Psychological Review
2003, Vol. 110, No. 2, 203–219

0033-295X/03/$12.00 DOI: 10.1037/0033-295X.110.2.203

The Theoretical Status of Latent Variables

Denny Borsboom, Gideon J. Mellenbergh, and Jaap van Heerden
University of Amsterdam

This article examines the theoretical status of latent variables as used in modern test theory models. First, it is argued that a consistent interpretation of such models requires a realist ontology for latent variables. Second, the relation between latent variables and their indicators is discussed. It is maintained that this relation can be interpreted as a causal one but that in measurement models for interindividual differences the relation does not apply to the level of the individual person. To substantiate intraindividual causal conclusions, one must explicitly represent individual level processes in the measurement model. Several research strategies that may be useful in this respect are discussed, and a typology of constructs is proposed on the basis of this analysis. The need to link individual processes to latent variable models for interindividual differences is emphasized.

Consider the following sentence: "Einstein would not have been able to come up with his $e = mc^2$ had he not possessed such an extraordinary intelligence." What does this sentence express? It relates observable behavior (Einstein's writing $e = mc^2$) to an unobservable attribute (his extraordinary intelligence), and it does so by assigning to the unobservable attribute a causal role in bringing about Einstein's behavior. In psychology, there are many constructs that play this type of role in theories of human behavior; examples are constructs like extraversion, spatial ability, self-efficacy, and attitudes. Such variables are usually referred to as *latent variables*. It is common to investigate the structure and effect of unobservables like intelligence through the analysis of interindividual differences data by statistically relating covariation between observed variables to latent variables. This is done, for example, in the widely used factor model. The idea is that although the fit of a latent variable model to the data may not prove the existence of causally operating latent variables, the model does formulate this as a hypothesis; consequently, the fit of such models can be adduced as evidence supporting this hypothesis. Finally, it is often suggested that the type of causal relation tested in latent variable modeling is similar to the relation between Einstein's intelligence and behavior in the above example; that is, the latent variable exerts influence at the level of the individual.

Given the intuitive appeal of explaining a wide range of behaviors by invoking a limited number of latent variables, it is not surprising that latent variables analysis has become a popular technique in postbehaviorist psychology. The conceptual framework of latent variables analysis, however, is older than cognitive psychology and originates with the work of Spearman (1904), who developed factor analytic models for continuous variables in the context of intelligence testing. The basic statistical idea of latent variables analysis is simple. If a latent variable underlies a number of observed variables, then conditionalizing on that latent variable will render the observed variables statistically independent. This is known as the *principle of local independence*. The problem of latent variables analysis is to find a set of latent variables that satisfies this condition for a given set of observed variables.

With these insights, Spearman (1904) opened up a paradigm, and the development of this paradigm in the 20th century has been spectacular. The factor analytic tradition continued with the work of Lawley (1943), Thurstone (1947), and Lawley and Maxwell (1963), and it entered into the conceptual framework of confirmatory factor analysis (CFA) with Jöreskog (1971); Wiley, Schmidt, and Bramble (1973); and Sörbom (1974). In subsequent years, CFA became a very popular technique, largely because of the LISREL program by Jöreskog and 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 (1971). 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 and Steinberg (1984). Meanwhile, in yet another parallel research program, methods were developed to deal with categorical latent variables. In this context, Lazarsfeld (1950), Lazarsfeld and Henry (1968), and Goodman (1974) developed latent structure analysis. Latent structure models may involve categorical observed variables, in which case one speaks of latent class analysis or metrical observed variables giving rise to latent profile analysis (Bartholomew, 1987). After boundary-crossing investigations by McDonald (1982), Thissen and Steinberg (1986), Takane and de Leeuw (1987), and Goldstein and Wood (1989),

Denny Borsboom, Gideon J. Mellenbergh, and Jaap van Heerden, Department of Psychology, University of Amsterdam, Amsterdam, the Netherlands.

We thank Gitta Lubke, Sanneke Schouwstra, Maarten Speekenbrink, and the members of the SEMNET electronic discussion group for many useful discussions on latent variables analysis. Ingmar Visser, Helma van den Berg, Conor Dolan, and Keith Markus have provided useful comments on early versions of this article. Special thanks go out to Peter Molenaar, whose work has stimulated many of the developments in this article.

Correspondence concerning this article should be addressed to Denny Borsboom, Department of Psychology, Faculty of Social and Behavioral Sciences, University of Amsterdam, Roetersstraat 15, 1018 WB Amsterdam, the Netherlands. E-mail: dborsboom@fmg.uva.nl

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 and Muthén (1998), McDonald (1999), and Moustaki and Knott (2000). Different terms are used to indicate the general latent variable model. For example, Goldstein and Wood (1989) use the term *generalized linear item response model* (GLIRM), whereas Mellenbergh (1994) speaks of *generalized linear item response theory* (GLIRT), and Moustaki and Knott (2000) follow McCullagh and Nelder (1989) in using the term *generalized linear model* (GLIM). We 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 21st 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.

The growing use of latent variables analysis in psychological research 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, e.g., Edwards & Bagozzi, 2000; Pearl, 1999; Sobel, 1994), the status of a strongly related concept, namely the true score of classical test theory (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, during which battles were fought over the conceptual status of unobservable entities such as electrons (for some contrasting views, see Cartwright, 1983; Devitt, 1991; Hacking, 1983; and Van Fraassen, 1980). However, the theoretical status of the latent variable as it appears in models of psychological measurement has not received a thorough and general analysis as yet.

The following questions, for example, are relevant but seldom addressed in detail. 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 important, 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, we will argue 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. Further, we develop a distinction between several kinds of latent variables on the basis of their relations with individual processes.

Before we start out on this investigation, some qualifications are in order. Latent variable models for psychological measurement are generally used in research in which a number of items, or tests, are administered to a number of subjects at a single time point. This type of model, which explains between-subjects covariation by invoking latent variables on which subjects differ from each other, is the primary topic of this paper. There are three reasons for this. First, it is the most widely used model in psychology; second, its formal theory is has been developed in great detail; and third, it is the basis for some of the most influential latent variable models around. These include those used in intelligence testing (with the general intelligence model as a primary example) and those used in personality research (with the five factor model as a primary example). We denote this model as the standard measurement model.

The structure of this article is as follows. First, it is argued that the latent variable typically appears in two distinct ways: as a formal–theoretical concept and as an operational–empirical concept. In applications, these two concepts have to be connected. To do this, however, we need a third—ontological—concept. We distinguish three ontological frameworks that may be applied: realism, constructivism, and operationalism. It is argued that a realist account of the latent variable is required to maintain a consistent connection between the formal and empirical concept of a latent variable. The realist view requires an account of the relation between the latent variable and its indicators, for which causality is a natural candidate. We inquire whether such an interpretation can be defended, and if so, how this causal relation should be interpreted. Finally, we discuss the implications of our analysis for research in psychology.

Three Ways of Looking at the Latent Variable

If one carefully examines the practice of testing, it appears that there are at least two distinct ways in which the concept of a latent variable is used. The first is as a formal, technical term, and the second as an empirical term. The formal concept figures in mathematical treatments, whereas the empirical concept is a function of the observed scores (often a weighted sumscore). For example, a five factor model may be fitted to personality data. On the basis of this model, factor scores can be constructed by summing appropriately weighted item (or subtest) scores. It is natural to connect the formal and empirical concepts by conceiving of such a weighted sumscore as an "estimate" of or as a "proxy" for the latent variable of interest, as is customary in the literature; in the example, the weighted sumscore of all items loading on the factor *extraversion* would be considered an estimate of the level of extraversion.

We will argue that this position is not without problems. Specifically, to make the connection, we need an ontology for the latent variable. This requires an account from a third stance, which we term *the ontological stance*. We will argue that the ontology must be realist in nature. To clarify the problem situation, we will discuss the formal and empirical connotations of the term *latent variable* before establishing a connection between the two.

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. This regression function may differ in form (e.g., it is linear for the factor model but logistic for the Rasch, 1960, model; see also Mellenbergh, 1994). For instance, in a factor model for general intelligence, one would specify 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 although 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.

Now there are two things one can do on the basis of the set of formal assumptions underlying latent variables analysis. First, one 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, one 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 often implicitly suggested that the formal derivations reveal something about reality, but this is not the case. Each supposition inside the formal system is a tautology, and tautologies in themselves cannot reveal anything about the world. So this is all in the syntactic domain; that is, it has no meaning outside the formal theory. We will denote the latent variable as it appears in this "formal stance" (i.e., the concept indicated by θ, in the IRT literature and by η or ξ in the structural equation modeling [SEM] literature) as the *formal latent variable*.

The Formal Stance: 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?

The syntax of latent variables analysis is taken from statistics, and so are its semantics. Statistics is concerned with the behavior of random variables, that is, with variables whose actual realization is determined in a chance experiment. It is clear that the interpretation of such variables as random, and the statistical treatment that is based on that interpretation, is related to the unpredictability of the processes that lead to the outcome of the chance experiment. The justification for using statistical techniques depends, in general, on the plausibility of such an interpretation. This means that one has to show that the variable of interest can, in some sense, be conceived of as a variable whose values are determined by a chance experiment, so that the variable can be considered a proper random variable.

In psychological measurement, the outcome variable that must be conceived of as random is the item response. After all, it is the expectation of the item response that goes into the regression formulas. At first sight, however, it is not at all clear why a response to an item in a psychological test should be considered a random variable. It is therefore important to interpret the item response in such a way as to justify this approach. This is rarely stated explicitly in treatments of psychological measurement, but it is crucial to the applicability of statistical models. This paragraph is concerned with possible interpretations of the response to an item, say, the item "2 + 2 = . . .," that may be used to justify treating such a response as a random variable.

The main question is, how does one interpret the conditional probability distribution of the observed variables, given the latent variable? Although there may be many possible interpretations of this distribution, we focus on two consistent interpretations that were distinguished by Holland (1990). The first interpretation, known as the *stochastic subject interpretation*, takes the probability distribution as applying to the individual subject. This interpretation implies a series of hypotheticals of the form, "Given that Subject A has Value X on the latent variable, A has Probability Distribution Y over the item responses." Supposing that the imaginary subject John takes an intelligence test item, this would become something like, "Given that John's level of intelligence is 2 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 1 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 second interpretation we discuss is the *repeated sampling interpretation*, which is more common in the literature on factor analysis (see, e.g., 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 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 As with Value X on the latent variable follows Distribution Y over the item responses." Now, the probability distribution over the item responses that pertains to a specific Value X of the latent variable arises from repeated sampling from the population of As having this value. 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 2 standard deviations below the population mean) will answer the item correctly with probability 1, and 30% of those people will answer the item correctly with probability 0. 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 one is modeling: in the stochastic subject formulation, one is modeling characteristics of individuals, whereas in the repeated sampling interpretation, one is modeling between-subjects variables. However, if one follows the stochastic subject interpretation and assumes 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 who 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 and Van den Wollenberg (1993) have shown, however, that the local homogeneity assumption is an independent assumption that in no way follows 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 and 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, we will argue 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 paper.

The Empirical Stance

Before we discuss the ontology of the latent variable, we make an observation in the empirical domain. This observation is simple: If observed variables behave in the right way, a latent variable model will fit. By "in the right way," we mean that the pattern of scores behaves according to the model. For some models, this requirement is more stringent than for others. In a standard CFA, for example, only first-and second-order moments are involved in the analysis, so that the requirement applies only to this part of the data structure; for a Rasch (1960) model, additional requirements concerning the pattern of scores are necessary. However, the central point is both simple and instructive: The *explanandum* (observed scores) can be discussed separately from the *explanans* (the model).

The well known problem of underdetermination (any set of data can be explained by an indefinite number of theories) illustrates why the model cannot be considered identical with or implied by the corresponding empirical structure and, as a matter of fact, should be considered strongly distinct from that structure. In a statistical context, the problem of underdetermination translates into the idea that many data-generating mechanisms (i.e., models) may lead to the same dataset. There is a connection here with the issue of equivalent statistical models (see, e.g., 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. The models are conceptually different: The factor model posits continuous latent variables (i.e., it specifies that subjects vary in degree but not in kind), whereas 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. That such theoretically distinct models can be practically equivalent in an empirical sense urges a strong distinction between the formal and empirical structure of latent variables analysis.

We make this point 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* (after Sobel, 1994). 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 can 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 (1960) model. A more general treatment is given in Suppes and Zanotti (1981) who show that for three two-valued 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 and 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. 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.

The Ontological Stance

The formal latent variable is a mathematical entity. It figures in mathematical formulas 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 require. The assumption that it was this model, and not some other model, that generated the data must precede the estimation process. In other words, if one considers the weighted sumscore as an estimate of the position of a given subject on a latent variable, one does so under the model specified. Now this weighted sumscore is not an estimate of the formal latent variable; one does not use an IQ score to estimate the general concept usually indicated by the Greek letter θ, but to estimate intelligence. Thus, one uses the formal side of the model to acquire knowledge about some part of the world; then it follows that one estimates something that is also in that part of the world. What is that something?

It will be clear that in answering this question, one 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. We 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 we intend the latter term to indicate a broader class of views (e.g., the more sophisticated empiricist view of Van Fraassen, 1980). In fact, we think that only the first of these views can be consistently attached to the formal content of latent variable theory.

Note that our discussion of these views is not meant to constitute an exhaustive categorization of the possible positions one may take. For present purposes, however, the gap between realism and constructivism is more important than the fine line separating various forms of each position. For this reason, we limit our attention to these views.

Operationalism and the Numerical Trick

We will first discuss the last view—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 flawed is to take any standard textbook on latent variable theory and to replace the term *latent variable* with *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 one should not take such texts literally or, worse, that one 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 one constructs. This is a direct consequence of the operationalist view (Bridgman, 1927), which holds that the meaning of a concept is synonymous with the set of operations used to measure it. Therefore, distinct sets of operations define distinct concepts (Suppe, 1974). In the present context, this implies that different sets of items must necessarily measure different latent variables. This is inconsistent with the basic idea of latent variable theory. To see this, consider a simple test consisting of three items *a*, *b*, and *c*. In the operationalist view, the latent variable that accounts for the item responses on the subtest consisting of items *a* and *b* is different from the latent variable that accounts for the item response pattern on the subtest consisting of items *b* and *c*. So, the test consisting of items *a*, *b*, and *c* does not measure the same latent variable and therefore cannot be unidimensional. In fact, in 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, in which different tests are administered to different subjects with the objective to measure a single latent variable. We conclude that operationalism and latent variable theory are fundamentally incompatible.

A related view holds that the use of latent variable theory is merely instrumental, a means to an end. This is the 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 reduction technique, they 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 of 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. 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, in 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 constructivists' denial of one or more of the realists' claims. Realism exists in a number of forms, but in general a realist will maintain one or several of the following theses (Devitt, 1991;

[1] This should not be read as a value judgment. We think data reduction techniques are very important, especially in the exploratory phases of research. That these techniques are important, however, does not entail that they are not ad hoc.

Hacking, 1983). 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. We will argue 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 that we call *electrons* correspond to particles in reality. This does not imply 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 nonexistent entities is difficult to conceive of. We 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 depicted in the left panel of Figure 1. (We borrow the symbolic language 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, in which 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 conceptualization of measurement in psychology. An alternative model that is more customary in sociological and economical modeling is the model in the right panel of Figure 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 socioeconomic status (SES). In such a model a researcher would, for example, record the variables income, educational level, and neighborhood as indicators of SES.

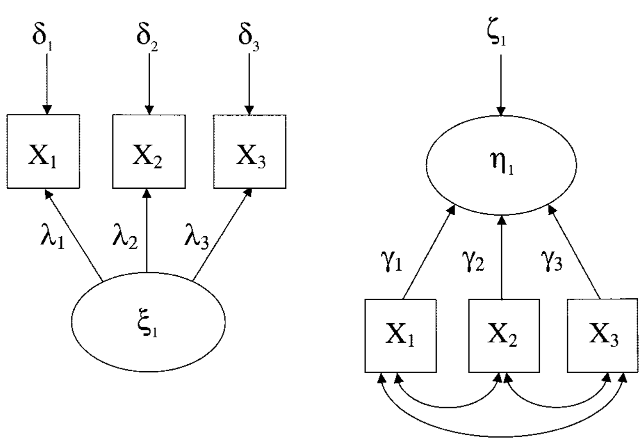

Figure 1. Two models for measurement. The left panel is the reflective measurement model. The Xs are observed variables, ξ is the latent variable, λs are factor loadings, and the δs are error terms. The right panel shows the formative model. The latent variable is denoted η, the γs are the weights of the indicators, and ζ is a residual term.

The models in Figure 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 use 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 that 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 one is not forced in any way to make such an assumption. Now, if one wanted to know how to weigh the relative importance of each of the measurements comprising SES in predicting, say, health, one could use a formative model like the one depicted in the right panel of Figure 1. In such a model, one 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

[2] It is in itself an interesting (and neglected) question as to where to draw the line separating these classes of models at the substantive level. For example, which of the formal models should be applied to the relation between diagnostic criteria and mental disorders in the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.; American Psychiatric Association, 1994)?

variable. However, nowhere in this development has it been shown that SES exists independent of the measurements.

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 on the hypothetical side of the argument (the assumptions of the model). Maybe more important than this is 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 to have a match between theoretical relations and relations in reality, these relations in reality have to exist quite independently 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 positing 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 possible only with 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 2 subjects, say John and Jane. You find that the weighted sumscore (i.e., the operational latent variable) is greater for John than for Jane, and you tentatively conclude that John occupies 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, "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 one finds 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 meaningful only 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 that this is a problem only 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 (Lee, 1997; Novick & Jackson, 1974) 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. Because it is a probability distribution, however, the total probability over the set of parameter values must be equal to 1. This means that, in specifying a prior distribution, 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 1. 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 in which one encounters 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 goodness-of-fit chi-square test is computed under the null hypothesis that the model is true. Psychometricians are, of course, aware 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 (Browne & Cudeck, 1992; Cudeck & Browne, 1983). For example, McDonald and Marsh (1990) stated, “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 that 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 although this is not the case? So interpreted, one is still a realist; there is a true model, but it is a different model from the one we specified, that is, one in which the latent variable is not normally distributed. The fact that the model is false is, in this sense, contingent on 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, in 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 because the whole notion of misspecification requires the existence of a true model, for how can we misspecify if there is no true model? Now, one may say that one judges the (misspecified) model close enough to reality to warrant the estimation procedures. One then interprets the model as “approximately true.” So, with this interpretation, one is firmly in the realist camp, even though one acknowledges that one has not succeeded in formulating the true model. This is as far as realists could go in the acknowledgement that their 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 one 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 its ugly head.

For example, take a simple case of statistically equivalent covariance structure models such as the ones graphically represented in Figure 2 (based on Hershberger, 1994). These models are empirically equivalent. This means that if one of them fits the data,

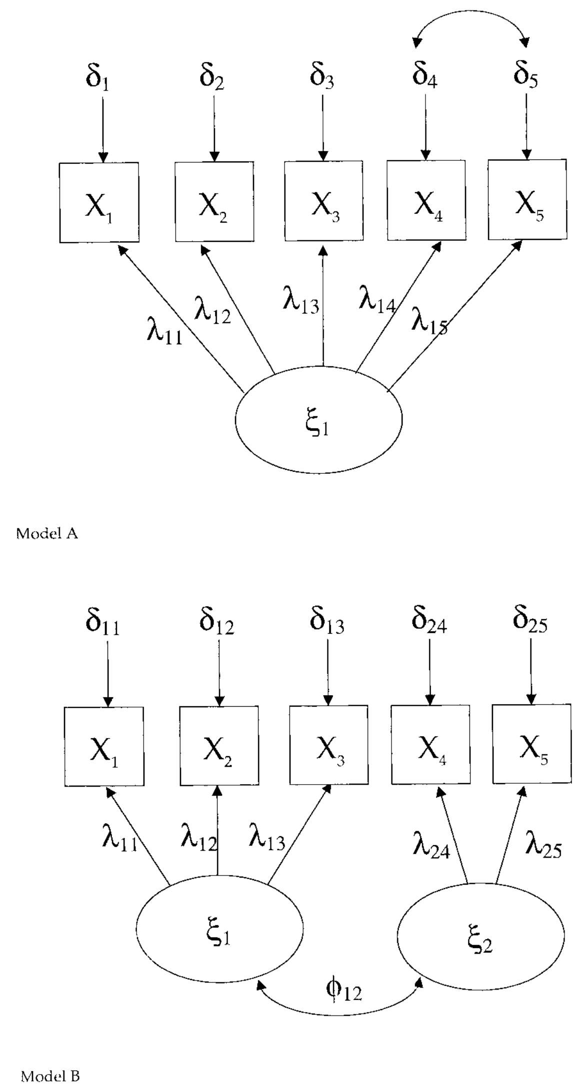

Figure 2. Two equivalent models. The structural equation models in the figure predict the same variance–covariance matrix and are thus empirically equivalent. Xs indicate observed variables, ξs are latent variables, λs are factor loadings, δs are error terms, and ϕ is the correlation between latent variables.

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.

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 in 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 our view, this is not particularly enlightening. But it must be said that the situation need not necessarily bother the constructivist, because 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 summary, 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 would be difficult to interpret these procedures without an appeal to some sort of correspondence truth. However, what we have shown is only that the natural interpretation of what one is 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, we 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 people construct this relation themselves; it is not an actual but a mental relation, revealing the structure of the 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., Edwards & Bagozzi, 2000; Glymour, 2001; Pearl, 1999, 2000), and we 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, we will concentrate on the form of the relation in a standard measurement model. Specifically, we 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.

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 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) in 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). We 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 independently 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.

The individual account of causality, however, is problematic. Consider the statement that Subject A's position on the latent variable causes Subject A'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 A'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.

The obvious way out of this issue is to invoke a counterfactual account of causation (see, e.g., Lewis, 1973; Sobel, 1994). With 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, "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 A had had a different position on the latent variable, Subject A 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 statement, “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 statement as follows: “If Einstein had had John’s level of intelligence, he would have scored lower on the IQ items.” But does this counterfactual statement express the causal efficacy of intelligence within Einstein? It seems to us 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 fruit fly, 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 fruit flies. We know that fruit flies act rather stupidly, and so are inclined to agree that Einstein would act equally stupidly if he had the intelligence of a fruit fly. 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 statement is completely unintelligible except when interpreted as expressing knowledge concerning the difference between human beings (a population) and fruit flies (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 contains only the information that Einstein will score higher on an IQ test than a fruit fly because he is more intelligent than a fruit fly—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 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 statement (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 *n*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 our 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.

But what about variables like height? Is it not unreasonable to say, “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). 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-subjects 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 between-subjects variability are qualitatively the same: If we give Einstein platform shoes that make him taller, he is, in all relevant respects, exchangeable with the taller person in the example. We 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 true 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 because 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; moreover, 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 have 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 important, 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, our case against the individual causal account would be reduced from a substantial objection to philosophical hairsplitting. However, the required equivalence has not been shown, and the following reasons lead us 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 we have been discussing in the stochastic subject formulation of latent variable models, in which the model for the individual is counterfactually defined in terms of repeated measurements with intermediate brainwashing. We have already mentioned that Ellis and 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 although 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 and Ostendorf (1998). Mischel and Shoda (1998), 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 that, 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 (e.g., 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 motoric processes that 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.

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. We 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 on the basis of a between-subjects model does not refer to the same process in all subjects. This contrasts sharply with measures of, say, temperature, in which the same process is responsible for different readings on a thermometer. In such a case, the position on the latent variable could be taken as a proxy for a process, and the causal explanation of observed scores in terms of a latent variable could be viewed as an elliptical explanation.

In psychological measurement, however, such an elliptical explanation would 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, whereas Peter has difficulties in imagining spatial rotations). It is obvious that this problem is even more serious in personality testing, in which one generally does 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 shorthand for whatever process leads to person's response. In such an interpretation, the within-subject causal account would be trivially true, but uninformative.

On the basis of this analysis, we must conclude that the within-subjects causal statement, that Subject A'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.

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 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 increase if you undergo a sex change operation. It is 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.

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, we 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 to substantiate conclusions concerning these 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 perhaps 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 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 to both 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 and Van den Wollenberg (1993) we call this type of construct *locally homogeneous*, in which *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 homogeneous constructs will not often be encountered in psychology in which myriads of individual differences can be expected to be the rule rather than the exception. We 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, whereas 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 whether it 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. We emphasize 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. We 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, whereas locally heterogeneous constructs do not. In applications, it is imperative to find out which of the two are being discussed, especially when we are testing hypotheses concerning processes at the individual level with between-subjects models.

It will be immediately obvious that constructs that are hypothesized as stable traits, such as the factors in the five factor model, are not expected to exhibit either of these structures. If a trait is highly stable, covariation of repeated measurements will not obey a latent variable model at all. Most variance of the observed variables will be error variance, so that this implies that these observed variables will be almost independent over time. This hypothesis could and should be tested using time series analysis (for the five factor model, the data of Borkenau & Ostendorf, 1998, actually seem to reject it). If it holds, the latent variable in question would be one that produces between-subjects variability but does no work at the individual level. We call this type of construct a *locally irrelevant construct*. This terminology should not be taken to imply 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 one 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 focusing almost exclusively on between-subjects models. It seems that psychological research has adapted to the limitations of common statistical procedures (e.g., by abandoning case studies because analysis of variance requires sample sizes larger than 1) 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 (Durbin & Koopman, 2001). 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 (Collins & Sayer, 2001, provide an informative overview; for further information, see, e.g., Fischer & Parzer, 1991; McArdle, 1987; Molenaar, 1985; and Wilson, 1989). Powerful software for estimating and testing these models has been developed (Jöreskog & Sörbom, 1993; Muthén & Muthén, 1998; Neale, 1999), which makes this type of analysis relatively accessible to nonstatisticians. It would be especially 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.

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. We have shown 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) that 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. We 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, we 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 paragraph may therefore be read as a restatement of his call for integration of these schools.

Discussion

In this article, we have inquired what philosophical position is implied by latent variable theory. One may reframe this question as the question of whether latent variable models are philosophically neutral. It has been argued that this is not the case. The mathematical and empirical connotations of the latent variable may be considered neutral. In a sense, neither requires the word *latent*; the formal latent variable is a mathematical concept, and the operational latent variable is a weighted sumscore. It is in the connection between these two concepts when we use the syntax of latent variable theory to estimate something with the weighted sumscore that the theory takes side with realism. Entity realism about latent variables is needed to motivate the choice for the reflective model over the formative model. Theory realism follows from the observation that the formal side of the theory implies that it is possible to be wrong about the position of a subject on the latent variable, and that weaker formulations—using empirical adequacy instead of truth—are difficult to interpret. Finally, in a standard measurement model, the causal ingredient of realism can be defended in a between-subjects sense but not in a within-subject sense. The within-subjects causal interpretation may be viewed as a fallacious application of between-subjects results to individuals. To substantiate causal conclusions at the level of the individual, one must investigate patterns of covariation at the individual level, that is, one must fit within-subject latent variable models to repeated measurements in the sense of Cattell and Cross (1952) and Molenaar (1985).

On the basis of this line of thinking, the possible relations between within-subjects models and between-subjects models were used as the foundation for a classification of psychological constructs as locally homogeneous, locally heterogeneous, and locally irrelevant. The main implication of this analysis for psychological research is as simple as it is instructive: If one wants to know what happens in a person, one must study that person. This requires representing individual processes where they belong, namely at the level of the individual. On the other hand, if the study of the individual is dismissed as too difficult, too labor intensive, or simply as irrelevant, one cannot expect between-subjects analyses to miraculously yield information at this level.

Before we discuss some implications of these results, there are two important asides to make concerning what we are 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 we 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 article. 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 we are 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 we are 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, e.g., Cartwright, 1983; Devitt, 1991; Hacking, 1983; Van Fraassen, 1980). On the epistemological side of the problem, there are probably few latent entities in psychology that fulfill the epistemological demands of realists such as Hacking (1983).

The realism implicit in latent variables analysis resides in the hypothetical side of the argument. Here, the theory cannot do without theory realism. The assumption that a model is true must be taken literally, more literally, perhaps, than many latent variables theorists would be comfortable with. However, to do science means one has to immerse oneself in the scientific world picture—a fact that is admitted even by such antirealists as Van Fraassen (1980)—and that world picture is thoroughly realist. It does not mean that—in a rather trivial way—latent variables exist by fiat, as they would in a constructivist account. On the contrary, from the realist viewpoint, the existence of latent entities is an assumption that may or may not be fulfilled, and assuming their existence could be regarded as an "as if" approach to the data. This may be considered analogous to, for example, the treatment of data as if they were the result of random sampling; random sampling is extremely rare (if it exists at all), but the bulk of statistical analyses assume it. As a result, a researcher will approach the data as if they were the result of a random sampling procedure.

It will be felt that there are certain tensions in this article. We have not tried to cover these up, because we think they are indicative of some fundamental problems in psychological measurement and require a clear articulation. The realist interpretation of latent variable theory seems to lead to conclusions that we are not willing to draw. Psychology has a strong empiricist tradition, and we 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 variables 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 consistently formulated only 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, on 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.

The need to extend the conceptual framework of psychology by linking individual processes to between-subjects comparisons has been emphasized by a number of psychologists, for example by Sternberg (1985) in the context of intelligence research and by Eysenck and Eysenck (1985) in the field of personality theories. The need for models that can incorporate individual processes has also been acknowledged by psychometricians, such as Goldstein and Wood (1989). Modeling individual processes and linking them to between-subjects latent variables is possible and has become a growing field in psychometrics (Collins & Sayer, 2001; Embretson, 1994; Fischer & Parzer, 1991; McArdle, 1987; Molenaar, 1985; Wilson, 1989). These developments are promising, and we have indicated a number of ways in which research into latent variables structures could benefit from making the connection between individual processes and between-subjects latent variables. It is clear that such research will often have to involve the analysis of repeated measurements of individuals, because it is imperative to ascertain whether our constructs are locally homogeneous, locally heterogeneous, or locally irrelevant. Theory formation could also benefit greatly from an analysis along these lines, for in many fields, it is unclear what role psychological constructs play at the level of the individual. So, there is a substantial amount of work to do, both in theoretical analysis and in empirical research. For now, we have to acknowledge that individual processes are not represented in our standard measurement models, but we hope that, with respect to this issue, this article will soon be outdated.

References

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

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

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 (Eds.), *Statistical theories of mental test scores* (pp. 397–479). 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. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology, 53,* 605–634.

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

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.

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

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

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

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

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.

Collins, L. M., & Sayer, A. G. (Eds.). (2001). *New methods for the analysis of change.* Washington, DC: American Psychological Association.

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

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

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

Durbin, J., & Koopman, S. J. (2001). *Time series analysis by state space methods.* Oxford, England: Oxford University Press.

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

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.

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.

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

Eysenck, H. J., & Eysenck, M. W. (1985). *Personality and individual differences: A natural science approach.* New York: Plenum.

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.

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.

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

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

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

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, England: Cambridge University Press.

Hacking, I. (1983). *Representing and intervening.* Cambridge, England: Cambridge University 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: Applications for developmental research* (pp. 68–108). Thousand Oaks, CA: Sage.

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.

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.

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

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement* (Vol. 1). New York: Academic Press.

Lamiell, J. T. (1987). *The psychology of personality: An epistemological inquiry.* New York: Columbia University 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., & Henry, N. W. (1968). *Latent structure analysis.* Boston: Houghton Mifflin.

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

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

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

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

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

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

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

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.

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

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: Erlbaum.

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

Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin, 115,* 300–307.

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

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.

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

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

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. (1971). *A theory and procedure of scale analysis with applications in political research.* The Hague, the Netherlands: 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 life, behavioural and social sciences* (pp. 143–167). Thousand Oaks, CA: 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, the Netherlands: Kluwer.

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: Applications for developmental research* (pp. 226–242). Thousand Oaks, CA: Sage.

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: Muthén & Muthén.

Neale, M. C. (1999). *Mx: Statistical modeling* [Computer software and manual]. Retrieved from http://www.vcu.edu/mx

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

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

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 life, behavioural and social sciences* (pp. 240–284). Thousand Oaks, CA: Sage.

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

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

Popper, K. R. (1963). *Conjectures and refutations.* London: Routledge and Kegan Paul.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Paedagogiske Institut.
Reichenbach, H. (1956). *The direction of time.* Berkeley: University of California Press.
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.
Russell, J. A., & Carroll, J. M. (1999). On the bipolarity of positive and negative affect. *Psychological Bulletin, 125,* 3–30.
Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph, 17.*
Sobel, M. E. (1994). Causal inference in latent variable models. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 3–35). Thousand Oakes, CA: Sage.
Sörbom, D. (1974). A general method for studying differences in factor means and factor structures between groups. *Psychometrika, 55,* 229–239.
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, England: Cambridge University Press.
Suppe, F. (1974). *The structure of scientific theories.* Urbana: University of Illinois Press.
Suppes, P., & Zanotti, M. (1981). When are probabilistic explanations possible? *Synthese, 48,* 191–199.
Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52,* 393–408.
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.
Trout, J. D. (1999). Measurement. In W. H. Newton-Smith (Ed.), *A companion to the philosophy of science* (pp. 265–277). Oxford, England: Blackwell.
Van Fraassen, B. C. (1980). *The scientific image.* Oxford, England: Clarendon.
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.
Wilson, M. (1989). Saltus: A psychometric model of discontinuity in cognitive development. *Psychological Bulletin, 105,* 276–289.
Wood, R. (1978). Fitting the Rasch model: A heady tale. *British Journal of Mathematical and Statistical Psychology, 31,* 27–32.

Received December 8, 2000
Revision received June 14, 2002
Accepted June 17, 2002 ■