 'validity' is just that variations in the attribute to be measured produce variations in the test scores. This is most certainly not a property of test scores or of test score interpretations. It is the test that does the job here, and it is the test to which the property of validity should be assigned. The test functions as a gateway from the world to us, and if this gateway happens to convey the effect of variations in an intended attribute (like intelligence), and blocks the influence of many other attributes (like height, weight, extraversion, and so on), then *that* is the property we are interested in. I do not deny that the interpretation of test scores in theoretical terms is important, and there is a nontrivial sense in which modern validity theorists are justified in placing the emphasis here – namely, when one approaches the issue from an epistemological point of view. The argument put forward to convince other researchers that the test has the property of validity (Kane, 1992; 2001) is also important, and worthy of study in its own right. Likewise, the justification for using test scores in selection or placement is important; such justifications should be thoroughly scrutinized, and not just on scientific grounds (Messick, 1989). But not every *important* aspect of tests and test use is *relevant* to the validity concept. In fact, I think that most of the issues on which validity theory has focussed in recent years are not directly relevant to the concept of validity. Many of these concerns follow directly from shifting the emphasis from tests to test score interpretations. I think that this move is both unnecessary and inadequate. Validity should therefore be reconceptualized as a property of tests.

6.6 Simplicity versus completeness

The development of validity theory in the course of the 20th century shows a consistent movement towards a greater scope for the concept. The original formulations were more or less technical in nature, stressing primarily the size of validity coefficients. However, at least since the work of Cronbach & Meehl (1955), who made the concept depend on nomological networks, validity theory has aimed at completeness. I think that the concept should be kept as simple as possible. Validity is a central concept in psychological testing, but not in the sense that it embraces and incorporates every important consideration in test use.

The most elaborate attempt to present a complete theory of validity that covers every aspect of tests is the treatise by Messick (1989). This theory is much too big to warrant a fair discussion here, but I would like to highlight some of the major difficulties I see in this kind of unified validity concept. In particular, I want to consider briefly Messick's progressive reading of his famous faceted conception of validity (Messick, 1989; p. 20-21). This reading suggests that scientific evidence, predictive utility, value implications, and social consequences add up to form a kind of total validity. In Messick's words, 'evidence of the relevance and utility of test scores in specific applied settings, and evaluation of the social consequences of test use as well as of the value implications of test interpretation, all contribute in important ways to the construct validity of score meaning' (Messick, 1989, p. 21). This seems to imply that, say, the use of IQ-scores in personnel selection could be approached through a maximization of the evidential basis, predictive utility, and social benefits of such use. While there may be cases where such maximizations are possible in the additive sense suggested by Messick, in many cases we will be faced with a trade-off instead. To make this point clear, I briefly discuss two cases where the interests represented by the different cells of Messick's (1989) matrix will juxtapose each other, rather than align.

Case 1: Measurement invariance, prediction invariance, and fairness

First, consider the measurement versus prediction invariance paradox as discussed by Millsap (1997). The problem concerns the fact that equal regression lines across groups of, say, job performance on IQ, are generally inconsistent with the hypothesis that the same factor model underlies the IQ-scores in each group. This becomes especially problematic if groups differ in the variances of the latent variable (which will be the rule rather than the exception). In this case, if we have measurement invariance (no bias) across groups, then we will have unequal prediction slopes. If we have equal prediction slopes, however, we will have a test that violates the requirements of measurement invariance. Now suppose we are involved in a selection problem, say, we need to select people for access to an educational program in some medical specialization. Knowing the excellent predictive properties of IQ-tests, we decide to base the selection procedure on IQ-scores. Further suppose that we are selecting people from different ethnic groups. Obviously, we want the best people for the job; and at least from a public health perspective, this is surely in the best interest of society. This requires maximizing predictive utility. We also want to have a good scientific basis, which requires factorial invariance. Finally, we would like a fair and transparent selection procedure, which implies a rule like 'whatever ethnic group you belong to, if your IQ-score exceeds 120, then you are in'. Now, what strategy should we follow to jointly maximize construct validity, predictive utility, and fairness?

It is immediately obvious that no such strategy exists. If we, for example, maximize construct validity, then we ought to use an unbiased measurement instrument, i.e., a test that shows factorial invariance over ethnic groups. If we also want to maximize predictive utility, then we ought to allow the predictive regression lines to differ over ethnic groups. However, if we now set a desired standard on the criterion (job performance), we will have to set a different cut-off score in each of the groups. But this would imply that a person from ethnic group A is selected if her score exceeds, say, 115 on the IQ-test, while a person from ethnic group B is selected if her score exceeds 120. This does not seem fair to us, and it certainly would not seem fair to the general public (in whose interest we are doing this – remember that the goal is to get the best people for the job). It would take a very good psychometrician to explain this practice in court. On the other hand, if we select an instrument with equal regression slopes, we can jointly maximize predictive utility and fairness, but construct validity would have to suffer, because we would have to select a test that violates factorial invariance (i.e., a biased test). Finally, if we jointly optimize construct validity and fairness, by using a factorially invariant test but setting the same cut-off score for all applicants, we will have to give in on the predictive front, and our selection will not be optimal; in this case, we would not get the best doctors. Clearly, then, construct validity, predictive utility, and social consequences do not 'contribute' to the overall validity of the selection procedure, except for the fact that their interplay presents us with a difficult moral dilemma.

Case 2: Homogeneity in measurement is multicollinearity in prediction

The conjunction of scientific concerns and issues of social consequences will give rise to paradoxical situations and moral dilemma's. This is because the goals of science and society often conflict. However, conflicting interests can also arise when no social consequences are in play. An interesting case concerns the problem of choosing between optimizing measurement or predictive properties of tests. It has often been suggested that these go hand in hand, and that optimizing validity will pay itself back through improved prediction. There are actually strong reasons to suspect that the opposite is the case (Lord & Novick, 1968, p. 332; Smits, Mellenbergh, & Vorst, 2002). Optimizing measurement properties will, in general, lead to suboptimal predictive properties. More seriously, however, optimizing predictive properties will tend to destroy all measurement properties a test might have.

Measurement is typically approached through latent variable models, such as the various item response theory models (Hambleton & Swaminathan, 1985) or common factor models (Jöreskog, 1971; Bollen, 1989). Inspecting the structure of any latent variable model will show that items that measure the same latent variable must be correlated. Now suppose that we want to construct a test to measure extraversion. When optimizing measurement properties, we will construct a homogeneous test, i.e., a test with correlated items. If we now were to use the test for prediction, this property, which is desirable from a measurement point of view, would translate into a problem from a prediction point of view. In essence, any set of items selected on the basis of measurement properties will show multicollinearity in prediction: The items will not add independently to the regression equation. Thus, focussing on the measurement properties by necessity leads to suboptimal predictive properties.

On the other hand, if we focus on prediction instead of measurement, we will turn up with a completely different test. Suppose that we construct a test to maximally predict extravert behavior, for instance, the tendency to engage in group discussions. When optimizing this prediction, we are likely to select items that add to the regression equation independently. That is, we would avoid rather than produce multicollinearity, and therefore we would select uncorrelated items rather than correlated ones (Lord & Novick, 1968, p. 271). Optimizing prediction necessarily produces a set of items that add to the prediction equation independently of one another, and therefore such procedures are likely to select items that measure different things. This problem will occur even if we are selecting items for predicting a highly relevant behavioral domain. Thus, in selecting items on the basis of predictive criteria, we will produce a heterogeneous test and not a homogeneous one. And because all measurement models imply homogeneity, such models will hardly ever fit a test constructed for optimal prediction. The reason for this is that the selected items will not measure the same attribute, and therefore cannot possibly be valid for measuring one attribute.

Measurement and prediction will, in general, not go hand in hand. In fact, they will work against each other. This trade-off occurs because the structure of the prediction problem is simply radically different from the structure of the measurement problem. Optimal predictive properties imply suboptimal measurement properties, and optimal measurement properties imply suboptimal predictive properties.

Again, it is clear that we cannot have it all at the same time.

Is validity a matter of degree?

The above examples point to a serious problem for any 'overarching' conception of validity. Even such simple and commonplace problems as measurement and prediction are not structured in a way that allows for a simultaneous maximization of desirable measurement and prediction properties. It is therefore impossible to improve a test on all the relevant fronts at the same time. There is no inherent problem about this, but a serious problem will occur when we couple these observations with the notion of an overarching validity conception that is supposed to come in degrees. That validity is a matter of degree has become more or less a dogma of construct validity. Cronbach & Meehl (1955, p. 290) state that 'the problem is not to conclude that the test "is valid" for measuring the construct variable', but that 'the task is to state as definitely as possible the degree of validity'. Similarly, Messick (1989, p. 13) writes that 'it is important to note that validity is a matter of degree, not all or none'. In view of the above examples, I doubt whether this view is adequate.

What exactly does it mean to say that validity comes in degrees? It seems that a theorist who expresses this notion is saying that different tests, or test score interpretations, can be ordered in terms of their validity. But how does this work? Can we say, for example, that the Stanford-Binet is more valid for measuring intelligence than the WAIS? This seems relatively unproblematic, but appearances deceive here. In particular, it is unclear how we should determine this degree of validity. We could imagine that, in a given population, intelligence produces a proportionally larger amount of the variance in WAIS scores than in Stanford-Binet scores. This may be taken to imply that the WAIS is now 'more valid'. However, the situation here is by no means inconsistent with, for example, the presence of a large biasing effect in WAIS scores. So we could easily have a situation where the WAIS has the larger portion of variance produced by intelligence, but where it is also seriously biased against, say, females. The Stanford-Binet may be more unreliable but unbiased. How are we to weigh these different merits in determining which test has the higher degree of validity? It seems to me that this will be quite difficult. Moreover, the problems are going to multiply very quickly if we now move to different domains of interest. What kind of argument would it take to say that the WAIS is more valid for intelligence than Eysenck's extraversion scale is for extraversion? To be honest, I do not have a clue. And although one could imagine a kind of 'validity score' for each test, which could be considered a weighted sum of desirable characteristics (e.g., reliability, absence of bias, predictive utility, unidimensionality, etc.), the problem becomes insurmountable once we move to a comparison of tests which are intended for different purposes, such as measurement and prediction.

The Minnesota Multiphasic Personality Inventory (MMPI), for instance, has been explicitly constructed with the objective of maximizing its predictive performance with respect to clinical syndromes. The test could reasonably be said to exemplify a predictive instrument, rather than a measurement instrument. But how are we to determine whether the MMPI is more valid for predicting membership of various categories of mental disorders, than the WAIS is for measuring intelligence? This seems downright impossible. If we have two tests which were developed for the same objective, say, measurement, then we could at least imagine a kind of weighted sum of desirable measurement characteristics (provided that we agree on these). But if we have two tests which were developed for different objectives, like measurement and prediction, we can no longer do this, because there are few if any characteristics which are desirable both in measurement and in prediction. Because the structure of the measurement and prediction problems are different, the desirable test characteristics and their relative importance will be different, and therefore we cannot construct a meaningful comparison. The tests, or test score interpretations, are incommensurable: Attempting to place them on a 'validity scale' which comes in degrees is like trying to answer the question whether the U.S. baseball team is better at playing baseball than the Dutch soccer team is at playing soccer.

Saying that validity is a matter of degree implies that one can order tests, or test score interpretations, in terms of their validity. Coupled with the desire to apply the label of 'validity' to all possible instances of test use or test interpretation, this proves to be very difficult, if possible at all. It is questionable whether tests that are constructed with different purposes in mind (measurement, prediction, selection, etc.) are scalable on their 'degree of validity', and to the best of of my knowledge no procedure for doing this has been proposed. I submit that no such procedure can be found at all, simply because of the fact that we are dealing with different problems. Now, the conception of validity proposed here is one concerned with measurement. Therefore, I wish to exclude tests designed for prediction or selection from consideration, except for the case where one is explicitly claiming validity in terms of measurement for such tests (which one is by no means forced to do). This does not mean I want to go back to the 'tripartite' scheme, which distinguished different kinds of validity (namely, content, construct, and criterion validity). I think that the question of validity is one of measurement and that, as a result, content and criterion validity are of relatively minor importance. Thus, I strongly endorse Messick's (1989) suggestion to consider these as questions of content relevance and coverage instead of 'content validity', and of predictive utility instead of 'criterion validity'.

I differ in opinion, however, with respect to the question whether validity should incorporate all these aspects at the same time; I answer this question negatively. In particular, I see no reason to demand that a test, which is intended for prediction, should first be shown valid for measuring an attribute of interest. Thus, I do not endorse Messick's progressive reading of his validity matrix. Further, I consider the task raised by Cronbach & Meehl (1955), which is to state as precisely as possible the 'degree of validity', to be impossible. For a validity concept that comes in degrees requires an ordering of tests, or test score interpretations, with respect to their degree of validity. These tests and test score interpretations, however, will more often than not be incommensurable, so that it is meaningless to speak of their degree of validity. I think that most researchers realize this at some level or another, because nobody ever attempted to develop a construct validity coefficient to capture the implied ordering. Predictive utility does come in degrees, for the simple reason that it is a direct function of the association between test scores and the variable to be predicted. Validity does not, because it is not a function of any such association.

A simple concept for a simple question

The desire to create a validity concept that comes in degrees seems to result from the fact that there are two questions, both important, that can be asked of any measurement procedure. First, one may ask: '*does* the test measure the intended attribute?'. Second, one may ask: '*how well* does the test measure the intended attribute?'. In my view, the first question can only be answered dichotomously: Either differences in test scores are produced by variations in the attribute of interest, or they are not. The second question addresses the quality of the test, relative to several methodological concepts like reliability, measurement invariance, and unidimensionality. This question is conditional: it is sensible to ask how well a test measures an attribute only if the test does indeed measure it. Thus, asking for an overall evaluation of the quality of the test presupposes its validity for the intended attribute. The question is whether we should intend to cover both questions with one concept. I think it is more sensible to restrict the meaning of validity to apply only to the question whether variation in test scores is produced by variation in the attribute we intend to measure, but not the question how well we measure it. This is a natural consequence of conceptualizing validity in terms of causality: In contrast to a correlation, a causal relation does not come in degrees. I prefer to leave the question how well a test measures an attribute to the various technical approaches that have been proposed in the psychometric literature. I thus divorce validity from various psychometric concepts that are explicitly concerned with the question how well we are measuring the attribute of interest, such as reliability, unidimensionality, and bias.

Reliability Within the present conception, the problem of reliability is distinct from the problem of validity. The square root of the reliability coefficient is certainly not an 'upper limit' for validity, although it does pose an upper limit for predictive utility. But predictive utility is irrelevant to the question of validity as I construe it. The test can be valid in a given setting (i.e., it measures what it should measure), but very unreliable. The traditional question, how much of the total variation in test scores is accounted for by the intended attribute, cannot be an issue of validity because it is dependent on the variation in the population. In a population where there is no variation in the attribute, none of the variation in the test scores is produced by variation in the attribute. As I have argued above, this does not preclude the test from being valid, because the present account of validity is causal, not correlational, and in saying a test is valid, one is saying only that variations in the intended attribute will produce variations in test scores *if* the attribute variations are present. Validity is a property of the test, and not of the scores, so it should not vary with populations – except for the fact that there may be populations in which the attribute does have causal relevance for the test scores, and populations where it does not. However, while reliability does not place an upper limit on the validity of a test, validity does place a strong restriction on the applicability of the reliability concept. This is because reliability is an index of measurement precision (Mellenbergh, 1996). Thus, the question that reliability is concerned with is 'how precise are our measurements?'. Obviously, the entire notion of measurement precision presupposes that the test is valid: We cannot say that the IQ-scores measure intelligence with a certain precision, but that they do not measure intelligence. Reliability is not an upper limit for the 'degree of validity', but it is the case that invalidity prohibits any statement concerning the reliability of test scores. If a test is invalid, then the scores cannot reliably measure the attribute of interest, because they do not measure the attribute of interest at all.

Unidimensionality A second question that is highly relevant to the question how well we measure the intended attribute, but not to the question whether we measure that attribute, is the concept of unidimensionality. No psychological test is unidimensional in the sense that the test measures only one attribute in every imaginable situation. All psychological test scores depend on many attributes that, if they were to vary in a population, would cause variation in test scores. This is clear from the fact that we may always create a second source of variation in addition to the one we are studying, thereby creating multidimensionality. For instance, one may measure the genetic quality of seeds by recording the height of the grass they produce. This test is clearly valid because variations in genetic quality, if present, will produce variations in the height of the grass. But if a gardner has just mown half of the lawn, we will in addition measure the effect of his presence, which obviously also produced variation in the height of the grass. The reason is that the test is also valid for measuring the presence of lawn mowers: if there is variation in this presence (i.e., the lawn mower covers some areas, but not others), then this will produce variations in the height of the grass.

Tests themselves are therefore always 'multidimensional' insofar as this term applies to tests at all. IQ-tests may be valid for intelligence; but they will certainly also be valid for dyslexia, motivation, and reading ability. What we must always do when we are trying to single out one attribute, is to secure that, in the population we are working in, there is no variation in other attributes. This means that we must ensure that these other attributes, for which the test is *also* valid, do not come into play. We may do this in two ways. First, by carefully selecting a population where we can assume zero variation on the other attributes that may cause variation in scores. For the intelligence example, this means that we should exclude people with dyslexia or impaired reading ability from consideration. Second, we can attempt to block the effects from other variables that may cause variation in scores. This is commonly done in personality tests, for example, by telling the subject that there are no correct or incorrect answers. In that case, we are trying to block effects from, say, variations in the tendency to answer in a socially desirable manner. Now, if we succeed in our attempt to exclude other causally relevant variables from operating, then we would expect the resulting scores to fit a unidimensional model. It is clear, however, that whether such a model fits is completely dependent upon our success in blocking effects of other variables.

It is also clear that we can use a test to measure different attributes in different situations. When we try to measure intelligence, it is important to ensure that systematic variation in other causally effective attributes is minimal. However, if variations in IQ-scores also depend on variations in dyslexia, this means that we could, in principle, also use a verbal IQ-test to measure dyslexia. This could be done by administering the IQ-test to populations that are homogeneous in all attributes (including intelligence) except for dyslexia, or by blocking all other effects. Of course, this would be utterly impractical, and I am not recommending any such use of IQ-tests, but the conceptual point is clear: A test can be used, in principle, to measure any attribute that produces variations in the test scores. Unidimensionality can be created, in a sense, by administering tests to populations that are homogeneous with respect to systematic variation in all attributes but the intended one, or by blocking the effects of all unintended sources of variation. However, a test can never be used to measure an attribute that does not causally produce variations in test scores. Thus, the relation between validity and unidimensionality is similar to the relation between validity and reliability: Unidimensional measurement of intelligence presupposes the validity of the test for intelligence, but validity does not – and cannot – presuppose unidimensionality.

Bias The above comments are directly related to the problem of bias, or differential item functioning (Mellenbergh, 1989; Meredith, 1993; Millsap & Everson, 1993). Bias, like unreliability and unidimensionality, refers to the question how well we measure an attribute. In particular, it is concerned with the question whether there exist subpopulations which induce multidimensionality (Shealy & Stout, 1993). An intelligence test is biased against ethnic groups, for example, if the scores depend on an attribute different from intelligence (say, the familiarity with the English vocabulary), on which the ethnic groups score systematically lower. If the presence of bias is established, however, it is not established that the test is invalid. It is shown that the test is sensitive to variations in more than one attribute, but not that the test is insensitive to variation in the intended attribute. If the intelligence test in the above example is biased, this does not imply that the test does not measure intelligence; it may be that the test measures more than intelligence alone, or that it measures intelligence differently in the different groups (Borsboom, Mellenbergh, & Van Heerden, 2002-b). The presence of bias is thus not directly relevant to the question of validity as we have construed it here. This may sound counterintuitive, but it is in accordance with the technical formulation of item bias. This is because the formulation of bias involves an effect of the grouping variable on the expected item response, conditional on the latent variable (Mellenbergh, 1989; Meredith, 1993). However, if no latent variable underlies the test scores at all, the concept of bias cannot even be formulated. Therefore, it seems that the formulation of bias presupposes validity, rather than that validity presupposes the absence of bias.

Validity and psychometrics Validity, as it is conceptualized here, is thus the point of departure rather than of arrival. It is central to test development and use, but it does not embrace all aspects involved in these practices. It is underlying rather than overarching, simple rather than complicated, and basic rather than unified. Psychometric approaches, which generally deal with the question how well we are measuring, cannot be considered to formulate necessary or sufficient conditions for validity because they presuppose validity. Therefore, it is plausible to separate the question whether we are measuring the right attribute from the question how well we are measuring that attribute. I think that validity theory should refrain from trying to answer both questions with a single concept; not only are they radically different – one is substantive, the other methodological – but, in the past century, the technical issues involved in assessing the latter question have been largely taken over by psychometric approaches. These approaches handle the questions involved in a much more sophisticated manner than conceptual treatments of validity could ever hope for. How precise we are measuring the attribute is a question for theories of measurement precision (Mellenbergh, 1996); whether we measure only one attribute is a question of unidimensionality (Hambleton & Swaminathan, 1985); whether we measure primarily one attribute is a question of essential unidimensionality (Stout, 1991); to what extent our measurements are unbiased is a question of measurement invariance (Millsap & Everson, 1993); and so on. If there is a function here for a unified validity concept employing degrees of validity at all, it would have to involve a method of translating these characteristics into a single number. How this should be achieved is, to my knowledge, unknown. And it is very interesting to see that psychometrics has not developed the need for a validity concept; while concepts like measurement precision, unidimensionality, and invariance flourish, there is almost no psychometric literature which explicitly uses the validity concept itself. The reason for this is that validity theory has no business in psychometrics. Not because validity is irrelevant, but because the entire undertaking of psychometrics presupposes it.

6.7 Discussion

I have proposed a simple conception of validity that concerns the question whether the attribute to be measured produces variations in the measurement outcomes. This concept of validity is based on reference and causation, rather than on meaning and correlation. As a result, it is an all-or-none property. Moreover, it is a property of tests, and not of scores or of test score interpretations. Although epistemological issues are central to validation, and consequential issues are central to test use, both are considered irrelevant to the concept and definition of validity itself. The conjunction of these theses produces a viewpoint that is almost diametrically opposed to the currently endorsed conceptions of validity, which state that the concept applies to test score interpretations, that it depends on nomological networks or is at the very least theory-dependent, that it is complex and faceted, and that social, ethical, and political consequences are relevant to validity. I do not see the need for a 'unified' validity concept (Messick, 1989; Moss, 1992; Shepard, 1993; Ellis & Blustein, 1991), because I think there is nothing to unify.

Although the proposed validity concept may be dissonant with the current validity literature, few of its ingredients are truly new. In particular, several related ideas have been put forward by a number of scholars in the previous century (e.g., Cattell, 1946; Campbell, 1960; Loevinger, 1957; Kelley, 1927; Popham, 1997). A realist reading of construct validity also comes very close to the conception proposed here. In addition, I am under the impression that most researchers operate with a validity concept that is highly similar to the one I am proposing. It seems, however, that nobody has yet consistently followed through the consequences of a realist conception of psychological attributes for the concept of validity; and the emphasis on causality as opposed to correlation seems never to have been stressed. As I have argued in the present work, the consequences of such a conception are far-reaching, but the overall picture that emerges is consistent and fits the intuitive notions most researchers have about validity quite well. I therefore think that the proposed validity concept is a viable alternative to the current consensus in validity theory.

The philosophical assumptions involved in the present conception are strong; stronger, perhaps, than in any previous discussion of validity. Therefore, it may be argued that, by invoking real entities and causal relations, I am engaging in metaphysical speculation. I concede this point, but it does not bother me. The very idea, that metaphysics and science are necessarily opposed, is a relic that stems from logical positivism; in fact, I think that science is the best way of doing metaphysics we know. To the hard-boiled empiricist, I reply that it is naive to think that *any* scientific theory can get off the ground without introducing an ontological picture of how the world works, which will always contain metaphysical ingredients. Given that this is the case, the metaphysics better be good. Other objections may come from the postmodern or social constructivist camp. An obvious one is the objection that psychological attributes are social constructions, and that I am engaging in an unjustified reification of such constructions. To this objection I reply that the position taken here is indispensable for rendering a coherent picture of measurement. It is an ontological attitude one has to take.

To see this, consider the following example. We may measure the degree of aggressive behavior displayed by Donald Duck, Mickey Mouse, and Woody Woodpecker, by rating the number of aggressive acts in a randomly sampled five minutes of film. I am surely not going to deny that Donald Duck, Mickey Mouse, Woody Woodpecker, as well as their aggressive behavior are social constructions; but this is completely besides the point. The point is that, even in this highly contrived situation, the logic we are following is that differences between Donald Duck and Mickey Mouse in their universe scores of aggressive behavior will lead to differences between them in the five minutes of film we sampled. We are thus presupposing the existence of such a universe score, and we are also presupposing a causal relation between this score and the number of aggressive behaviors we have observed: If there are no differences in universe scores, then we expect no differences in the number of aggressive behaviors, but if there are, then we expect these to lead to differences in the number of aggressive behaviors. If we are going to measure something, then we will have to suppose its existence and causal impact. Whether that something is, in itself, more properly conceptualized as a construction or as some kind of natural phenomenon is irrelevant to this issue.

Although I have separated the ontological concerns in psychological measurement, among which is validity, from the epistemological ones, which include validation strategies, the present developments do have some relevance in the area of validation research. In particular, it seems that the emphasis on the role of constructs in theories, and their place in nomological networks, has prompted validation research to adopt what has been called a top-down strategy (Cervone, 1997). This basically comes down to the fact that much validation research is concerned with creating tables of correlation coefficients, and then checking whether these go in the right direction. While I do not deny the relevance of such macro-level relations, it would seem to me that the primary objective of validation research is not to establish that the correlations go in the right directions, but to offer a theoretical explanation of the processes that lead up to the measurement outcomes. That is, there should be at least a hypothesis concerning the causal processes that lie between the attribute variations and the differences in test scores. To use Embretson's (1983) terminology, validation should be concerned primarily with construct representation and only secondarily with nomothetic span.

In this view, validation is not, and cannot be, a purely or even mainly methodological enterprise. This does not mean that methodological and psychometric techniques are irrelevant to validation research, but that the primary source for understanding how the test works must be substantive and not methodological. Thus, I consider it impossible to argue for test validity solely on the basis of a multi-trait multi-method matrix. Such a matrix is helpful, but I do not view a favorable matrix configuration as constitutive of validity. What is constitutive of validity is the existence of an attribute and its causal impact on our scores. Therefore, if one does not have an idea of how the attribute variations produce variations in measurement outcomes, one cannot have a clue as to whether the test measures what it should measure. No table of correlations, no matter how big, can be a substitute for knowledge of the processes that lead to item responses. The knowledge of such processes must be given by substantive psychological theory and cannot be based on methodological principles. There are certainly tests for which a considerable body of knowledge has accumulated in this respect. Examples of research in this direction are, for instance, the cognitive modeling approach in spatial reasoning tests (Embretson, 1994) and the latent class approach in the detection of developmental stages (Jansen & Van der Maas, 1997). I think we are more likely to find evidence of validity in such explicit attempts to model respondent behavior, than in tables of correlations.

The upshot of this line of reasoning for test construction is also clear. Purely empirical methods, like those used in the construction of the MMPI, are very unlikely to generate tests that can be considered valid measurements. This is because focussing on predictive properties will destroy, rather than enhance, measurement properties such as validity (note that this does not preclude that these tests may be highly useful for prediction). Thus, it seems that one has to start with an idea of how differences in attributes will lead to differences in test scores; otherwise the project of test construction is unlikely to generate tests that are valid for more than prediction. This may be one of the few instances where psychology may actually benefit from looking at the natural sciences. In the more exact quarters, nobody starts constructing measurement instruments without the faintest idea of the processes that lead to the measurement outcomes. And, interestingly, the problem of validity appears never to have played the major and general role it has played in psychology. These two observations may well be related: The concept of validity may never have been necessary because the instruments were generally set up based on an idea of how they would work. In that case, the question what it is, precisely, that is measured, can simply be resolved by pointing to the processes that lead to the measurement outcomes.

In contrast, the question what psychological instruments measure is generally not answered by pointing to the way the instruments work, but by pointing to the relation they have with other instruments. This way of working makes the question 'what is measured?' a question to be answered after the test has been constructed. Thus, the contrast here is between a conception that sees validity as something that one puts into an instrument, and a conception that views validity as something to be discovered afterwards. Construct validity theorists have tended to construe validity as an empirical matter, that is, the question what is measured is to be answered by data. However, a century of experience with test construction and analysis clearly shows that it is very hard to find out where the scores are coming from, if tests are not constructed on the basis of a theory of item response processes in the first place. Therefore, I would like to push the proposed validity conception one step further, and to suggest not only that epistemological issues are irrelevant to validity, but that their importance may well be overrated in validation research too. A large part of test validity must be put into the test at the stage of test construction (see also Schouwstra, 2000), a stage of the testing process that has received little attention compared with the enormous emphasis that has been placed on test analysis. Thus, it is suggested here that the issue may not be first to measure, and then to find out what it is you are measuring, but rather that the process must run the other way. It does seem that, if one knows exactly what one intends to measure, then one will probably know how to measure it, and little if any validation research will be necessary. If this is correct, then the problem of validation research is not that it is difficult to find out what is measured; the problem is that it is difficult to find out what we intend to measure.

7. APPENDIX A. FUNCTIONAL THOUGHT EXPERIMENTS

Abstract

The literature on thought experiments has been mainly concerned with thought experiments that are directed at a theory, be it in a constructive or a destructive manner. This has led some philosophers to argue that all thought experiments can be formulated as arguments. The aim of this paper is to draw attention to a type of thought experiment that is not directed at a theory, but fulfills a specific function within a theory. Such thought experiments are referred to as functional thought experiments, and they are routinely used in applied statistics. An example is given from frequentist statistics, where a thought experiment is required to establish the probability space. It is concluded that a) not all thought experiments can be formulated as arguments, and b) the role of thought experiments is more general and more important to scientific reasoning than has previously been recognized.

7.1 Introduction

It could be argued that all science begins with counterfactual thinking. For the most basic question of inquiry, ‘why is the world as it is?’, can only originate from the idea that the world could have been different. That is, an explanation need only be considered if there are phenomena to be explained, and phenomena require explanation only if they are not taken for granted. Not taking phenomena for granted requires one to consider the possibility that the world could have been different, which is only possible upon the consideration of counterfactual alternatives. This, in effect, means one has to perform a thought experiment - where a thought experiment is loosely defined as a line of reasoning that proceeds from counterfactual premises.

In the light of the importance of counterfactual reasoning and thought experimenting to such basic issues of inquiry, it seems somewhat surprising that the role of the thought experiment in science has for long been neglected by philosophers of science. Apart from the pioneering work of Mach (1905/1976), and a paper by Kuhn (1977), it has only been for the last decade that a considerable body of conceptual research has emerged on the thought experiment as a philosophical, mathematical, and scientific strategy (Brown, 1991; Horowitz & Massey, 1991; Sorensen, 1992; Wilkes, 1988).

One of the most clarifying achievements in this emerging body of literature is the taxonomy proposed by Brown (1991). He classifies thought experiments as being *destructive* (aimed at refutation of a theory), *constructive* (providing support for a theory), or *platonic* (destructive for all theories but one). An example of a destructive thought experiment is Einstein's refutation of Maxwell's theory of light. Einstein reasoned that, if Maxwell's theory were correct, he would have to see a light beam as a spatially oscillatory electromagnetic field at rest, when running at the speed of light. The thought experiment is destructive, because it is used to derive a contradiction from the premises of Maxwell's theory. An example of a constructive thought experiment is the silicon-brain experiment, in which it is argued that, if all the neurons in your brain were gradually replaced by computerchips, you would still be conscious after the replacement. This thought experiment has been used as an argument for functionalism by various philosophers (see, for example, Dennett, 1991). Finally, an example of a platonic thought experiment is Galileo's famous refutation of Aristotle's theory of motion. Aristotle's theory stated that heavier objects fall with greater acceleration than lighter objects. Galileo reasoned that, if Aristotle's theory were true, a heavier object tied to a lighter object should, when falling from a given height, simultaneously reach the ground sooner and later than the heavier object alone. This thought experiment is platonic in the sense that it does not only refute Aristotle's theory, but at the same time establishes a single alternative theory, namely the theory that acceleration does not depend on the mass of an object. Platonic thought experiments may therefore be conceived of as destructive thought experiments that are constructive for a single alternative. The majority of thought experiments that have thus far been considered in the literature can be classified in Brown's taxonomy, a possible exception being what Bunzl (1996) has called the consistency thought experiment. The essential feature of consistency thought experiments is that 'typically, such thought experiments result in a modification of background assumptions rather than any change in the theory itself' (p. 234). An example is the Einstein-Podolsky-Rosen (EPR) thought experiment, which eventually did not serve to refute or support quantum mechanics, but resulted in a modification of our background assumptions.

All thought experiments that have thus far been considered in the literature are of the type that is directed at a theory, be it in a constructive or a destructive manner (notwithstanding the fact that consistency thought experiments do not result in a modification of the theory, the EPR thought experiment was clearly directed at quantum theory). This has led some philosophers, such as Norton (1991), to claim that all thought experiments can be formulated as arguments. The aim of the present paper is to draw attention to a type of thought experiment that is not directed at theories, and therefore cannot be formulated as an argument, nor subsumed under Brown's taxonomy. These thought experiments may best be characterized as functional, in the sense that they create a conceptual framework that allows for the application of a theory. A functional thought experiment that we will consider in some depth is employed routinely in the application of frequentist statistics in order to establish a probability space. As such, it plays an important role in frequentist statistics as well as in the areas where statistics is applied. Before we discuss the characteristics of the thought experiment, we will shortly scetch the frequentist conception of probability, and illustrate the problem by an example drawn from the theory of mental testing.

7.2 The Frequentist Conception of Probability

It is not extremely difficult to provide a syntax for probability theory in the form of a calculus. Axiomatized systems have, for example, been provided by Kolmogorov (1933) and Rényi (1970). However, the interpretation of the probability syntax is not straightforward, and to some extent arbitrary. Several interpretations have survived up to this day (see Nagel, 1939, or Fine, 1973, for an overview of possible interpretations). The most influential interpretation of probability in applied statistics is the so-called frequentist account, which is in terms of long run frequencies.

The frequentist's long run interpretation of probability is seemingly uncomplicated. Toss a coin infinitely many times, and the limiting value of the relative frequency of heads in these trials is the probability of heads. For many applications it is important that these trials satisfy an independence condition. It is, however, not easy to specify what 'independent' means before probability itself has been defined. We cannot use the concept of probability to define the independence of trials, because probability itself is defined in terms of these trials. Hacking (1965) escapes the vicious circle by deducing the independence of trials from the independence of the outcomes of these trials. Following Hackings brand of frequentism, one requires that the trials are 'unrelated' and defines the probability of an outcome as the relative frequency of that outcome. Then the concept of probability can be applied to define the statistical independence of the outcomes, from which the independence of trials themselves can be deduced.

The definition of probability as the limit of a relative frequency in an infinitely long run of independent observations yields a very general framework for the application of the probability calculus. Its general applicability, together with its 'objective' character, have made the frequentist conception the most widely held view in applied statistics.

7.3 The Imaginary Long Run: Lord & Novick's Brainwash

The frequentist idea of probability as relative frequency in the long run is, upon closer examination, problematic. It is certainly intuitively plausible for games of dice, but it is not at all straightforward for many areas in which statistics is needed, and indeed has proven successful. We will illustrate this statement by an analysis of the problem as it occurs in mental testing, because it clarifies what the problem with the long run actually is.

In the theory of psychological testing, a basic assumption is that observed test scores contain measurement error. A subject's score on a psychological test will be influenced by factors that are not of interest to the researcher. Some of these are systematic (to be understood as stable over time, for example, certain personality characteristics), and some are random (i.e., accidental, for example, the subject had a headache at the testing occasion). Here, we will be concerned only with the random part of the error.

The basic idea of classical test theory (Lord & Novick, 1968) is that there exists a 'true score', which is to be conceived of as the observed score, stripped of its random error. The true score t is thus defined as the observed score X minus the random error E. This leads to the classic equation $X = \mathrm{t} + E$. Of course, this definition is empty unless some procedure is specified to define what error actually is. This procedure is borrowed from the theory of random errors as developed in astronomy (Edgeworth, 1888; see also Stigler, 1986, and Hacking, 1990). The idea is that, if we take measurements on many occasions, it is plausible to define the true score as the expectation of the observed scores over repeated measurements, so that $\mathrm{t} = \mathcal{E}(X)$. This is unproblematic in the context of astronomical measurements, where the repeated observations can reasonably be assumed to be independent (in the sense of the previous section). Consequently, the frequentist conception of probability as long-run relative frequency can be utilized. It is then reasonable to define the true score as the expectation over repeated observations, which is, by its definition, a constant.

This line of reasoning fails in psychological testing. Disregarding the fact that performing many measurements on the same subject is unrealistic, people learn, get tired, become familiar with the testing procedure, and so on. As a consequence, trials are not unrelated and the outcomes of the trials will not, in general, be independent. So, Hacking's (1965) method of deducing the independence of trials from the independence of outcomes does not work here. However, if we want to apply a long run frequency interpretation of probability, trials must be independent.

Lord & Novick solve this problem by introducing a thought experiment originally proposed by Lazarsfeld (1959). It runs as follows:

> 'Suppose we ask an individual, Mr. Brown, repeatedly whether he is in favour of the United Nations; suppose further that after each question we 'wash his brains' and ask him the same question again. Because Mr. Brown is not certain as to how he feels about the United Nations, he will sometimes give a favorable and sometimes an unfavorable answer. Having gone through this procedure many times, we then compute the proportion of times Mr. Brown was in favor of the United Nations.' (Lord & Novick, 1968, pp. 29-30)

In the thought experiment, the observations are rendered independent as a result of the brainwashing procedure. Now we may apply the frequency interpretation of probability. It then becomes possible to take the expectation of the observed scores and to define this expectation as the true score, which is again a constant. In the particular case of Mr. Brown, the expectation equals the probability of him giving a favorable answer, which is estimated by the proportion of times he was in favor of the United Nations. The introduction of this thought experiment has proven extremely useful in the development of both classical (Lord & Novick, 1968) and modern (Hambleton & Swaminathan, 1985) test theory.

Lord and Novick are very plain in admitting that they use a thought experiment. They are forced to do so by their subject matter. A long sequence of repeated testing occasions is so obviously implausible that they cannot ignore the problem.

The thought experiment, however, is not symptomatic for psychological testing. It is implicitly present in many applications of frequentist statistics. A long run of independent observations on the same unit does not exist anywhere in the real world. Almost independent, yes; practically independent, yes; truly independent, no. The notion of independent observations is an idealization, although it often is a useful assumption (it would certainly be a pathological case of hair-splitting to criticize the assumption of independent trials in throwing dice). In virtually every application of inferential statistics, however, the thought experiment is needed. Hacking (1965; p. 10) hinted at this when he said that long run frequency is concerned with 'what the long run frequency is or would be or would have been'. In order to be able to invoke the expectation of the outcome of measurements in medicine, economics, or social research, one always has to talk about 'what the long run frequency would have been if ...'. The conditional part of the sentence contains, in these cases, counterfactual premises. It is a thought experiment.

7.4 The Nature of Statistical Thought Experiments

The frequentist thought experiment is different from those used in physics or philosophy. The primary characteristic that distinguishes the thought experiment from the type that has received attention in the literature so far, is that it is not directed at any theory in particular. Although the thought experiment is necessary for employing the frequentist scheme of statistical inference, it is not used to support the frequentist view (in the sense of showing that the frequentist theory is 'true'). Rather, it is an integral part of that view: It creates the conceptual framework rather than supporting it. This type of thought experiment may be best characterized as functional: A functional thought experiment is not aimed at refuting or supporting a theory, but has a specific function within a theory. In the case of frequentist statistics, it functions as a semantic bridge, providing a real world interpretation for the abstract syntax of probability.

The distinction between functional and constructive/destructive thought experiments runs parallel to the distinction between theory and model. A theory can be true or false: a constructive or destructive thought experiment is intended to show that it is true or false. A statistical model, if it is internally consistent, can only be shown to be applicable or not applicable; and the functional thought experiment is used to convince the reader that it actually is applicable. This, in the frequentist thought experiment, is established by an appeal to analogy. The frequentist conception of probability is generally deemed applicable to games of dice, and as a result the outcome of trials in such a game can be considered a random variable, satisfying the required independence conditions. In the Lord & Novick thought experiment, we are asked to imagine a situation where a subject's response could be considered analogous to the outcome of trials in throwing dice. The crucial point is that an important structural characteristic (the 'randomness' of the trial outcomes) is preserved in the new domain. If this much is granted, the rest of the theory follows smoothly.

The use of functional thought experiments is not limited to the semantic bridge function in frequentist statistics. Actually, there are many statistical models and techniques that are not interpretable without a thought experiment. Because it is not within the scope of this paper to give a complete and thorough overview of the use of functional thought experiments, we only mention briefly some of the models in which they are used. One example is the causal model of Rubin (1974; see also Holland, 1986), in which it is necessary to consider a concept called the counterfactual expectation. In a standard experimental setup employing an experimental and a control condition, this would for example be the counterfactual expectation of the dependent variable in the control condition, that would have 'existed' if the subjects in that condition had been assigned to the experimental condition. A related statistical technique where thought experiments are needed is the use of covariates as control variables. In this technique, the observed means for a given variable are corrected for a covariate. For example, a mean difference between men and women on annual income may disappear, once the observed difference is corrected for the sex difference in educational level. Such a result is interpreted as 'there would not have been a sex difference in annual income, had men and women had the same average eductional level', which is clearly a counterfactual statement. All these statistical thought experiments are functional, since they render a model applicable but are not directed at a theory.

7.5 Discussion

Functional thought experiments are not aimed at a theory, but create a conceptual framework for the application of a theory or model. Therefore, they cannot be formulated as arguments in the sense of Norton (1991). Another consequence is that functional thought experiments cannot be described as constructive or destructive for a theory, and are not captured in the taxonomy of Brown (1991). It therefore seems that the functional thought experiment specifies a distinct class of thought experiments. This may be one of the reasons that it has gone unnoticed in the philosophical literature on thought experiments. Another reason may be that statistical thought experiments are, in most cases, not explicitly presented as counterfacual lines of reasoning. Most treatments of statistics do not explicate the counterfactuals that are employed in statistical arguments, Lord and Novick (1968) being one of the rare exceptions to this rule.

There are several implications of the present discussion that are of interest to the role of thought experiments in science. In the literature on thought experiments, it is generally contended that the method of thought experiment is used almost exclusively in philosophy and physics. Upon the present discussion, however, this does not seem to be the case. Thought experiments are used incidentally in physics and regularly in philosophy, but they are commonplace in medicine, biology, and the social and behavioral sciences. Confidence intervals, p-values, reliabilities, and likelihood ratios all result from the same kind of thought experiment: The counterfactual long run. We therefore think it is not unreasonable to say that the long run frequency thought experiment, although generally not recognized as such by those who employ it, is the most common thought experiment in science. The role of thought experiments and counterfactual reasoning may therefore be more general and more important to the development of science than has been previously recognized.

8. APPENDIX B. DIFFERENT KINDS OF DIF: A DISTINCTION BETWEEN ABSOLUTE AND RELATIVE FORMS OF MEASUREMENT INVARIANCE AND BIAS

Abstract

In this paper, a distinction is made between absolute and relative measurement. Absolute measurement refers to the measurement of traits on a group-invariant scale, and relative measurement refers to the within-group measurement of traits, where the scale of measurement is expressed in terms of the within-group position on a trait. Relative measurement occurs, for example, if an item induces a within-group comparison in respondents. We discuss these distinctions within the framework of measurement invariance, differentiating between absolute and relative forms of measurement invariance and bias. It is shown that items for relative measurement will produce bias as classically defined if the mean and/or variance of the trait distribution differ between groups. This form of bias, however, does not result from multidimensionality but from the fact that measurement is on a relative scale. A logistic regression procedure for the detection of relative measurement invariance and bias is proposed, as well as a model that allows for the incorporation of items for relative measurement in test analysis. Implications of the distinction between absolute and relative measurement are discussed, and prove to be especially relevant for the domain of personality research.

8.1 Introduction

Questions concerning test validity are central to test theory and scientific progress, but also to ethical, legal, and political issues related to test use (Cronbach, 1988; Messick, 1989). Within validity theory, the development of concepts such as measurement invariance and item bias has provided an important conceptual framework for thinking about these issues. However, the relation between construct validity and measurement invariance is not yet entirely clear. This paper purports to provide some insight into this relation by presenting a distinction between different kinds of measurement invariance and bias, and by evaluating these within a construct validity perspective. Especially, we are concerned with the meaning of measurement invariance and bias in the domains of personality and attitude testing.

The ideas of item bias and measurement invariance were first conceived of in Item Response Theory (IRT) by Lord (1980), who proposed that measurement invariance with respect to group membership holds if an item follows the same Item Characteristic Curve (ICC) in all groups. In IRT for dichotomous item responses, this requirement means that the probability of a given response is the same for members of different groups with the same position on the trait measured by the test (Mellenbergh, 1989; Millsap & Everson, 1993). The notion of measurement invariance can be generalized to cover a wider range of models by making the more general requirement that the distribution function of the item response is invariant across groups, conditional on the latent trait (Meredith, 1993). Thus, an item j, answered by subject i and assumed to measure latent trait θ, is measurement invariant with respect to selection on variable V if and only if the following equation holds for the distribution function F of the item response U_{ij}:

$$F(U_{ij} = u_{ij} \mid \Theta = \theta_i, V = v_i) = F(U_{ij} = u_{ij} \mid \Theta = \theta_i) \qquad (8.1)$$

for all u, θ, v.

This definition corresponds to unobserved conditional invariance (UCI) as discussed by Millsap and Everson (1993). Whenever the above condition is violated, the item is said to be biased. In the IRT literature, the more neutral term Differential Item Functioning (DIF) is often preferred. In this paper, we use the terms interchangeably. Thus, item bias occurs if and only if

$$F(U_{ij} = u_{ij} \mid \Theta = \theta_i, V = v_i) \neq F(U_{ij} = u_{ij} \mid \Theta = \theta_i) \qquad (8.2)$$

for some u, θ, v.

As can be seen from Formula 8.2, item bias amounts to an influence of group membership on the item response in subpopulations with the same position on the trait. Item bias may, for example, occur in an IQ-test if men score better on an item than women, although there is no difference in intelligence. Item bias is to be sharply distinguished from impact, which amounts to differences in test scores that are due to differences in trait distributions (Millsap & Everson, 1993). For the above example, impact would occur if the better scoring of men was due to higher mean intelligence. In this case, the differential performance can be entirely attributed to a difference in location of the latent trait distributions.

Item bias bears directly on construct validity. When the intention is to measure a unidimensional concept, one would intuitively expect that a biased item is necessarily invalid. Indeed, item bias has been equated with multidimensionality (Kok, 1988). Shealy and Stout (1993, p. 198) remark that 'test bias occurs if the test under consideration is measuring a quantity in addition to the one the test was designed to measure, a quantity that both groups do not possess equally'. On the item level, item bias is seen as the effect of an unwanted, additional variable on the item response. In this view, bias with respect to group membership is produced by an association of this additional variable with group membership, thus influencing item responses differently in each group. Consequently, if the intention is to measure a single trait, removing items that are biased with respect to group membership from the test seems a plausible strategy to enhance construct validity. One of the objectives of this study, however, is to show that this is not always the case.

The conceptual framework of measurement invariance has been developed from the perspective of cognitive testing, and this is the primary field where DIF-analyses are used. One reason for this is that the concepts of measurement invariance and bias are most salient in individual decisions that are 'high-stake', for example when tests are used for college admissions or personnel selection - domains where cognitive tests are of primary importance. For scientific purposes, however, the importance of questions concerning measurement invariance and bias is not restricted to any specific theoretical domain. Indeed, there have been some recent applications in personality testing (Ellis, Becker & Kimmel, 1993; Huang, Church & Katigbak, 1997; Smith & Reise, 1998), and the screening for DIF is equally important in the field of personality psychology as in any other domain of psychological measurement.

Now, the technical aspects of measurement invariance and bias can be applied to domains other than cognitive testing without any specific problems, since they are of a mathematical nature and thus entirely syntactical. However, the meaning of measurement invariance and bias may change with the field of application. We will be concerned with one specific shift of meaning that occurs when the concepts of measurement invariance and bias are used in the area of personality and attitude testing. Especially, we will look at the meaning of measurement invariance when items invoke a frame of reference, for example by inducing a within-group comparison. It will be argued that such items will display bias as defined above. However, from a construct validity perspective, such items are not necessarily invalid, but rather there is a misfit between the model that is used and the cognitive processes that are involved in the item response. To deal with this problem, we extend the conceptual framework of measurement invariance. We introduce a distinction between relative and absolute forms of measurement and define the corresponding forms of measurement invariance and bias. We show that items inducing a within group comparison lead to absolute, but not to relative bias. Following this distinction it will be argued that items showing absolute, but no relative bias, do not necessarily have to be eliminated from a test. Upon proper analysis, these items - although biased according to current standards - can enhance test validity and do not necessarily produce test bias.

8.2 Absolute and relative forms of bias

Consider the following thought experiment. Imagine a world where the development of measurement theory in the social sciences has preceded measurement in the natural sciences. In this world, psychological research on attitudes and self-efficacy is common practice, whereas concepts such as 'height' or 'weight' are still to be invented. A psychologist might then conceive of a person's 'height' as a useful construct for the explanation of certain types of behavior, such as the predisposition of some individuals to participate in basketball, and the difficulty others experience when reaching for the upper shelves of a closet. However, because a measurement apparatus for the assessment of height has not yet been invented, he can only use social science's measurement methods to assess height. For this reason, he would probably go about constructing a questionnaire consisting of items like 'I have trouble getting a book from the upper shelves in a library', 'Sometimes I have to bend over in order to see my face in a mirror', and 'When sitting on somebody else's chair, I cannot usually reach the ground with my feet'. Suppose he would have constructed a questionnaire consisting of the aforementioned three items, and would add a fourth on the basis of his intuitions concerning the relation between height and basketball:

'I would do well on a basketball team'.

Although this item has high face validity, a formal test of DIF points out that the item shows DIF with respect to sex; women have a higher probability of answering 'yes' than do men of the same height. Formally, if we call the item response (scored dichotomously with 'yes':1 and 'no':0), take height to represent the latent trait θ, let V denote sex (say, $V= 0$ for men and $V= 1$ for women), and define the probability of the item response u_{ij} as $P(U_{ij} = u_{ij})$, then

$$P(U_{ij} = 1 \mid \Theta = \theta_i, V = 0) < P(U_{ij} = 1 \mid \Theta = \theta_i, V = 1), \tag{8.3}$$

for at least some values of θ, so the item has DIF. To increase test validity, our psychologist removes the item from the test. But is this a sensible thing to do? We think it is not, and this has to do with the nature of the sex difference. A woman, 5.8 feet tall, may imagine a basketball team consisting of women, and conclude that she would do good because she is relatively tall - considering her sex. A man of the same height may correctly judge himself to be relatively short - considering his sex - and conclude the opposite. Because of the within-group comparison made by both sexes, the item has absolute bias: men and women of the same height do not have the same probability of an affirmative answer. However, men and women with the same relative height within their own group (for example, a standard deviation above the group mean) do have identical probabilities of an affirmative answer. Thus, although the item is biased with respect to absolute height, it is not biased with respect to relative height.

We now formalize this notion. Denote the relative position on the trait by Ω, taking on values ω_i. Then, for the item under consideration, although it is true that

$$P(U_{ij} = 1 \mid \Theta = \theta_i, V = 0) < P(U_{ij} = 1 \mid \Theta = \theta_i, V = 1), \tag{8.4}$$

for some θ, it is also true that

$$P(U_{ij} = 1 \mid \Omega = \omega_i, V = 0) = P(U_{ij} = 1 \mid \Omega = \omega_i, V = 1), \tag{8.5}$$

for all ω. Following this insight, we can distinguish two forms of measurement. Absolute measurement refers to a procedure to measure the trait on an absolute scale (e.g., 'I have trouble getting a book from the upper shelves in a library'), and relative measurement refers to a procedure to measure the trait on a relative scale (e.g., 'I would do well on a basketball team'), where the measurement unit is expressed in terms of the relative position within the group to which the subject belongs. The different forms of measurement imply different definitions of measurement invariance and bias. Accordingly, we differentiate between absolute and relative measurement invariance and their corresponding forms of bias as follows:

Definition 1. For an item j, generating item response U_{ij} and measuring trait θ, *absolute measurement invariance* with respect to selection on variable V holds if and only if

$$F(U_{ij} = u_{ij} \mid \Theta = \theta_i, V = v_i) = F(U_{ij} = u_{ij} \mid \Theta = \theta_i) \tag{8.6}$$

for all u, θ, v. *Absolute bias* with respect to selection on variable V occurs if and only if

$$F(U_{ij} = u_{ij} \mid \Theta = \theta_i, V = v_i) \neq F(U_{ij} = u_{ij} \mid \Theta = \theta_i) \tag{8.7}$$

for some u, θ, v. Note that these are the usual definitions of measurement invariance and bias (Mellenbergh, 1989; Millsap and Everson, 1993).

Definition 2. For an item j, generating item response U_{ij} and measuring trait θ, *relative measurement invariance* with respect to selection on variable V holds if and only if, for the item response conditional on ω (the relative, within-group position on θ),

$$F(U_{ij} = u_{ij} \mid \Omega = \omega_i, V = v_i) = F(U_{ij} = u_{ij} \mid \Omega = \omega_i) \tag{8.8}$$

for all u, ω, v. *Relative bias* with respect to selection on variable V occurs if and only if

$$F(U_{ij} = u_{ij} \mid \Omega = \omega_i, V = v_i) \neq F(U_{ij} = u_{ij} \mid \Omega = \omega_i) \tag{8.9}$$

for some u, ω, v.

Now the problem occurs how to specify ω. This depends primarily on the nature of the cognitive processes involved in answering personality items, which at present is unknown for most tests. However, it is obvious that ω should be some transformation of the trait θ. The form of this transformation might be different for different tests, items, and it could, in principle, even vary over groups. So, in the general definitions the exact form of the transformation should not play a role. However, in order to apply the concepts of relative measurement and bias, we have to assume some form for the transformation. We will conceive of ω as the within-group standardized transformation of θ. There are three reasons for this. First, this assumption leads to precise and testable hypotheses. Second, the Z-transformation has many desirable mathematical properties that will become apparent in the next section. Third, even if the actual comparison is not made on a standardized within-group scale (for example, if it is in terms of absolute deviations from the mode), the Z-transformation will often provide a reasonable approximation. In the remainder of this paper, therefore, ω will be equated with the within-group standardized transformation of θ, which we will denote as Ξ, taking on possible values ξ_i. Note that the first two moments of the distribution of ξ are - by definition - the same across groups: It has mean 0 and variance 1 within each of the groups. Further, if the original trait distributions are normal, the distribution of ξ is the same in each group. This observation has implications for the theory of multidimensionality, which are discussed below.

By equating ω with the within-group standardized transformation of θ, definitions 8 and 9 are altered by substituting Ξ and ξ_i for Ω and ω_i, respectively, and the resulting concepts may be coined 'standardized relative measurement invariance and bias'. To avoid an overload of terminology, however, in text we will continue to speak of relative measurement invariance and bias, with the understanding that an assumption concerning the form of the transformation has been made; consequently, we will use ξ instead of ω in the formulae. Finally, we would like to stress that different forms of the transformation could be used, and that the appropriateness of the chosen form of the transformation represents a testable hypothesis. Thus, although the exact form of the transformation does not play a role in the general definitions given above, it does play a role in the consequences and assessment of relative measurement invariance and bias. As a consequence, the results derived hereafter do depend on the appropriateness of the Z-transformation.

8.3 The Relation between Absolute and Relative Bias

In this section, we examine the relation between absolute and relative measurement invariance and bias. This paragraph is primarily intended to show the mutual incompatibility of absolute and relative measurement invariance. The terminology of IRT will be used because it allows for a clear and comprehensible expression of the concepts of measurement invariance and bias. Later in this paper, we return to the more general case and also discuss a structural equation modeling (SEM) approach to modeling relative measurement invariance.

In parametric IRT, the probability of a correct response to an item is expressed as a function of a person characteristic (the position on the latent trait) and a number of item characteristics (e.g., the difficulty of the item and the item's ability to discriminate between subjects with different trait values). A common form for this relation between the probability of a correct response, the position on the latent trait, item difficulty, and item discrimination, is provided by Birnbaum's (1968) two parameter logistic model:

$$P(U_{ij} = 1 \mid \theta_i, \alpha_j, \beta_j) = \frac{e^{\alpha_j(\theta_i - \beta_j)}}{1 + e^{\alpha_j(\theta_i - \beta_j)}}, \tag{8.10}$$

where β_j indicates the difficulty of item j, α_j is its discrimination parameter, and θ_i denotes subject i's position on the latent trait θ. For a single item, model Formula 8.10 gives the Item Characteristic Curve (ICC), which results from plotting the response probabilities for this item against the latent trait values. The parameter β_j determines the location of the ICC and the parameter α_j its slope in the point $\theta_i = \beta_j$, hence their interpretation as item difficulty and item discrimination. Absolute measurement invariance can be expressed as the requirement that the ICC's for different groups are identical: If ICC's are identical across groups, the probability of a correct response, conditional on the latent trait, is the same for subjects with the same latent trait values, regardless their group membership.

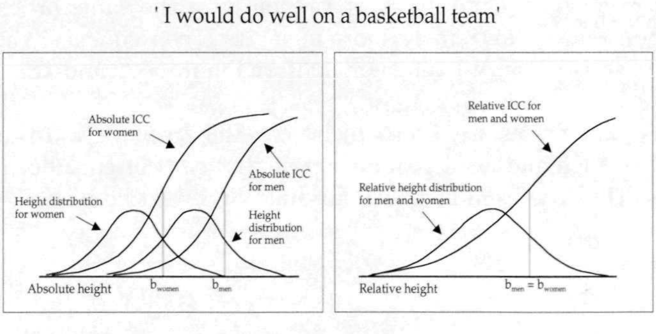
The ICC results from plotting the probability of a correct response against the latent trait, for which absolute trait values are used. Following our distinction between absolute and relative measurement, we will refer to the 'classical' ICC discussed above as an absolute ICC. However, it is also possible to plot the probability of a correct response against relative trait values. This gives us a relative ICC. The relative ICC relates the probability of a correct response to the within-group standardized latent trait ξ. Like the absolute ICC, its form is determined by two parameters indicating the relative (within-group) difficulty and slope. We will denote these parameters as $\beta_{j_{rel}}$ and $\alpha_{j_{rel}}$, and refer to the original absolute parameters as $\beta_{j_{abs}}$ and $\alpha_{j_{abs}}$. The form of the two-parameter variant of the relative ICC is determined by the following formula:

$$P(U_{ij} = 1 \mid \xi_i, \alpha_{j_{rel}}, \beta_{j_{rel}}) = \frac{e^{\alpha_{j_{rel}}(\xi_i - \beta_{j_{rel}})}}{1 + e^{\alpha_{j_{rel}}(\theta_i - \beta_{j_{rel}})}}. \tag{8.11}$$

As is the case with absolute measurement invariance, the requirement of relative measurement invariance, that the relative ICC's must be equal across groups, can be reformulated as the requirement that the parameters of the relative ICC's, $\beta_{j_{rel}}$ and $\alpha_{j_{rel}}$, are equal across groups.

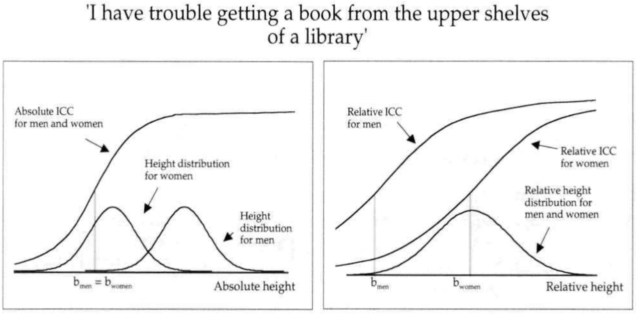
The question arises how the relative ICC relates to the absolute ICC, or, alternatively, how the relative item difficulty and discrimination parameters relate to the absolute item difficulty and discrimination parameters. In particular, it is interesting to inquire under which conditions absolute and relative measurement invariance may both hold. We discuss the relation between absolute and relative measurement at an intuitive level before turning to a more precise formulation of the relation between absolute and relative parameters.

Figure 8.1. Absolute and relative ICC's for an item with relative measurement invariance but absolute bias.


' I would do well on a basketball team'
Absolute ICC for women
Absolute ICC for men
Height distribution for women
Height distribution for men
Absolute height
b_women
b_men
Relative ICC for men and women
Relative height distribution for men and women
Relative height
b_men = b_women

Consider the item for relative measurement in the height test ('I would do well on a basketball team'). The left half of Figure 8.1 shows, in a single graph, the population distributions of the latent trait and the absolute ICC's for men and women. (The population distributions and the ICC's can be drawn in a single graph because, in IRT, trait parameters and item difficulty parameters are on the same scale.) The ICC's for men and women differ in location (i.e., item difficulty), indicating absolute bias. The right half of Figure 8.1 shows the relative ICC's, that is, the item response probabilities plotted against relative trait values. Also shown are the population distributions of the relative trait values. These are identical because the trait has been standardized within groups (the distribution has mean 0 and variance 1 in each of the groups). Because the locations of the absolute ICC's relative to the within group distributions are the same, the relative ICC's are identical for men and women. This indicates that there is no relative bias; the item has relative measurement invariance.

Figure 8.2. Absolute and relative ICC's for an item with absolute measurement invariance but relative bias.

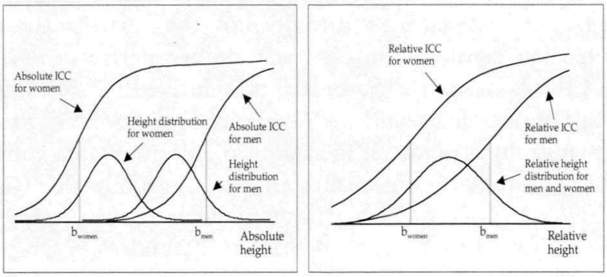


In contrast, Figure 8.2 shows an item with absolute measurement invariance ('I have trouble getting a book from the upper shelves in a library', scored yes:0 and no:1). The absolute ICC's, shown in the left half of the figure, are identical for men and women, indicating absolute measurement invariance. However, the absolute ICC is located relatively further away from the mean of the trait distribution for men than it is for women; moderately short women have the same probability of an affirmative response as do extremely short men. As a consequence, the relative ICC's are different, as is shown in the right half of Figure 8.2, and the item has relative bias.

Finally, Figure 8.3 shows an item for which both the absolute and the relative ICC's are different for men and women, for example 'In bed, I often suffer from cold feet'. This indicates that the item has both absolute and relative bias.

Figure 8.3. Absolute and relative ICC's for an item with both absolute bias and relative bias.

'In bed, I often suffer from cold feet'



The figures suggest that absolute and relative measurement invariance cannot hold simultaneously if the distribution of the latent trait differs across groups. This is due to the fact that the absolute ICC's cannot be simultaneously located at the same position on the absolute trait (absolute measurement invariance) and have the same location relative to the group means (relative measurement invariance). We now turn to a more precise formulation of the relation between absolute and relative parameters.

We have defined a relative ICC in model Formula 8.11. The relative parameters can be expressed as functions of the absolute parameters, because the relative trait values are linear transformations of the absolute trait values; they are defined by the within-group standardization

$$\xi_i = \frac{\theta_i - \mu_{\theta_v}}{\sigma_{\theta_v}} \tag{8.12}$$

where μ_{θ_v} and σ_{θ_v}represent the mean and standard deviation of the trait distribution in group v, to which subject i belongs. This standardization is performed separately for each group, which means that a possibly different linear transformation of the trait values is performed in each group. The relation between absolute and relative item parameters can be expressed as the effect of these transformations on the item parameters.

The absolute difficulty parameter $\beta_{j_{abs}}$ is defined as the latent trait value for which the probability of a correct response, given the latent trait, is 0.5. In the standardization all trait values are rescaled through Formula 8.12. It follows that the relative difficulty parameter is the relative trait value that is associated with the absolute trait value through the linear transformation given in model Formula 8.12. So,

$$\beta_{j_{rel}} = \frac{\beta_{j_{abs}} - \mu_{\theta_v}}{\sigma_{\theta_v}} \tag{8.13}$$

The relative and absolute difficulty parameters are related through a linear transformation that is possibly different for each group. Whether the transformation is different depends on differences in the trait distribution between groups. It follows that, if the mean and/or variance of the trait distribution differ between groups, absolute and relative measurement invariance in the difficulty parameters cannot hold simultaneously.

A similar effect holds for the discrimination parameters. It is intuitively plausible that differences in trait variances have an effect on the slope of the relative ICC. Since the standardization changes the distances between trait values by a factor $1/\sigma_{\theta_v}$, we can expect the slopes of the absolute and relative ICC's to differ by a factor σ_{θ_v}. Formally, we can derive this result as follows. We may set Equations 8.10 and 8.11 equal within each group, because the standardization of trait values is a linear transformation, and consequently the probability of a response for each value of θ and the corresponding value of ξ must be the same within each group. Substituting the right hand sides of Equations 8.12 and 8.13 for ξ_i and $\beta_{j_{rel}}$, respectively, and solving for $\alpha_{j_{rel}}$, we obtain

$$\alpha_{j_{rel}} = \sigma_{\theta_v}\alpha_{j_{abs}} \tag{8.14}$$

From this relation it follows that, if the variance of the trait distribution differs over groups, either absolute or relative bias in the discrimination parameter will occur.

Thus, absolute and relative measurement invariance cannot hold simultaneously if groups differ in means and/or variances of the latent trait. This relation also holds for other than dichotomous item responses, e.g. polytomous or continuous item responses. We do not formally prove this statement, because we think it is rather obvious: Any type of item can only be simultaneously measurement invariant with respect to θ and with respect to ξ if the transformation that leads from θ to ξ is identical across groups. This transformation can only be identical if the means and variances of the population distributions on the latent variable are the same. Thus, absolute and relative measurement invariance can hold simultaneously, but only if there are no differences in the means and variances of these population distributions. If there are differences in the means and/or variances of these distributions, absolute measurement invariance will lead to relative bias, and relative measurement invariance will lead to absolute bias.

8.4 DIF-Detection and Modeling

The question arises how to detect relative DIF in an empirical situation. The formulation of relative measurement invariance as the requirement that relative ICC's are identical across groups opens a range of possibilities. It makes IRT-based techniques for the assessment of absolute DIF available for the detection of relative DIF. Thus, methods based on area measures, such as signed and unsigned area tests, as well as statistics for the equality of item parameters, or Mantel-Haenszel based procedures, could in principle be used to assess relative DIF (see Holland and Wainer, 1993, or Camilli and Shepard, 1994, for overviews of available techniques for the detection of absolute DIF).

For the dichotomous case, we will present an adaptation of the logistic regression approach (Swaminathan & Rogers, 1990) for the detection of relative measurement invariance and DIF, because it is simple and instructive. The logistic regression approach is based on the idea that, in a regression of the binary item response on the continuous latent trait, group membership should not contribute significantly to the prediction, once the trait has been included as a predictor in the regression equation. If, as is usual, a sumscore X is used as a proxy for θ, an item can be tested for DIF by fitting the full regression

$$P(U_j = 1) = \frac{e^{c_0 + c_1 X + c_2 V + c_3 XV}}{1 + e^{c_0 + c_1 X + c_2 V + c_3 XV}} \tag{8.15}$$

where the c_0 to c_3 are regression parameters, X is a sumscore, and V is a dummy variable coding for group membership. In this procedure, one checks whether the parameters 2 and c_3 differ from zero. Here, c_2 represents the main effect of group membership and c_3 the interaction between group membership and the sumscore. A significant parameter value for c_3 would indicate non-uniform DIF, which occurs when the amount of DIF changes across levels of X (Mellenbergh, 1982). If the parameter value for c_3 is not significant, but the parameter value for c_2 is, this indicates uniform DIF, i.e., a constant amount of DIF across levels of X.

Relative measurement invariance can also be tested using logistic regression. Since the concept of relative measurement invariance requires that there is no effect of group membership, given the relative position on the trait, we substitute Z for X in the regression (where Z is the within-group standardized value of X, and is taken as a proxy for ξ - note that one needs a set of absolute items to compute X before this procedure can be carried out). This gives

$$P(U_j = 1) = \frac{e^{c_0 + c_1 Z + c_2 V + c_3 ZV}}{1 + e^{c_0 + c_1 Z + c_2 V + c_3 ZV}} \tag{8.16}$$

Again one proceeds by checking the significance of the parameters c_2 and c_3, but now significant parameter values indicate relative DIF instead of absolute DIF. Analogous to the absolute case, a significant value for the parameter c_3 indicates an interaction between group membership and the latent trait, corresponding to non-uniform relative DIF. A significant value for the parameter c_2 without a significant value for c_3 indicates uniform relative DIF.

If an item shows relative measurement invariance but absolute DIF, the item may be used as a relative indicator of the trait in question. This requires modeling relative measurement, which implies that the absolute and relative items be treated differently. For absolute items, item parameters should be equal across groups as usual. For relative items, however, the (absolute) item parameters will differ across groups if the trait distributions differ (see the previous section). Now, under relative measurement invariance, the differences in discrimination and slope are functions of the difference in trait distributions. The relations between the absolute parameters in both groups are simple and can be deduced from Formulae 8.13 and 8.14. Setting the right hand side of Formula 8.13 equal for two groups and solving for the absolute difficulty in group 1 gives

$$\beta_{j1_{abs}} = \mu_{\theta_1} + \sigma_{\theta_1} \left[\frac{\beta_{j2_{abs}} - \mu_{\theta_2}}{\sigma_{\theta_2}} \right] \tag{8.17}$$

for the difficulty parameters, where the second subscript on these parameters indicates group. For the discrimination parameters we obtain

$$\alpha_{j1_{abs}} = \frac{\sigma_{\theta_2}}{\sigma_{\theta_1}} \alpha_{j2_{abs}} \tag{8.18}$$

Modeling relative item responses can be carried out using these relations. A (slightly ad hoc) method for doing this would consist of the following three steps. First, estimate the means and variances of the trait distributions in both groups using only a set of absolute items. This provides estimates for the means and variances of the trait distribution in the different groups. Second, estimate the absolute item parameters for the relative items in one group (use the largest group for better parameter estimation). This provides estimates for the absolute item parameters for the relative items in one group, so that the difficulty and discrimination parameters for each relative item can be inserted into the right hand side of Formulae 8.17 and 8.18. Finally, fix the absolute parameters for the relative items in the second group at the values given by Formulae 8.17 and 8.18. This method is somewhat ad hoc, but has the advantage of being simple and easy to implement in widely available software. Also, this procedure yields the possibility to assess the fit of the entire model with absolute and relative items, thus testing the fit of the absolute and relative part of the model simultaneously.

Another option that may be taken, which is especially useful in a SEM approach, is to conceptualize the relative, within group dimension as a separate latent variable. SEM programs such as LISREL (Jöreskog & Sörbom, 1993) are flexible enough to specify an absolute latent variable for the absolute items and a relative latent variable for the relative items. The relative latent variable is then restricted in such a way that it becomes a within-group standardized rescaling of the absolute latent variable. This requires that the relative latent variable correlates perfectly with the absolute latent variable within groups, and that it has a mean of zero and a variance of one within each of the groups. To provide a within-group correlation of one between the absolute and relative latent variable, the covariance matrix of these latent variables must be subjected to nonlinear restrictions. Further, the mean and variance of the relative latent variable are fixed at zero and one, respectively, and specified to be invariant across groups. The between-group differences in means and variances for the absolute latent variable, however, are freely estimated. Then one subjects the entire model to a test for strict factorial invariance (Meredith, 1993) to test for relative measurement invariance. The formal details of this model are outlined in the Appendix. This is an elegant procedure for fitting the relative model and a useful extension of the SEM framework. To our knowledge, widely available IRT software does not allow the required restrictions to be imposed. For dichotomous item responses, this approach can therefore only be taken indirectly through the analysis of tetrachoric correlations with SEM programs.

8.5 Illustration 1

We will illustrate some of the ideas and procedures set forth in this paper by analyzing a Dutch version of the Personality Research Form-E (PRF-E), a widely used personality questionnaire due to Jackson (1974). The PRF-E was administered to 157 male and 279 female undergraduate psychology students. We will assess absolute and relative measurement invariance with respect to sex.

Table 8.1. p-values for the items in the PRF-E subscale 'impulsivity'. Negative items (items 9 to 16) have been recoded, so that all p-values represent the proportion of indicative responses.

Item	p_{males}	p_{females}
1. Often I stop in the middle of one activity in order to start something else.	.64	.62
2. I often say the first thing that comes into my head.	.50	.63
3. When I go to a store, I often come home with things I had not intended to buy.	.42	.58
4. Many of my actions seem to be hasty.	.47	.46
5. I have often broken things because of carelessness.	.50	.47
6. Most people feel that I act impulsively.	.41	.45
7. Sometimes I get several projects started at once because I don't think ahead.	.59	.59
8. I find that thinking things over very carefully often destroys half the fun of doing them.	.44	.56
9. I am careful to consider all sides of an issue before taking action.	.52	.64
10. I am pretty cautious.	.31	.34
11. Rarely, if ever, do I do anything reckless.	.71	.71
12. Emotion seldom causes me to act without thinking.	.41	.66
13. I have a reserved and cautious attitude toward life.	.32	.46
14. My thinking is usually careful and purposeful.	.36	.60
15. I am not one of those people who blurt things out without thinking.	.46	.61
16. I generally rely on careful reasoning in making up my mind.	.31	.45

A nice property of the concept of relative measurement invariance is that it is possible to do a quick scan of a scale to see whether it may contain items for relative measurement - which is difficult with absolute measurement invariance. The reason for this is that the restrictions of relative measurement invariance imply that the p-values of relative items are equal across groups. So, if a scale consists of a number of items with unequal p-values across groups, but there is also a set of items with equal p-values, this may indicate that the items with equal p-values are items for relative measurement of the trait in question.

We found several scales in the PRF-E that showed this pattern, but it is most pronounced in the subscale 'impulsivity'. Since this analysis is a mere illustration of some of the ideas presented in this paper, we limit our analysis to this scale. The pattern of p-values for males and females is shown in Table 8.1.

These results suggest the existence of relative and absolute items in the scale. The items 1, 4, 5, 6, 7, 10, and 11 show almost identical p-values for males and females, which may indicate relative measurement. On the other hand, items 2, 3, 8, 9, 12, 13, 14, 15, and 16 show higher p-values for females, which may indicate a sex difference in latent trait distributions - females being more impulsive. We can check this by applying the logistic regression procedure. The items hypothesized to be items for absolute measurement are combined in a subscale, generating an absolute total score. This sumscore is then standardized within each group to generate a relative score. Subsequently, the amount of DIF for each item is evaluated with respect to the absolute score (to detect absolute DIF), and the relative score (to detect relative DIF) by assessing the effect of sex on the item response. The results yielded by this procedure are reported in Table 8.2. Only the results concerning uniform DIF are reported, since none of the items showed non-uniform absolute or relative DIF.

Table 8.2. Standardized parameter estimates for the effect of sex in the logistic regression procedure. A positive parameter estimate indicates that females have a higher probablility of an affirmative answer, conditional on their absolute/relative score.

Item	Absolute bias	Relative bias
1	-1.33	-0.29
2	-0.59	3.04
3	0.82	3.48
4	-3.15	0.43
5	-1.95	0.58
6	-2.12	0.63
7	-2.83	-0.13
8	-0.21	2.45
9	0.65	2.85
10	-1.85	0.63
11	-2.13	-0.05
12	10.03	5.65
13	-0.55	2.85
14	1.37	5.56
15	-0.65	3.18
16	-0.79	3.13

The results are in line with the initial hypothesis. The items that we hypothesized to be items for relative measurement conform to the idea that they measure relative to the other items in the scale, consistently showing absolute but no relative DIF. An exception is item 1, showing neither absolute nor relative DIF. The theoretical impossibility of such a result, given the difference in absolute score distributions, implies that this is due to a lack of power. The absolute items also behave as expected, consistently showing relative DIF but no absolute DIF, except for item 12. This item shows both absolute and relative DIF - presumably caused by the explicit use of the word 'emotion' - and should probably be removed.

Of course, these results should be interpreted with some caution; although the items do behave as relative items, inspection of the content of the items does not yield obvious reasons why this should be so. Further research should give more insight into the item features that trigger relative measurement. A research strategy that could give some insight in the response processes involved would be to present the items with and without explicit instructions for comparison. So, items could be administered with the instruction to compare oneself to a fixed reference group (e.g., males), with the instruction to compare oneself to a variable reference group (e.g., the group to which one belongs), and without any instruction at all. Comparing ICC's across these situations should provide information on the relevant response processes, which would in turn strengthen the validity of this and other personality scales.

8.6 Illustration 2

As mentioned, the concepts of absolute and relative measurement invariance generalize to other latent variable models such as the congeneric model often used in SEM. To illustrate the approach for the SEM model, we use a subset of data collected by Rodriguez Mosquera, Manstead, & Fischer (2000). They constructed scales to measure several types of honor concerns. A total of 61 male and 61 female Dutch undergraduate psychology students completed the scales.

Table 8.3. Means and standard deviations for males (n=61) and females (n=61) on items in the scale for honor concerns.

Item: 'How bad would you feel if the following description applied to you?'	Mean (SD) for males	Mean (SD) for females
1. Wearing provocative clothes	2.10 (1.14)	2.16 (1.39)
2. Sleeping with someone without starting a serious relationship with that person	2.48 (1.63)	2.77 (1.42)
3. Changing partner often	2.82 (1.38)	3.44 (1.32)
4. Being known as having different sexual contacts	2.77 (1.57)	3.90 (1.38)

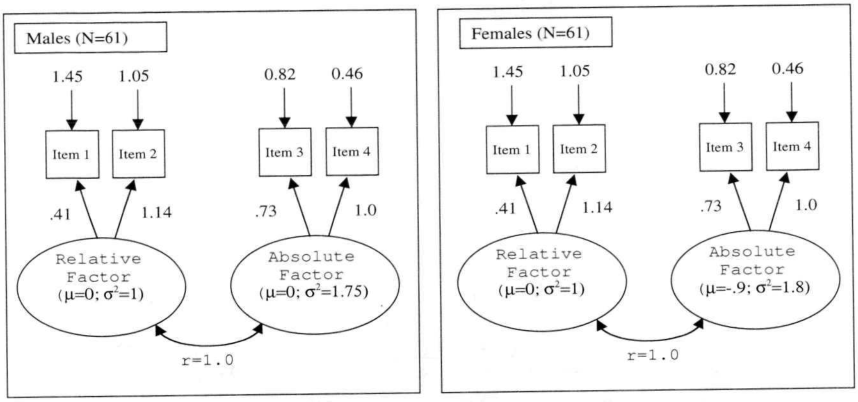
We analyze a subset of items of a scale called 'feminine honor concerns' and evaluate measurement invariance with respect to sex. The items request the participant to rate, on a 7-point scale, how bad he/she would feel if the descriptions given in the items applied to him/her. The content of the items is given in Table 9.3 along with the means and standard deviations for both sexes.

We fitted a unidimensional model with strict factorial invariance constraints across groups (Meredith, 1993) to test for measurement invariance. Although the model cannot be rejected ($\chi^2(14) = 21.77$; p = .08), overall fit is not satisfactory (RMSEA = .08), and inspection of modification indices suggests the presence of DIF. Likelihood ratio tests, conducted by individually freeing intercept parameters, reveal uniform bias for item 2 ($\chi^2(1) = 6.85; p < .05$) and for item 4 ($\chi^2(1) = 7.00; p < .05$).

As can be seen from Table 3, however, the observed means of items 1 and 2 are almost equal across groups. The content of the items, 'wearing provocative clothes' and 'sleeping with someone without starting a serious relationship with that person' suggest that these items may be interpreted differently by men and women. It is not implausible that subjects interpret the content of the items conditional on their sex.

If this is the case, it implies that these items may be treated as relative, within-group indicators. We fitted a model specifying these items as relative indicators using the SEM procedure described in the previous section (see the Appendix for the technical details). A graphical representation of the model is given in Figure 8.4.

Figure 9.4. A structural equation model for relative measurement. Items 1 and 2 load on the relative factor, and items 3 and 4 on the absolute factor. The relative factor is obtained by standardization within groups, and correlates unity with the absolute factor.



The model cannot be rejected ($\chi^2(14) = 15.84; p = .32$) and fits the data very well (RMSEA $< .01$). In accordance with these results, inspection of modification indices does not reveal substantial misfit anywhere in the model. Given the fact that the number of parameters in the model is equal to the number of parameters in the absolute model with strict factorial invariance, the better fit of the relative model suggests that this model should be preferred. This may indicate that items 1 and 2 do indeed measure relative to items 3 and 4, which may teach us more about the structure of honor concerns in male and female populations. This, in turn, may provide valuable information for theory development in this area.

8.7 Discussion

The theory and research presented in this paper provide some insight into the complicated relation between measurement invariance and construct validity. It has been argued that not all items that show DIF in the classical sense are invalid. Rather, a failure to distinguish between absolute and relative forms of measurement will lead to apparent bias of items for relative measurement. Items for relative measurement can be valid indicators of a trait within groups, but because of their relative nature, these items are bound to produce bias as defined in the classical sense. If the relative nature of the items is recognized, they do not have to be eliminated from a test. Instead, they can be used as relative indicators of the trait in question.

The distinction between absolute and relative measurement has some implications for the theory of measurement invariance and bias. If the latent trait distribution differs across groups, an item will show either absolute bias, relative bias, or both: Absolute measurement invariance and relative measurement invariance cannot simultaneously hold, unless the trait distributions are identical. If the trait distributions differ, relative measurement invariance of a given item will cause that item to show absolute bias. Bias in the classical sense can therefore result from relative measurement invariance. This is an intriguing result because it contradicts the view that all bias results from multidimensionality.

The relation between bias and multidimensionality should be constructed as follows. Bias is a group difference in the distribution of item responses conditional on the latent trait. Multidimensionality is a possible explanation for the presence of bias. Now, it is sometimes suggested that bias is multidimensionality because a biased item 'measures' group membership in addition to the variable of interest. So, in a very general sense, group membership is then conceived of as the second dimension. This line of reasoning may be maintained, but in this case multidimensionality is no longer an explanation of item bias: such an explanation would be circular because the group difference is exactly the phenomenon that requires an explanation. Thus, in this line of reasoning, all bias is multidimensionality, all multidimensionality is bias, and there does not seem to be a good reason for entertaining two words for the same concept. As a consequence, either of the terms should be dropped from the psychometric vocabulary. We do not endorse such a point of view, and take the relation between multidimensionality and bias to be of an explanatory nature. This implies that the second variable that the item measures in addition to the intended trait must be a variable that is distinct from group membership, although it must in some way be related to group membership (otherwise the variable could not influence the item responses differentially). The most sophisticated theory of the relation between this second variable and bias is the theory presented in Shealy & Stout (1993). Shealy & Stout show that a second variable could produce bias, if the groups differ in the distribution on this variable. In their theory of multidimensionality, group differences in the distribution of the second trait are therefore a necessary condition for bias to occur (Shealy & Stout, 1993, p.209 ff.). In other words, there has to be some association between this second trait and group membership. However, this is obviously not the case in relative measurement, because the distribution on a relative latent variable will often not be associated with group membership - for example, if the absolute trait distributions are normal. In view of this problem, there are two ways to proceed. Either Shealy & Stout's theory has to be revised in order to accommodate for the relative position on the measured variable as a second dimension producing bias, or we have to conclude that relative measurement does not imply multidimensionality. The first of these options requires that we consider, for example, absolute height and relative height to be two different traits. In our view this would render the concept of multidimensionality rather trivial. We therefore take the second option and submit that relative measurement invariance does not imply multidimensionality, but unidimensional measurement of the intended trait within groups. We conclude that not all bias results from multidimensionality.

A failure to recognize the fact that items provide relative measurement may produce distortions in the interpretation of data. For instance, in personality research, researchers obviously assume that the items in a personality scale are items for absolute measurement. This assumption is, however, not self-evident. If the assumption is not fulfilled, this may lead to incorrect conclusions regarding personality differences between groups. This is a direct result from the fact that absence of impact cannot be distinguished from relative measurement invariance without a substantial number of absolute items or a separate criterion. Consider, for example, an assertiveness scale in which most or all items are actually items for relative measurement (i.e., the item responses result from an explicit or implicit comparison of subjects with other members of a relevant group). A psychologist obtains responses from American and Japanese subjects. Suppose that the American subjects are in fact more assertive than the Japanese. What would happen if he started looking for an effect of nationality on assertiveness? He would never find any, since both groups answer the items by comparing themselves to their own reference group, which automatically results in comparable mean scores on the test. This is not an academic point, because virtually nothing is known about the cognitive processes involved in responses to personality items. Whether this kind of distortion occurs, and if so, how grave its consequences are, is of course a question for empirical research. Nevertheless, research in this area may profit from taking the relative nature of items in personality scales into account. An interesting line of research would consist in assessing absolute and relative measurement invariance of personality items with respect to a behaviorally inspired matching criterion. Such research, of course, requires the evaluation of tests at the item level. In this respect, the advantages of the generalized item response theory models (Mellenbergh, 1994) over classical test theory cannot be overemphasized.

The concept of relative measurement invariance could further be applied in a range of other situations. One could, for example, think of cross-cultural research into subjective well-being or happiness: It is not unlikely that people, in responding to items used in these scales, compare themselves to other people in their environment. I may consider myself depressed compared to the people around me, but if I get really depressed and I am admitted for hospitalization, I may consider myself rather happy compared to the people surrounding me there. Concepts such as satisfaction and happiness do seem to have an inherently relative component, and are therefore susceptible to relative measurement.

In sum, items with relative measurement invariance but absolute bias are not multidimensional and may be valid within-group indicators of the construct to be measured. Also, the fact that such items occur may lead to theory formation on item response processes outside the cognitive realm. The question then becomes what the practical implications of these findings are, and how they could be of help in practical situations. Should we drastically change the way we make personality tests? Should we be telling subjects not to make within group comparisons? In our opinion, no definitive answers to these questions can, at present, be given. How often the response processes we outlined occur, and which item features and person characteristics trigger these processes, are questions open to empirical research. Obviously, however, the relation between item response models and item response processes is not clear in domains outside cognitive testing. Within the field of cognitive testing, there is at least a raw image of the response processes that lead to item responses, and to a certain extent these processes have been successfully modeled (see Embretson, 1994, for a good example). Retaining items with relative measurement invariance in cognitive tests does not seem to be a very good idea, for there is little theoretical foundation for such practice. In fact, selecting items with relative measurement would technically be comparable to the item selection rules specified in the Golden Rule Settlement (McAllister, 1993), where the Educational Testing Service agreed to construct tests by giving priority to items showing the least differences between groups. Most psychometricians would agree that this was not a psychometrically sound basis for item selection, because it was based on the presumption that all group differences in the performance on these tests reflect bias. The main reason why retaining items with absolute bias in cognitive tests is not a very good idea, however, is precisely because relative measurement invariance conflicts with the construct definitions. Indeed, if one approaches such items from the perspective of cognitive processes in problem solving, the nature of these processes suggests, or even prescribes, that absolute measurement invariance should hold. This is in sharp contrast with construct definitions and response processes outside the realm of cognitive testing. In fact, we find it somewhat disturbing that the demands of measurement invariance are often generalized to the measurement of personality traits and attitudes, while this paper clearly shows how a rather simple, and not implausible, response process would destroy measurement invariance in the classical sense. Coupled with the fact that, in many research areas, there is very little theory on what happens between item administration and item response, relative measurement invariance may be an important concept, although we cannot, at present, determine its scope or usefulness in practical situations. However, we can safely conclude that the relation between construct validity and measurement invariance is rather intricate, since items without measurement invariance may very well be valid indicators of the construct in question. Therefore, the relation between measurement invariance and construct validity needs to be reconsidered, and theory formation on this subject is called for. Especially, the need to extend the work of Embretson (1994), on the relation between cognitive theories on response processes and latent trait models, to fields other than cognitive testing, seems pertinent.

8.8 Appendix

Evaluating measurement invariance for continuous item responses requires testing the strict factorial invariance model of Meredith (1993). This model involves the modeling of mean structures through multigroup analysis as introduced by Sörbom (1974). Strict factorial invariance with respect to a selection variable V (here we take V to indicate group) holds if

$$\mathbf{y_v} = \tau + \mathbf{\Lambda}\alpha_{\mathbf{v}} \tag{8.19}$$

and

$$\mathbf{\Sigma_v} = \mathbf{\Lambda}\mathbf{\Phi_v}\mathbf{\Lambda}' + \mathbf{\Delta} \tag{8.20}$$

where the $\mathbf{y_v}$ is a vector of means on observed variables in group v; τ is a vector of intercepts; $\mathbf{\Lambda}$ is a matrix of factor loadings; $\alpha_{\mathbf{v}}$ is the vector of factor means for group v; $\mathbf{\Sigma_v}$ is the covariance matrix of the observed variables in group v; $\mathbf{\Phi_v}$ is the covariance matrix of the factors in group v; and $\mathbf{\Delta}$ is a diagonal matrix containing the variances of the residuals. The strict factorial invariance model specifies that only the factor means and variances may differ between groups. Meredith (1993) has shown that strict factorial invariance with respect to V implies weak measurement invariance with respect to V. These conditions can be weakened to conditions of strong factorial invariance by allowing the matrix $\mathbf{\Delta}$ to vary over groups. Strong factorial invariance, however, no longer implies weak measurement invariance, so that we limit our attention to the strict factorial invariance model.

We take a simple unidimensional model with one latent variable as the point of departure. This renders $\alpha_{\mathbf{v}}$ and $\mathbf{\Phi_v}$ scalars. The model is identified by setting one of the elements in $\mathbf{\Lambda}$ to one and the factor mean $\alpha_{\mathbf{v}}$ to zero in one of the groups; $\alpha_{\mathbf{v}}$ is free to vary in the other groups. This is the unidimensional model fitted to the data in Illustration 2. We modify the model to cope with relative items as follows. Partition the observed variables into a set of absolute items and a set of relative items. For the absolute items, the strict factorial invariance model is maintained as above. We term the original single factor the *absolute factor.* For the relative items we invoke a new factor, so that $\alpha_{\mathbf{v}}$ is now a 1 x 2 vector and $\mathbf{\Phi_v}$ a 2 x 2 symmetric matrix. The relative items are allowed to load only on this second factor. We term this factor the *relative factor.* To ensure that the relative factor is the within-group standardized variant of the absolute factor we add the following restrictions to the model. First, we require that the relative factor has the standard normal distribution in each group. Second, we require that the relative factor correlates unity with the absolute factor within each of the groups. This gives the restrictions

$$\alpha = \begin{bmatrix}\alpha_v & 0\end{bmatrix} \tag{8.21}$$

and

$$\mathbf{\Phi_v} = \begin{bmatrix}\phi_v & \\ \sqrt{\phi_v} & 1\end{bmatrix}. \tag{8.22}$$

Equation 8.22 is a nonlinear restriction that can be readily implemented in LISREL (Jöreskog & Sörbom, 1993). However, admissability checks should be turned off because $\mathbf{\Phi}$ is not positive definite. The restriction ensures that the correlation r_{12} between the absolute and the relative factor is equal to $r_{12} = \phi_{12}/\sqrt{\phi_1 \times \phi_2} = \sqrt{\phi_1}/\sqrt{\phi_1 \times 1} = 1$, as required. For the relative items these constrains imply the following equations:

$$\mathbf{y_v} = \tau \tag{8.23}$$

and

$$\mathbf{\Sigma_v} = \mathbf{\Lambda\Lambda}' + \mathbf{\Delta}. \tag{8.24}$$

Thus, the vector of means as well as the covariance matrix are invariant over groups for the relative items. This parallels the dichotomous case where relative measurement implies that p-values for relative items are invariant over groups. Because for the absolute items in the model the original equations 8.19 and 8.20 hold, the part of the covariance matrix containing covariances between absolute and relative items looks as follows: For an absolute item j and a relative item k, the element σ_{jk} of $\mathbf{\Sigma_v}$ is equal to $\lambda_j \lambda_k \sqrt{\phi_v}$. The model with restrictions 8.21 and 8.22 is the relative model fitted to the data in Illustration 2.

9. REFERENCES

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38*, 123-140.

Bartholomew, D. J. (1987). *Latent variable models and factor analysis.* London: Griffin.

Batitsky, V. (1998). Empiricism and the myth of fundamental measurement. *Synthese, 116*, 51-73.

Bechtold, H. P. (1959). Construct validity: A critique. *American Psychologist, 14*, 619-629.

Bentler, P. M. (1982). Linear systems with multiple levels and types of latent variables. In K. G. Jöreskog & H. Wold (Eds.), *Systems under indirect observation* (pp. 101-130). Amsterdam: North Holland.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51.

Bollen, K. A., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110*, 305-314.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology, 53*, 605-634.

Bollen, K. A., & Ting, K. (2000). A tetrad test for causal indicators. *Psychological Methods, 5*, 3-22.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the social sciences.* Mahwah, NJ: Lawrence Erlbaum Associates.

Borkenau, P., & Ostendorf, F. (1998). The big five as states: how useful is the five factor model to describe intraindividual variations over time? *Journal of research in personality, 32*, 202-221.

Borsboom, D. (2002). The structure of the DSM. *Archives of General Psychiatry, 59*, 569-570.

Borsboom, D., Van Heerden, J., & Mellenbergh, G. J. (*in press*). *Validity and truth.* Proceedings of the International Meeting of the Psychometric Society in Osaka.

Borsboom, D., & Mellenbergh, G. J. (2002). True scores, latent variables, and constructs: A comment on Schmidt and Hunter. *Intelligence, 30*, 505-514.

Borsboom, D., & Mellenbergh, G. J. (*submitted*). Why psychometrics is not pathological: A comment on Michell.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2002-a). Functional thought experiments. *Synthese, 130*, 379-387.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2002-b). Different kinds of DIF: A distinction between absolute and relative forms of measurement invariance and bias. *Applied Psychological Measurement, 26*, 433-450.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (*in press*). The theoretical status of latent variables. *Psychological Review.*

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (*submitted*). The problem of validity.

Brennan, R. L. (2001). An essay on the history and future of reliability. *Journal of Educational Measurement, 38*, 295-317.

Bridgman, P. W. (1927). *The logic of modern physics.* New York: Macmillan.

Brogden, H. E. (1977). The Rasch model, the law of comparative judgment, and additive conjoint measurement. *Psychometrika, 42*, 631-634.

Brown, J. R. (1991). *The laboratory of the mind: Thought experiments in the natural sciences.* London: Routledge.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research, 21*, 230-258.

Bunzl, M. (1996). The logic of thought experiments. *Synthese, 106*, 227-240.

Cacioppo, J. T., & Berntson, G. G. (1999). The affect system: Architecture and operating characteristics. *Current Directions in Psychological Science, 8*, 133-137.

Campbell, D. T. (1960). Recommendations for APA test standards regarding construct, trait, or discriminant validity. *American Psychologist, 15*, 81-105.

Campbell, N. R. (1920). *Physics, the elements.* Cambridge: Cambridge University Press.

Cartwright, N. (1983). *How the laws of physics lie.* Oxford: Clarendon Press.

Cattell, R. B. (1946). *Description and measurement of personality.* New York: World Book Company.

Cattell, R. B., & Cross, K. (1952). Comparisons of the ergic and self-sentiment structures found in dynamic traits by R- and P-techniques. *Journal of Personality, 21*, 250-271.

Cervone, D. (1997). Social-cognitive mechanisms and personality coherence: Self-knowledge, situational beliefs, and cross-situational coherence in perceived self-efficacy. *Psychological Science, 8*, 43-50.

Cliff, N. (1992). Abstract measurement theory and the revolution that never happened. *Psychological Science, 3*, 186-190.

Coombs, C. (1964). *A theory of data.* New York: Wiley.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12*, 671-684.

Cronbach, L. J. (1980). *Validity on Parole: How can we go straight. New directions for testing and measurement: Measuring achievement over a decade.* Paper presented at the Proceedings of the 1979 ETS Invitational Conference, San Fransisco.

Cronbach, L. J. (1988). Five perspectives on validation argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, New Jersey: Erlbaum.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: The theory of generalizability.* New York: Wiley.

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

Cudeck, R., & Browne, M. W. (1983). Cross validiation of covariance structures. *Multivariate Behavioral Research, 18*, 147-167.

De Finetti, B. (1974). *Theory of Probability (Vol. 1).* New York: Wiley.

De Groot, A. D. (1961). *Methodologie.* 's Gravenhage: Mouton.

Dennett, D. C. (1991). *Consciousness explained.* Boston: Little Brown.

Devitt, M. (1991). *Realism and truth (2nd edition).* Cambridge: Blackwell.

Dolan, C. V., Jansen, B., & Van der Maas, H. (*submitted*). Constrained and unconstrained normal finite mixture modeling of multivariate conservation data.

Ebel, R. L. (1956). Must all tests be valid? *American Psychologist, 16*, 640-647.

Edgeworth, F. Y. (1888). The statistics of examinations. *Journal of the Royal Statistical Society, 51*, 598-635.

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods, 5*, 155-174.

Ellis, B. B., Becker, P., & Kimmel, H. D. (1993). An item response theory evaluation of an English version of the Trier Personality Inventory. *Journal of Cross Cultural Psychology, 24*, 133-148.

Ellis, J. L. (1994). *Foundations of monotone latent variable models.* Unpublished doctoral dissertation.

Ellis, J. L., & Van den Wollenberg, A. L. (1993). Local homogeneity in latent trait models: A characterization of the homogeneous monotone IRT model. *Psychometrika, 58*, 417-429.

Ellis, M. V., & Blustein, D. L. (1991). The unificationist view: A context for validity. *Journal of Counseling and Deveopment, 69*, 561-563.

Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179-197.

Embretson, S. (1994). Applications of cognitive design systems to test development. In C. R. Reynolds (Ed.), *Cognitive assessment: A multidisciplinary perspective* (pp. 107-135). New York: Plenum Press.

Epstein, S. (1994). Trait theory as personality theory: Can a part be as great as the whole? *Psychological Inquiry, 5*, 120-122.

Feldman, L. A. (1995). Valence focus and arousal focus: Individual differences in the structure of affective experience. *Journal of personality and social psychology, 69*, 153-166.

Fine, T. L. (1973). *Theories of probability.* New York: Academic Press.

Fischer, G. (1995). Derivations of the Rasch model. In G. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 15-38). New York: Springer.

Fischer, G. H., & Parzer, P. (1991). An extension of the rating scale model with an application to the measurement of change. *Psychometrika, 56*, 637-651.

Fisher, R. A. (1925). *Statistical methods for research workers.* London: Oliver and Boyd.

Foucault, M. (1970). *The order of things: An archeology of the human sciences.* New York: Pantheon.

Frege, G. (1952/1892). On sense and reference. In P. Geach & M. Black (Eds.), *Translations of the philosophical writings of Gottlob Frege.* Oxford: Blackwell.

Gaito, J. (1980). Measurement scales and statistics: Resurgence of an old misconception. *Psychological Bulletin, 87,* 564-567.

Gigerenzer, G. (1993). The Superego, the Ego, and the Id in statistical reasoning. In G. Keren & L. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: methodological issues.* Hillsdale: Lawrence Erlbaum Associates.

Glymour, C. (1980). *Theory and evidence.* Princeton, NJ: Princeton University Press.

Glymour, C. (2001). *The mind's arrows.* Cambridge, Massachusets: MIT Press.

Goldstein, H., & Wood, R. (1989). Five decades of item response modelling. *British Journal of Mathematical and Statistical Psychology, 42,* 139-167.

Goodman. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika, 61,* 215-231.

Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement, 6,* 427-439.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10,* 255-282.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stoufer, L. Guttman, E. A. Suchman, P. L. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Studies in social psychology in World War II: Vol. IV. Measurement and prediction* (pp. 60-90). Princeton, NJ: Princeton University Press.

Hacking, I. (1965). *Logic of statistical inference.* Cambridge: Cambridge Univeristy Press.

Hacking, I. (1983). *Representing and intervening.* Cambridge: Cambridge University Press.

Hacking, I. (1990). *The taming of chance.* Cambridge: Cambridge University press.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principles and applications.* Boston: Kluwer-Nijhoff.

Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika, 62,* 331-347.

Hempel, C. G. (1962). Deductive-nomological vs. statistical explanation. In H. Feigl & G. Maxwell (Eds.), *Minnesota Studies in the Philosophy of Science, Vol. 3: Scientific explanation, space, and time* (pp. 98-169). Minneapolis: University of Minnesota Press.

Hempel, C. G. (1965). *Aspects of scientific explanation.* New York: The Free Press.

Hershberger, S. L. (1994). The specification of equivalent models before the collection of data. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis.*

Thousand Oaks: Sage.

Hintikka, J. (2001). Post-Tarskian truth. *Synthese*, *126*, 17-36.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*, 945-959.

Holland, P. W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika*, *55*, 577-601.

Horowitz, T., & Massey, G. J. (Eds.). (1991). *Thought experiments in science and philosophy.* Savage, MD: Rowman and Littlefield Publishers, Inc.

Huang, D. C., Church, T. A., & Katigbak, M. S. (1997). Identifying cultural differences in items and traits: Differential item functioning in the NEO Personality Inventory. *Journal of Cross-Cultural Psychology*, *28*, 192-218.

Jackson, D. N. (1974). *Personality Research Form-E.* London: Research Psychologist Press.

Jackson, P. H., & Agunwamba, C. C. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: I. Algebraic lower bounds. *Psychometrika*, *42*, 567-578.

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

Jensen, A. R. (1998). *The g factor: The science of mental abilities.* Westport, CN: Praeger.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, *36*, 109-133.

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8 User's reference guide.* Chicago: Scientific Software International, Inc.

Judd, C. M., Smith, E. R., & Kidder, L. H. (1991). *Research methods in social relations.* Fort Worth: Harcourt Brace Jovanovich College Publishers.

Kaiser, H. A. (1960). Book review of V.L. Senders, Measurement and Statistics. *Psychometrika*, *25*, 411-413.

Kane, M. T. (1992). An argument based approach to validity. *Psychological Bulletin*, *112*, 527-535.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*, 319-342.

Kelley, T. L. (1927). *Interpretation of educational measurements.* New York: Macmillan.

Kelly, K. T. (1996). *The logic of reliable inquiry.* New York: Oxford University Press.

Klein, D. F., & Cleary, T. A. (1967). Platonic true scores and error in psychiatric rating scales. *Psychological Bulletin*, *68*, 77-80.

Kline, P. (1998). *The new psychometrics: Science, psychology, and measurement.* London: Routledge.

Kok, F. (1988). Item bias and test multidimensionality. In R. Langeheine & J. Rost (Eds.), *Latent trait and latent class models.* New York: Plenum Press.

Kolmogorov, A. (1933). *Grundbegriffe der Warscheinlichkeitsrechnung.* Berlin: Springer.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement, Vol. I.* New York: Academic Press.

Kripke, S. A. (1972). *Naming and necessity.* Oxford: Blackwell.

Krueger, R. F. (1999). The structure of common mental disorders. *Archives of General Psychiatry, 56*, 921-926.

Kuhn, T. S. (1977). *The essential tension.* Chicago: The University of Chigago Press.

Lakatos, I. (1978). *The methodology of scientific research programmes.* Cambridge: Cambridge University Press.

Lamiell, J. T. (1987). *The psychology of personality: An epistemological inquiry.* New York: Columbia University Press.

Latour, B. (1987). *Science in action.* Cambridge: Harvard University Press.

Laudan, L. (1977). *Progress and its problems.* Berkeley: University of California Press.

Lawley, D. N. (1943). On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh, 62*, 74-82.

Lawley, D. N., & Maxwell, A. E. (1963). *Factor analysis as a statistical method.* London: Butterworth.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. A. Stoufer, L. Guttman, E. A. Suchman, P. L. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Studies in social psychology in World War II: Vol. IV. Measurement and prediction* (pp. 362-412). Princeton, NJ: Princeton University Press.

Lazarsfeld, P. F. (1959). Latent structure analysis. In S. Koch (Ed.), *Psychology: A study of a science.* New York: McGraw-Hill.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis.* Boston: Houghton Mifflin.

Lee, P. M. (1997). *Bayesian statistics: An introduction.* New York: Wiley.

Levy, P. (1969). Platonic true scores and rating scales: A case of uncorrelated definitions. *Psychological Bulletin, 71*, 276-277.

Lewis, D. (1973). *Counterfactuals.* Oxford: Blackwell.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*, 635-694.

Lord, F. M. (1952). *A theory of test scores.* New York: Psychometric Society.

Lord, F. M. (1953). On the statistical treatment of football numbers. *American Psychologist, 8*, 260-261.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, New Jersey: Erlbaum.

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

Luce, R. D. (1996). The ongoing dialog between empirical science and measurement theory. *Journal of Mathematical Psychology, 40*, 78-95.

Luce, R. D. (1997). Several unresolved conceptual problems of mathematical psychology. *Journal of Mathematical Psychology, 41*, 79-87.

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology, 1*, 1-27.

Lumsden, J. (1976). Test theory. *Annual Review of Psychology, 27*, 251-280.

Mach, E. (1905/1976). *Knowledge and Error*. Dordrecht: Reidel.

Maraun, M. D. (1999). Measurement as normative practice: Implications of Wittgenstein's philosophy for measurement in psychology. *Theory & Psychology, 8*, 435-461.

Markus, K. A. (1998). Science, measurement, and validity: Is completion of Samuel Messick's synthesis possible? *Social Indicators Research, 45*, 7-34.

Maxwell, G. (1962). The ontological status of theoretical entities. In H. Feigl & G. Maxwell (Eds.), *Minnesota studies in the philosophy of science, Vol 3: Scientific explanation, space, and time* (pp. 3-28). Minneapolis: University of Minnesota Press.

McAllister, P. H. (1993). Testing, DIF, and public policy. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning*. Hillsdale, NJ: Erlbaum.

McArdle, J. J. (1987). Latent growth curve models within developmental structural equation models. *Child Development, 58*, 110-133.

McCrae, R. R., & Costa, P. T. (1997). Personality trait structure as a human universal. *American Psychologist, 52*, 509-516.

McCrae, R. R., & John, O. P. (1992). An introduction to the five factor model and its applications. *Journal of Personality, 60*, 175-215.

McCullagh, P., & Nelder, J. (1989). *Generalized linear models*. London: Chapman & Hall.

McDonald, R. P. (1982). Linear versus nonlinear models in item response theory. *Applied Psychological Measurement, 6*, 379-396.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates.

McDonald, R. P., & Marsh, H. W. (1990). Choosing a multivariate model: Noncentrality and goodness of fit. *Psychological Bulletin, 107*, 247-255.

McGuiness, B. (Ed.). (1976). *Ludwig Boltzmann. Theoretical physics and philosophical problems*. Dordrecht: Reidel.

Meehl, P. E. (1978). Theoretical risks and tabular aterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*, 806-834.

Meiland, J. W. (1977). Concepts of relative truth. *Monist, 60*, 568-582.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13*, 127-143.

Mellenbergh, G. J. (1994). Generalized Linear Item Response Theory. *Psychological Bulletin, 115*, 300-307.

Mellenbergh, G. J. (1994). A unidimensional latent trait model for continuous item responses. *Multivariate Behavioral Research, 19*, 223-236.

Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods, 1*, 293-299.

Mellenbergh, G. J. (1999). Measurement models. In H. J. Adèr & G. J. Mellenbergh (Eds.), *Research methodology in the social, life, and behavioural sciences.*

London: Sage Publications.

Mellenbergh, G. J., & Van den Brink, W. P. (1998). The measurement of individual change. *Psychological methods, 3*, 470-485.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika, 58*, 525-543.

Messick, S. (1981). Constructs and their vissicitudes in educational and psychological measurement. *Psychological Bulletin, 89*, 575-588.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (pp. 13-103). Washington, DC: American Council on Education and National Council on Measurement in Education.

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research, 45*, 35-44.

Michell, J. (1986). Measurement scales and statistics: A clash of para-digms. *Psychological Bulletin, 100*, 398-407.

Michell, J. (1990). *An introduction to the logic of psychological measurement.* Hillsdale, NJ: Erlbaum.

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

Michell, J. (2000). Normal science, pathological science, and psychometrics. *Theory and Psychology, 10*, 639-667.

Michell, J. (2001). Measurement Theory: History and Philosophy. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social and behavioral sciences*: Elsevier Science.

Mill, J. S. (1843). *A system of logic.* London: Oxford University Press.

Millsap, R. E. (1997). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods, 2*, 248-260.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing bias. *Applied Psychological Measurement, 17*, 297-334.

Mischel, W. (1968). *Personality and assessment.* New York: Wiley.

Mischel, W. (1973). Toward a social cognitive learning reconceptualization of personality. *Psychological Review, 80*, 252-283.

Mischel, W., & Shoda, Y. (1998). Reconciling processing dynamics and personality dispositions. *Annual review of psychology, 49*, 229-258.

Mokken, R. J. (1970). *A theory and procedure of scale analysis.* The Hague: Mouton.

Molenaar, P. C. M. (1985). A dynamic factor model for the analysis of multivariate time series. *Psychometrika, 50*, 181-202.

Molenaar, P. C. M. (1999). Longitudinal analysis. In H. J. Adèr & G. J. Mellenbergh (Eds.), *Research methodology in the social, life, and behavioural sciences.* Thousand oaks: Sage.

Molenaar, P. C. M., & Von Eye, A. (1994). On the arbitrary nature of latent variables. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis.* Thousand Oaks: Sage.

Molenaar, P. C. M., Huizenga, H. M., & Nesselroade, J. R. (in press). The relationship between the structure of inter-individual and intra-individual variability: A theoretical and empirical vindication of developmental systems theory. In

U. M. Staudinger & U. Lindenberger (Eds.), *Understanding human development.* Dordrecht: Kluwer.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62*, 229-258.

Moustaki, I., & Knott, M. (2000). Generalized latent trait models. *Psychometrika, 65*, 391-411.

Muthén, L. K., & Muthén, B. O. (1998). *Mplus User's Guide.* Los Angeles, CA.

Nagel, E. (1939). *Principles of the theory of probability.* Chicago: The University of Chicago Press.

Nagel, E. (1961). *The structure of science.* London: Routledge & Kegan Paul.

Narens, L., & Luce, R. D. (1986). Measurement: The theory of numerical assignments. *Psychological Bulletin, 99*, 166-180.

Neale, M. C., Boker, S. M., Xie, G., & Maes, H. H. (1999). *Mx: Statistical Modeling.* (5th ed.). Richmond, VA: Department of Psychiatry.

Neisser, U., Boodoo, G., Bouchard, T. J., Boykin, A., Brody, N., Ceci, S. J., Halpern, D. F., Loehlin, J. C., Perloff, R., Sternberg, R. J., & Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51*, 77-101.

Neyman, J., & Pearson, E. S. (1967). *Joint statistical papers.* London: Cambridge University Press.

Norton, J. (1991). Thought experiments in Einstein's work. In T. Horowitz & G. J. Massey (Eds.), *Thought experiments in science and philosophy.* Savage, MD: Rowman and Littlefield Publishers, Inc.

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3*, 1-18.

Novick, M. R., & Jackson, P. H. (1974). *Statistical methods for educational and psychological research.* New York: McGraw-Hill.

Novick, M. R., Jackson, P. H., & Thayer, D. T. (1971). Bayesian inference and the classical test theory model: Reliability and true scores. *Psychometrika, 36*, 261-288.

O'Connor, D. J. (1975). *The correspondence theory of truth.* London: Hutchinson University Library.

Pearl, J. (1999). Graphs, causality, and structural equation models. In H. J. Adèr & G. J. Mellenbergh (Eds.), *Research methodology in the social, behavioural, and life sciences.* Thousand Oaks: Sage.

Pearl, J. (2000). *Causality: Models, reasoning, and inference.* Cambridge: Cambridge University Press.

Perline, R., Wright, B. D., & Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement, 3*, 237-255.

Pervin, L. A. (1994). A critical analysis of current trait theory (with commentaries). *Psychological Inquiry, 5*, 103-178.

Popham, W. J. (1997). Consequential validity: Right concern-wrong concept. *Educational Measurement: Issues and Practice, 16*, 9-13.

Popper, K. R. (1959). *The logic of scientific discovery.* London: Hutchinson Education.

Popper, K. R. (1963). *Conjectures and refutations.* London: Routledge and Kegan Paul.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Paedagogiske Institut.

Reese, T. W. (1943). The application of the theory of physical measurement to the measurement of psychological magnitudes, with three experimental examples. *Psychological Monographs, 55,* 6-20.

Reichenbach, H. J. (1938). *Experience and prediction.* Chicago: University of Chicago Press.

Reichenbach, H. J. (1956). *The direction of time.* Berkeley: University of California Press.

Rènyi, A. (1970). *Foundations of probability.* San Francisco: Holden-Day.

Rodriguez Mosquera, P. M., Manstead, A. S. R., & Fischer, A. (2002). The role of honor concerns in emotional reactions to offenses. *Cognition and Emotion, 16,* 143-163.

Rorer, L. G. (1990). Personality assessment: A conceptual survey. In L. A. Pervin (Ed.), *Handbook of personality: Theory and research* (pp. 693-720). New York: Guilford.

Roskam, E. E., & Jansen, P. G. W. (1984). A new derivation of the Rasch model. In E. Degreef & J. van Bruggenhaut (Eds.), *Trends in mathematical psychology.* Amsterdam: North-Holland.

Rozeboom, W. W. (1960). Studies in the empiricist theory of scientific meaning, Part I. Empirical realism and classical semantics: A parting of the ways. *Philosophy of Science, 27,* 359-373.

Rozeboom, W. W. (1966-a). *Foundations of the theory of prediction.* Homewood, Illinois: The Dorsey Press.

Rozeboom, W. W. (1966-b). Scaling theory and the nature of measurement. *Synthese, 16,* 170-233.

Rozeboom, W. W. (1973). Dispositions revisited. *Philosophy of Science, 40,* 59-74.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66,* 688-701.

Russell, J. A., & Carroll, J. M. (1999). On the bipolarity of positive and negative affect. *Psychological Bulletin, 125,* 3-30.

Ryle, G. (1949). *The concept of mind.* London: Penguin.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph,* 17.

Scheiblechner, H. (1999). Additive conjoint isotonic probabilistic models. *Psychometrika, 64,* 295-316.

Schmidt, F. L., & Hunter, J. E. (1999). Theory testing and measurement error. *Intelligence, 27,* 183-198.

Schouwstra, S. J. (2000). *On testing plausible threats to construct validity.* Amsterdam: Unpublished doctoral dissertation.

Scott, D., & Suppes, P. (1958). Foundational aspects of theories of measurement. *Journal of Symbolic Logic, 23,* 113-128.

Shealy, R., & Stout, W. (1993). An item response theory model for test bias. In P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp. 197-239). Hillsdale, New Jersey: Erlbaum.

Shepard, L. A. (1993). Evaluating test validity. *Review of research in education, 19*, 405-450.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*, 5-8.

Skinner, B. F. (1938). *The behavior of organisms: an experimental analysis.* New York: Appleton-Century.

Skinner, B. F. (1987). Whatever happened to psychology as the science of behavior? *American Psychologist, 42*, 780-786.

Smith, L. J., & Reise, S. P. (1998). Gender differences on negative affectivity: An IRT study of differential item functioning on the Multidimensional Personality Questionnaire Stress Reaction Scale. *Journal of Personality and Social Psychology, 75*, 1350-1362.

Smits, N., Mellenbergh, G. J., & Vorst, H. C. M. (*submitted*). The measurement versus prediction paradox in the application of planned missingness to psychological and educational tests.

Sobel, M. E. (1994). Causal inference in latent variable models. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis.* Thousand Oakes: Sage.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structures between groups. *Psychometrika, 55*, 229-239.

Sorensen, R. (1992). *Thought experiments.* Oxford: Oxford University Press.

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

Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence.* Cambridge: Cambridge University Press.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*, 667-680.

Stevens, S. S. (1968). Measurement, statistics, and the schemapiric view. *Science, 30*, 849-856.

Stigler, S. M. (1986). *The history of statistics.* Cambridge, Massachusets: Harvard University Press.

Stout, W. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 55*, 293-325.

Suppe, F. (1977). *The structure of scientific theories.* Urbana: University of Illinois Press.

Suppes, P., & Zanotti, M. (1981). When are probabilistic explanations possible? *Synthese, 48*, 191-199.

Suppes, P., & Zinnes, J. L. (1963). Basic measurement theory. In R. D. Luce, R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 3-76). New York: Wiley.

Takane, Y., & Leeuw, J. D. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*, 393-408.

Thagard, P. (1988). *Computational philosophy of science.* Cambridge, MA: MIT Press.

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika, 49*, 501-519.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51*, 567-577.

Thurstone, L. L. (1947). *Multiple factor analysis.* Chicago: University of Chicago Press.

Toulmin, S. (1953). *The philosophy of science.* London: Hutchinson & Co. Ltd.

Townshend, J. T., & Ashby, F. G. (1984). Measurement scales and statistics: The misconception misconceived. *Psychological Bulletin, 96*, 394-401.

Trout, J. D. (1999). Measurement. In W. H. Newton-Smith (Ed.), *A companion to the philosophy of science.* Oxford: Blackwell.

Van Fraassen, B. C. (1980). *The scientific image.* Oxford: Clarendon Press.

Van Heerden, J., & Smolenaars, A. (1989). On traits as dispositions: An alleged truism. *Journal for the theory of social behaviour, 19*, 297-309.

Van Lambalgen, M. (1990). The axiomatization of randomness. *Journal of Symbolic Logic, 55*, 1143-1167.

Velleman, P. F. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *The American Statistician, 47*, 65-72.

Watson, J. B. (1913). Psychology as the behaviorist views it. *Psychological Review, 20*, 158-177.

Wiley, D. E., Schmidt, W. H., & Bramble, W. J. (1973). Studies of a class of covariance structure models. *Journal of the American Statistical Association, 86*, 317-321.

Wilkes, K. (1988). *Real People.* Oxford: Clarendon Press.

Wilson, M. (1989). Saltus: A psychometric model of discontinuity in cognitive development. *Psychological Bulletin, 105*, 276-289.

Wittgenstein, L. (1922). *Tractatus logico-philosophicus.* London: Routledge & Kegan Paul.

Wood, R. (1978). Fitting the Rasch model: A heady tale. *British Journal of Mathematical and Statistical Psychology, 31*, 27-32.

Wright, B. D. (1997). A history of social science measurement. *Educational measurement: Issues and practice, 16*, 33-45.

10. NEDERLANDSE SAMENVATTING

Conceptuele problemen in de psychometrie

Meten speelt een belangrijke rol in de psychologie. Of het nu gaat om het toepassen van persoonlijkheidstests in sollicitatieprocedures, om onderzoek naar de effectiviteit van psychotherapie, of om het vaststellen van verschillen in intelligentie tussen bepaalde bevolkingsgroepen: de psycholoog meet veel en graag. Het meten van psychologische eigenschappen zoals intelligentie, persoonlijkheid, of mate van depressiviteit, verloopt ongeveer als volgt. Personen worden geconfronteerd met een aantal vragen, problemen, of stellingen. Die worden in de psychologie 'items' genoemd. Mensen geven dan een respons op die items. Vervolgens worden de responsen op de verschillende items op een of andere manier gecombineerd tot een totaalscore. Die totaalscore wordt dan beschouwd als een meting van de psychologische eigenschap in kwestie. Het idee achter zo'n procedure is eenvoudigweg dat personen die depressiever zijn eerder 'ja' zullen antwoorden op de vraag 'Slaapt U slecht?', en dat intelligentere mensen eerder zullen zien welk cijfer er moet volgen in de reeks '1, 1, 2, 3, 5, 8, ..'. Depressieve mensen zullen meer en ernstiger depressieve klachten hebben, en intelligentere mensen zullen meer en moeilijker problemen oplossen. Derhalve zal variatie in testscores samenhangen met variatie in de te meten eigenschap. Dat is het basisidee van meten in de psychologie.

Helaas zijn de scores op psychologische tests niet altijd even makkelijk te interpreteren. Hoe weet de psycholoog bijvoorbeeld dat de IQ-test daadwerkelijk intelligentie meet? En zo ja, hoe precies is de meting dan? Hoe kan hij dat nagaan? Wat is dat eigenlijk, intelligentie? Het beantwoorden van dit soort vragen is erg belangrijk, maar wordt bemoeilijkt door het feit dat psychologen geen goed inzicht hebben in de processen die in het hoofd plaatsvinden op het moment dat een persoon een vraag beantwoordt. Kort gezegd komt het erop neer dan men niet precies weet hoe een persoon tot zijn antwoord komt, en daardoor is onduidelijk wat er gemeten wordt. Dat is niet het enige probleem. Omdat mensen altijd wel antwoord geven op vragen als ze daartoe worden aangespoord – ook als die vragen helemaal niets met de te meten eigenschap te maken hebben – en omdat mensen niet bijster consistent zijn – sommige mensen maken bijna alle moeilijke vragen in een intelligentietest goed, maar missen nu net dat ene makkelijke item – kunnen testscores niet beschouwd worden als een perfecte meting van psychologische eigenschappen. Daarom worden, om na te gaan hoe goed de metingen zijn, statistische modellen gebruikt.

Als psychologen echt wisten hoe de geobserveerde responsen op items samenhingen met de te meten eigenschap, dan zouden zulke modellen wellicht niet nodig zijn. Maar ze zijn dus wel nodig. En het interessante probleem doet zich nu voor, dat de psycholoog in de analyse van testresultaten wordt gedwongen bepaalde veronderstellingen te doen over de verhouding tussen de testscores en de te meten eigenschap. Die veronderstellingen zitten in de statistische modellen, maar hoeven niets met inhoudelijke theorie te maken te hebben. Omdat er meerdere soorten modellen zijn, moet de psycholoog er een kiezen. Daarmee kiest hij, meestal impliciet, ook voor een bepaalde visie op wat een psychologische eigenschap is en hoe die te maken heeft met de testscores. Over die visies gaat dit proefschrift. Ik bekijk een aantal veel gebruikte, dan wel vaak gepropageerde, wiskundige modellen, en daarbij stel ik mij de vraag: Als een psycholoog voor model X zou kiezen, wat voor een relatie tussen de eigenschap (intelligentie) en de testresponsen (IQ-scores) moet hij dan veronderstellen? Drie modellen komen daarbij aan de orde: het klassieke testmodel, het latente-variabelenmodel, en het representationele meetmodel. Vervolgens worden de relaties tussen de modellen besproken. Tenslotte neem ik uit ieder model enkele ideeën die mij plausibel lijken, en voeg ik die samen tot een geïntegreerde visie op het meetproces, en in het bijzonder het validiteitsbegrip.

In **Hoofdstuk 2** komt het klassieke testmodel aan de orde. Dit model richt zich op een opdeling van geobserveerde scores in een ware score en een meetfout. De ware score wordt beschouwd als de verwachtingswaarde van de geobserveerde scores, en de meetfout is wat er overblijft. Dat gaat echter niet zomaar. Om met verwachtingwaardes te kunnen werken moeten de testscores aan bepaalde eigenschappen voldoen. Meer specifiek moeten zij opgevat kunnen worden als het resultaat van een kansexperiment. Een kansexperiment is bijvoorbeeld een worp met een dobbelsteen. Het is echter overduidelijk dat testresponsen niet beschouwd kunnen worden als het resultaat van een kansexperiment: Het oplossen van een probleem in een IQ-test heeft, als proces, niets gemeen met het gooien van een dobbelsteen. Dat realiseren klassieke testtheoretici zich ook, en daarom hebben zij een gedachte-experiment bedacht. Stel, zo zegt de klassieke testtheoreticus, dat we mensen zouden kunnen hersenspoelen tussen twee testafnames door, en hen dan iedere keer opnieuw de test laten maken. Dan zouden ze niet iedere keer dezelfde score halen. Die score zal dan toevallige variatie vertonen. En in dat geval kunnen we de verwachte testscore opvoeren als de ware score: Die ware score is dan de gemiddelde score over een zeer lange reeks herhaalde testafnames – met tussen iedere twee testafnames een hersenspoeling.

De klassieke testtheoreticus heeft nu iets heel vreemds gedaan. In plaats van bepaalde statistische patronen waar te nemen en daar een model bij te verzinnen, heeft hij een model bedacht en daar vervolgens de patronen bij verzonnen die erbij horen. Dat die patronen niet alleen onwaarschijnlijk zijn, maar in feite helemaal niets met psychologische processen te maken hebben, devalueert de status van het model en daarmee van de ware score. De ware score heeft op zichzelf niets met psychologische eigenschappen te maken, en al helemaal niets met de waarheid. De ware score heeft uitsluitend te maken met de test, en zelfs daarvoor moet nog behoorlijk wat theoretische acrobatiek uit de kast gehaald worden. Omdat de ware score uitsluitend gedefinieerd is in termen van herhaalde testafnames, moet de psycholoog die intelligentie ziet als een ware score ook intelligentie definieren in termen van herhaalde testafnames. In de wetenschapsfilosofie heet zo'n opvatting operationalistisch. Een operationalist zegt: Intelligentie is volledig gedefinieerd in termen van de IQ-test. Dat is precies zoals de klassieke testtheorie de ware score definieert. Omdat niemand in de psychologie operationalist is, is het raadselachtig – om niet te zeggen inconsequent – dat het klassieke model het meest gebruikte testmodel is. Nog vreemder is dat veel onderzoekers denken dat verschillen in ware scores op intelligentietests hetzelfde zijn als verschillen in intelligentie. Ik laat zien dat deze opvatting zelfs met de beste wil van de wereld niet houdbaar is. Dat komt niet zozeer omdat de klassieke testtheorie de relatie tussen testscores en psychologische eigenschappen verkeerd voorstelt, maar omdat zij die relatie helemaal niet beschouwt. Ik concludeer dat het klassieke testmodel, hoewel wiskundig elegant en makkelijk in het gebruik, niet geschikt is om de relatie tussen psychologische eigenschappen en testscores te conceptualiseren.

Een tweede kandidaat wordt besproken in **Hoofdstuk 3**, waar het latente-variabelenmodel aan de orde komt. Dit model neemt aan dat variatie in verwachte testscores – die in sommige visies gezien worden als ware scores – een functie is van variatie op een niet direkt waargenomen, dus latente, variabele. De psycholoog zou ervoor kunnen kiezen om psychologische eigenschappen te beschouwen als latente variabelen. In dat geval moet hij wel van tevoren aangeven wat de relatie tussen die latente variabelen en de testscores is. Om dat te kunnen doen moet hij aannemen dat deze variabelen min of meer onafhankelijk van de gebruikte test bestaan, en een bepaalde structuur hebben. In de wetenschapsfilosofie staat zo'n visie bekend als realisme. Omdat de aannames van latente-variabelenmodellen soms nogal streng zijn, wordt weleens gezegd dat onderzoekers voor het specificeren van zulke modellen niet hoeven aan te nemen dat latente variabelen zelfs maar zouden kunnen bestaan. Die visie bestrijd ik met een aantal argumenten. Het belangrijkste argument is dat het vrijwel onmogelijk is de latente variabele volledig te definieren in termen van de gebruikte test. Er blijven altijd keuzes over betreffende de structuur van het model die voor rekening van de psychologie komen, omdat ze nergens dwingend uit volgen. Om die aannames te kunnen motiveren moet de psycholoog aannemen dat psychologische eigenschappen onafhankelijk van de meetprocedure bestaan.

De vraag die zich daarop voordoet is: Als latente variabelen zouden bestaan, wat voor een relatie zouden ze dan met de testscores hebben? Een mogelijke interpretatie is dat die relatie als oorzakelijk moet worden gezien: variatie op de latente variabele veroorzaakt dan variatie in de testscores. Ik laat zien dat zo'n oorzakelijke interpretatie wel geformuleerd kan worden, maar dat deze voor de meest gebruikte modellen uitsluitend beschouwd kan worden in termen van verschillen tussen personen: Je kunt wel zeggen dat latente verschillen tussen personen oorzakelijk relevant zijn voor geobserveerde verschillen tussen personen, maar dat betekent niet dat de latente variabele bij een individu een causale rol speelt. Deze laatste hypothese wordt in het model niet getoetst. Deze situatie is niet geheel bevredigend, omdat psychologische theorieën vaak juist wel op individueel niveau geformuleerd zijn. Ik geef daarom op een aantal punten aan welke richting het onderzoek uit zou kunnen om verschillen tussen personen te relateren aan processen binnen personen. De conclusie van het hoofdstuk is, dat een realistische interpretatie van latente-variabelenmodellen, mits niet verkeerd opgevat, een redelijk beeld van het meetproces geeft.

In het **Hoofdstuk 4** van het proefschrift komt een zelden gebruikt, maar vaak gepropageerd, alternatief naar voren voor zowel het ware score model als het latente-variabelenmodel. Dit model heet het representationele meetmodel. Het representationalisme beschouwt meetschalen als weergaves (representaties) van geobserveerde relaties. Omdat niet hoeft worden aangenomen dat er zoiets als 'intelligentie' in de werkelijkheid bestaat, lijkt deze strategie veel op een stroming die in de wetenschapsfilosofie als empiricisme bekend staat. De onderzoeker observeert patronen in de data, en representeert deze patronen in een wiskundige constructie, namelijk de meetschaal. Omdat de onderzoeker deze schaal expliciet zelf construeert, moet de psycholoog die intelligentie opvat als een meetschaal deze psychologische eigenschap opvatten als zijn eigen constructie. Intelligentie is dan dus niet iets, dat onafhankelijk van de onderzoeker in de wereld bestaat, maar iets dat de onderzoeker zelf geconstrueerd heeft.

Hoewel het representationalisme zowel wiskundig als filosofisch gezien zeer krachtig is, kleven er een aantal bezwaren aan die de benadering minder geschikt maken voor het meten van psychologische eigenschappen. Geobserveerde verschillen tussen mensen zijn nogal chaotisch, en in een strikte interpretatie van het representationalisme moeten we daarom concluderen dat psychologische metingen zeldzaam zijn of zelfs helemaal niet bestaan: De structuur die voor gebruik van het woord 'meting' noodzakelijk is wordt in psychologische testscores namelijk vrijwel niet aangetroffen. De eisen die het representationalisme stelt zijn echter zo streng, dat ze op de keper beschouwd bijna ieder vorm van meten uitsluiten. Een belangrijke reden daarvoor is dat het model principiele bezwaren tegen het introduceren van meetfouten heeft. Daardoor kan het model moeilijk in statistische termen geformuleerd te worden. Als dat wel gebeurt, dan wordt het model een speciaal soort latente-variabelenmodel, en wordt realisme over psychologische eigenschappen door de achterdeur weer binnen gebracht. Wanneer dat realisme geaccepteerd wordt, dan is er echter geen goede reden om de restricties, die voor het representationalisme noodzakelijk zijn, te handhaven, waardoor de hele onderneming in het water valt.

Omdat het representationalisme vrijwel geen praktische toepassingen in de psychologie kent, en gebaseerd is op een sterk geidealiseerd beeld van meten, stel ik voor het representationalisme niet als een praktisch model te beschouwen, maar als een geidealiseerde reconstructie van het meetproces zoals het plaatsvindt in de natuurwetenschappen. Uit het feit dat metingen in de natuurkunde min of meer gereconstrueerd kunnen worden in termen van het representationele model, volgt echter niet dat psychologen dat model in de praktijk van testanalyse moeten gebruiken. Daarvoor zijn de psychologie en de natuurkunde te verschillend. Recente pogingen van een aantal theoretici om het model als normatief model voor de psychologische praktijk te introduceren wijs ik daarom af als ongegrond.

Hoewel de besproken modellen in filosofisch opzicht verschillend zijn, lijken ze formeel gezien soms erg sterk op elkaar. In **Hoofdstuk 5** bespreek ik de voorwaarden waaraan moet worden voldaan om de modellen met elkaar in overeenstemming te laten zijn. Uit deze analyse blijkt, dat de modellen elkaar niet hoeven tegen te spreken, maar dat ze zich wel op een ander gedeelte van het meetproces concentreren. Het latente-variabelenmodel kan gezien worden als een hypothese over hoe de verschillen in testscores tot stand komen, het ware score model behandelt de structuur van de meetfouten, en het representationele model geeft een representatie van relaties tussen ware scores door die relaties af te beelden in een meetschaal. Om deze verbinding tot stand te kunnen brengen, moet echter worden aangenomen dat verwachtingswaardes op het individuele niveau betrekking hebben. Dat vereist een soortgelijk gedachte-experiment als in de klassieke testtheorie. Het is echter ook mogelijk om verwachtingswaardes te zien als gemiddelden, die gedefinieerd zijn op subgroepen van mensen met dezelfde positie op de latente variabele. Het proces dat tot de respons op een vraag leidt wordt dan niet opgevat als een kansexperiment. In deze interpretatie hebben de modellen vrijwel niets met elkaar te maken. Het ware score model kan dan namelijk niet worden gedefinieerd, en daarom werkt het representationele meetmodel ook niet meer: als er geen verschillen in ware scores zijn om af te beelden in de meetschaal, dan kan die meetschaal niet worden geconstrueerd. Het latente-variabelenmodel kan dan nog wel opgesteld worden, maar is dan niet langer een model voor het item-respons proces, maar voor verschillen tussen subpopulatiegemiddelden.

De vraag die zich nu voordoet is: wat is een zinnige manier om naar de relatie tussen psychologische eigenschappen en testscores te kijken? In het tweede gedeelte van hoofdstuk 5 maak ik met betrekking tot deze vraag een aantal keuzes. Het ware score model geeft helemaal geen beeld van de betreffende relatie, behalve in een operationalistische interpretatie van psychologische eigenschappen, en aangezien die interpretatie onzinnig is moet zij afgewezen worden. Het representationele model is ongeschikt omdat het nauwelijks statistisch geformuleerd kan worden, en de aanname dat de relatie tussen psychologische eigenschappen en testscores deterministisch is al te sterk. In een nadere beschouwing wordt echter opgemerkt dat het representationele model, strikt genomen, de aanname doet dat experimentele controle mogelijk is. En die experimentele controle kan gezien worden als een interventie in een causaal systeem. Wanneer het model statistisch geformuleerd wordt, dan moet het worden uitgebreid wordt met latente variabelen en een realistische interpretatie. Hoewel experimentele controle over latente variabelen zowel praktisch onmogelijk als een inhoudelijk theoretisch ongemotiveerde aanname is, kan de zwakkere aanname, dat de relatie tussen latente variabele en geobserveerde score causaal van aard, is wel gehandhaafd worden. Dat betekent dat het latente-variabelenmodel en het representationalisme zeer dicht bij elkaar komen. In feite komt het erop neer, dat het niet onredelijk is dat, om van een meting van een psychologische eigenschap te kunnen spreken, aan twee voorwaarden voldaan moet zijn: de betreffende psychologische eigenschap moet bestaan, en variatie op deze eigenschap moet de oorzaak zijn van variatie in de testscores.

Deze conclusie wordt in **Hoofdstuk 6** gebruikt om een nieuwe inhoud te geven aan het validiteitsbegrip. De validiteitsliteratuur houdt zich bezig met de vraag: meten psychologische tests de juiste psychologische eigenschappen? Hoewel er zeer veel geschreven is over onderzoeksprocedures om dit na te gaan, is er in mijn ogen te weinig aandacht geweest voor de vraag wat het betekent als je zegt dat IQ-tests intelligentie meten. Ik beweer dat deze stelling waar is als verschillen in intelligentie verschillen in testscores veroorzaken, en anders niet. Deze opvatting geeft een inhoud aan het validiteitsbegrip die radicaal afwijkt van de huidige consensus in de literatuur. Waar de literatuur de betekenis van testscores in termen van een theorie als definierende karakteristiek van het validiteitsbegrip aanvoert, is in mijn opvatting niet de betekenis, maar het bestaan van psychologische eigenschappen cruciaal. Waar de literatuur het heeft over de overeenstemming tussen scores op verschillende tests, voer ik de causale relatie tussen eigenschap en score aan als essentieel element van het validiteitsbegrip. Waar de validiteitsliteratuur validiteit ziet als een eigenschap van de interpretatie van testscores, zie ik validiteit als een eigenschap van de test zelf. En waar in de literatuur geprobeerd wordt vrijwel ieder belangrijk aspect van testgebruik onder het validiteitsbegrip te laten vallen, beperk ik de betekenis van het begrip aanzienlijk: validiteit gaat over de vraag of de test de bedoelde eigenschap meet, en nergens anders over. Relevante andere vragen, zoals de vraag hoe precies een test de bedoelde eigenschap meet, laat ik voor rekening van de technisch georienteerde psychometrie, die er veel meer over te zeggen heeft dan de filosofisch georienteerde validiteitsliteratuur. Deze voorstelling van zaken leidt tot een geheel andere kijk op de vraag waar het validiteitsprobleem in de psychologie vandaan komt. Dit probleem ontstaat misschien niet zozeer doordat de psycholoog in veel gevallen niet weet wat een test meet, maar doordat hij niet goed weet wat hij wil meten.

Stellingen

behorend bij het proefschrift
'Conceptual Issues in
Psychological Measurement'
door Denny Borsboom

1. Ware scores hebben met waarheid niets te maken (Hoofdstuk 2 van dit proefschrift).

2. De structuur van interindividuele verschillen zegt weinig tot niets over de structuur van intraindividuele processen (Hoofdstuk 3 van dit proefschift).

3. N=1 onderzoek is niet onwetenschappelijk, maar de verkettering ervan wel (Hoofdstuk 3 van dit proefschrift).

4. Het representationele meetmodel schrijft niet voor hoe het meetproces eruit moet zien, maar geeft daarvan een geïdealiseerde logische reconstructie (Hoofdstuk 4 van dit proefschrift).

5. Het validiteitsprobleem kan alleen worden opgelost door een causale theorie over het itemresponsproces op te stellen (Hoofdstuk 6 van dit proefschift).

6. Validiteit is geen methodologisch, maar een inhoudelijk probleem (Hoofdstuk 6 van dit proefschrift).

7. De maatschappelijke relevantie van dit proefschrift is wetenschappelijk irrelevant.

8. 'Fundamenteel onderzoek' is een pleonasme; 'toegepast onderzoek' een oxymoron.

9. Een universiteitsbestuurder, die meer verdient dan de beste onderzoekers en docenten in zijn organisatie, zou 's nachts niet moeten kunnen slapen.

10. Gitaarsolo's en wetenschappelijke artikelen hebben gemeen dat zij spannender worden naarmate er meer weggelaten wordt.

11. Wie gelooft in levitatie moet nodig eens proberen een 7a te klimmen.

Reviewers' comments

"This paper hits everyone smack on the head with the problem, and it does so with an analysis that is particularly well informed by work in the philosophy of science."

"This engaging, imaginative, and provocative essay makes the 'modern' thinkers of the last generation seem like misguided old fuddy-duddies bogged down in their own complexities."

"I would rethink the ending. To read all of this stuff and have the authors tell you it doesn't make any difference is somewhat discouraging. I am sure they don't really mean that."

"This paper raises a red herring. It is completely divorced from what happens in real research and in that critical sense is irrelevant."

"This article is outstanding in the force and correctness of its logic and deserves publication as is."

"The primary issue that concerns me is the big one: So what?"