Conceptual issues in psychological measurement

Borsboom, D.

Citation for published version (APA):
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 philos-
1.2 Theoretical terms in psychology

...
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
1.3 The function of models in psychology

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
Introduction
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 $\alpha$, 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 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.