Conceptual issues in psychological measurement

Borsboom, D.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
3. LATENT VARIABLES

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

3.1 Introduction

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

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

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

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

The growing use of latent variables analysis in psychological research is interesting from a philosophical point of view, exactly because latent variable theory, in contrast to classical test theory, is typically aimed at constructing an explanatory model to account for relations in the data. This means that explanations that make use of unobservable theoretical entities are increasingly entertained in psychology. As a consequence, the latent variable has come to play a substantial role in the explanatory structure of psychological theories. Now, concepts closely related to the latent variable have been discussed extensively. These concepts include the meaning of the arrows in diagrams of structural equation modeling (see, for example, Sobel, 1994; Pearl, 1999; Edwards & Bagozzi, 2000), the status of true scores (Klein & Cleary, 1967; Lord & Novick, 1968; Lumsden, 1976), definitions of latent variables (Bentler, 1982; Bollen, 2002), specific instances of latent variables such as the Big Five Factors in personality research (Lamiell, 1987; Pervin, 1994), and the trait approach in general (Mischel, 1968; 1973). Also, the status of unobservable entities is one of the major recurrent themes in the philosophy of science of the past century, where battles were fought over the conceptual status of unobservable
3.2 Three perspectives on latent variables

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

3.2.1 The formal stance

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

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

(3.1)
where \( g \) is a link function, \( \mathcal{E}(U_{ij}) \) is interpreted either as the expected item response of subject \( i \) on item \( j \), or as the expectation of the item response in a population of subjects with position \( \theta_i \) on the latent variable, and \( \alpha_j \) and \( \beta_j \) are an item-specific regression weight and intercept term, respectively.

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

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

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

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

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

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

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

Now there are two things we can do on the basis of our set of assumptions. First, we can determine how observed scores would behave if they were generated under
3.2 Three perspectives on latent variables

our model (this applies not only to mathematical derivations but also to simulation studies). Second, we can develop plausible procedures to estimate parameters in the model on the basis of manifest scores, given the assumption that these scores were generated by our model. It is sometimes implicitly suggested that the formal derivations tell us something about reality, but this is not the case. Each supposition ‘inside’ the formal system is a tautology, and tautologies in themselves cannot tell us anything about the world. So this is all in the syntactic domain, that is, it has no meaning outside the formal theory. Let us denote the latent variable as it appears in this formal stance (that is, the concept indicated by \( \theta \), in the IRT literature, or by \( \xi \), in the SEM literature) as the formal latent variable.

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

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

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

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

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

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

individuals, and even less about item response processes. This will prove crucially important for the ontology of latent variables, to be discussed later in this chapter.

3.2.2 The empirical stance

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

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

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

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

### 3.2.3 The ontological stance

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

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

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

Operationalism and the numerical trick

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

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

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

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 con-
structivist's denial of one or more of the realist claims. Realism exists in a number
of forms, but a realist will in general maintain one or several of the following theses
(Hacking, 1983; Devitt, 1991). First, there is realism about theories: the core thesis
of this view is that theories are either true or false. Second, one can be a realist
about the entities that figure in scientific theories: the core thesis of this view is
that at least some theoretical entities exist. Third, realism is typically associated
with causality: theoretical entities are causally responsible for observed phenomena.
These three ingredients of realism offer a simple explanation for the success of sci-
ence: 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 ques-
tion is whether a realist commitment is implied in latent variables analysis. It will
be argued that this is the case: latent variable theory maintains both theses in the
set of assumptions underlying the theory.

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

Entity realism

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

---

1 This should not be read as a value judgement. Data reduction techniques are very important,
especially in the exploratory phases of research. The fact that these techniques are important,
however, does not entail that they are not ad hoc.
Figure 3.1. (The symbolic language is borrowed from the structural equation modeling literature, but the structure of the model generalizes to IRT and other latent variable models.) The model specifies that the pattern of covariation between the indicators can be fully explained by a regression of the indicators on the latent variable, which implies that the indicators are independent after conditioning on the latent variable (this is the assumption of local independence). An example of the model in the left panel of the figure would be a measurement model for, say, dominance, where the indicators are item responses on items like ‘I would like a job where I have power over others’, ‘I would make a good military leader’, and ‘I try to control others’. Such a model is called a reflective model (Edwards & Bagozzi, 2000), and it is the standard latent variable model in psychology - employed in prominent models such as the general intelligence and Big Five models. An alternative model, that is more customary in sociological and economical modeling, is the model in the right panel of Figure 3.1. In this model, called a formative model, the latent variable is regressed on its indicators. An example of a formative model is the measurement model for social economic status (SES). In such a model a researcher would, for example, record the variables income, educational level, and neighborhood as indicators of SES.

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

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

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

\[\text{It is in itself an interesting (and neglected) question where to draw the line separating these classes of models at a content-level. For example, which of the formal models should be applied to the relation between diagnostic criteria and mental disorders in the DSM?}\]
could also test whether SES acts as a single variable in predicting health. In fact, this predictive value would be the main motivation for conceptualizing SES as a single latent variable. However, nowhere in this development have we been forced to admit that SES exists independent of our measurements.

Figure 3.1. Two models for measurement. The figure in the left panel is the reflective measurement model. The $X$'s are observed variables, $\xi$ is the latent variable, $\lambda$'s are factor loadings and the $\delta$'s are error terms. The right panel shows the formative model. The latent variable is denoted $\eta$, the $\gamma$'s are the weights of the indicators, and $\zeta$ is a residual term.

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

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

Theory realism

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

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

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

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

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

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

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

For example, take a simple case of statistically equivalent covariance structure models such as the ones graphically represented in Figure 3.2. (taken from Hershberger, 1994). These models are empirically equivalent. This means that, if one of them fits the data, the other will fit the data equally well. If the assumption that model A is true is restated as the assumption that it is empirically adequate (i.e., it fits the item responses perfectly at the population level), the assumption that model A is true is fully equivalent to the assumption that model B is true.
Figure 3.2. Two equivalent models. The SEM-models in the figure predict the same variance-covariance matrix and are thus empirically equivalent. X’s indicate observed variables, ξ’s latent variables, λ’s are factor loadings, δ’s error terms, and φ is the correlation between latent variables.

Now try to reconstruct the estimation procedure. The estimation of the correlation between the latent variables ξ₁ and ξ₂ 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,
3.2 Three perspectives on latent variables

This would mean that the correlation between the latent variables $\xi_1$ and $\xi_2$ can be estimated under the assumption that model A is true. In model A, however, there is only one latent variable. It follows that, upon the empirical adequacy view, the correlation between two latent variables can be estimated under the assumption that there is only one latent variable underlying the measurements. In my view, this is not particularly enlightening. But it must be said that the situation need not necessarily bother the constructivist, since the constructivist did not entertain a realist interpretation of these latent variables in the first place. However, it would take some ingenious arguments to defend this interpretation.

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

Causality

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

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

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

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

**Within-subject causal accounts**  The individual account of causality is problematic. Consider the statement that subject i’s position on the latent variable causes subject i’s item response. The main problem here is the following. One of the essential ingredients of causality is covariation. All theories of causality use this concept, be it in a real or in a counterfactual manner. If X is to cause Y, X and Y should covary. If there is no covariation, there cannot be causation (the reverse is of course not the case). One can say, for example, that striking a match caused the house to burn down. One of the reasons that this is possible, is that a change in X (the condition of the match) precedes a change in Y (the condition of the house). One cannot say, however, that subject i’s latent variable value caused his item responses, because there is no covariation between his position on the latent variable and his item responses. An individual’s position on the latent variable is, in a standard measurement model, conceptualized as a constant, and a constant cannot be a cause. The same point is made in a more general context by Holland (1986) when he says that an attribute cannot be a cause.
Counterfactuals The obvious way out of this issue is to invoke a counterfactual account of causation (see, for example, Lewis, 1973; Sobel, 1994). On this account, one analyzes causality using counterfactual alternatives. This is done by constructing arguments such as ‘X caused Y, because if X had not happened, ceteris paribus, Y would not have happened’. This is called a counterfactual account because X did in fact happen. For the previous example, one would have to say that ‘the striking of the match caused the house to burn down, because the house would not have burned down if the match had not been struck’. For our problem, however, this account of causality does not really help. Of course, we could construct sentences like ‘if subject \(i\) had had a different position on the latent variable, subject \(i\) would have produced different item responses’, but this raises some difficult problems. Suppose, for example, that one has administered Einstein a number of IQ-items. Consider the counterfactual ‘if Einstein had been less intelligent, he would have scored lower on the IQ-items’. This seems like a plausible formulation of the hypothesis tested in a between-subjects model, and it also seems as if it adequately expresses the causal efficacy of Einstein’s intelligence, but there are strong reasons for doubting whether this is the case. For example, we may reformulate the above counterfactual as ‘if Einstein had had John’s level of intelligence, he would have scored lower on the IQ-items’. But does this counterfactual express the causal efficacy of intelligence within Einstein? It seems to me that what we express here is not a within-subject causal statement at all, but a between-subjects conclusion in disguise, namely, the conclusion that Einstein scored higher than John because he is more intelligent than John. Similarly, ‘if Einstein had had the intelligence of a fruitfly, he would not have been able to answer the IQ-items correctly’ does not express the causal efficacy of Einstein’s intelligence, but the difference between the population of humans and the population of fruitflies. We know that fruitflies act rather stupidly, and so are inclined to agree that Einstein would act equally stupidly if he had the intelligence of a fruitfly. And it seems as if this line of reasoning conveys the idea that Einstein’s intelligence has some kind of causal efficacy. However, the counterfactual is completely unintelligible except when interpreted as expressing knowledge concerning the difference between human beings (a population) and fruitflies (another population). It does not contain information on the structure of Einstein’s intellect, and much less on the alleged causal power of Einstein’s intelligence. It only contains the information that Einstein will score higher on an IQ-test than a fruitfly because he is more intelligent than a fruitfly – but this is exactly the between-subjects formulation of the causal account. Clearly, the individual causal account transfers knowledge of between-subjects differences to the individual, and posits a variable that is defined between-subjects as a causal force within-subjects.

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

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

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

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

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

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

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

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

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

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

This is a problem for an elliptical account. For instance, one can say that the Titanic has rusted after so many years on the bottom of the sea, because it was made of iron. This explanation is elliptical, because it does not specify all processes that actually lead to the phenomenon we call rust. The reason why the explanation works, however, is that the explanation subsumes the Titanic under the category of iron things, and this category is homogeneous with respect to the processes that will occur when such things are left on the bottom of the ocean. Thus, one may look up the details of the reaction between Fe and H\textsubscript{2}O that leads to rust, and unproblematically take these processes to apply to the Titanic. One could say that the category of iron things displays *process homogeneity* with respect to the situation at hand.

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

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

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

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

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

Locally homogeneous constructs  First, theoretical considerations may suggest that a latent variable is at the appropriate level of explanation for both between-subjects and within-subjects differences. Examples of psychological constructs that could be conceptualized in this manner are various types of state-variables such as mood, arousal, or anxiety, and maybe some attitudes. That is, it may be hypothesized, for differences in the state variable ‘arousal’, that the dimension on which I differ from myself over time, and the dimension on which I differ from other people at a given time point, are the same. If this is the case, the latent variable model that explains within-subjects differences over time must be the same model as the model that explains between-subjects differences. Fitting latent variable models to time series data for a single subject is possible (Molenaar, 1985), and such techniques suggest exploring statistical analyses of case studies in order to see whether the structure of the within-subject latent variable model matches between-subjects latent variables models. If this is the case, there is support for the idea that we are talking about a dimension that pertains both to variability within a subject and between-subjects variability. Possible states of a given individual would then match possible states of different individuals, which means that, in relevant respects, the exchangeability condition discussed in the previous section holds. Thus, in this situation we may say that a latent variable does explanatory work both at the within-subject and the between-subjects level, and a causal account may be set up at both of these. Following the terminology introduced by Ellis & Van den Wollenberg (1993) I propose to call this type of construct locally homogeneous, where ‘locally’ indicates that the latent variable structure pertains to the level of the individual, and ‘homogeneous’ refers to the fact that this structure is the same for each individual.
Locally heterogeneous constructs  Locally homogeneous constructs will not often be encountered in psychology, where myriads of individual differences can be expected to be the rule rather than the exception. I would not be surprised if, for the majority of constructs, time series analyses on individual subjects would indicate that different people exhibit different patterns of change over time, which are governed by different latent variable structures. So, for some people, psychological distress may be unidimensional, while for others it may be multidimensional. If this is the case, it would seem that we cannot lump these people together in between-subjects models to test hypotheses concerning psychological processes, for they would constitute a heterogeneous population in a theoretically important sense. At present, however, we do not know how often and to what degree such a situation occurs, which makes this one of the big unknowns in psychology. This is because there is an almost universal - but surprisingly silent - reliance on what may be called a uniformity of nature assumption in doing between-subjects analyses; the relation between mechanisms that operate at the level of the individual and models that explain variation between individuals is often taken for granted, rather than investigated. For example, in the attitude literature (Cacioppo & Berntson, 1999; Russell & Carroll, 1999) there is currently a debate on whether the affective component of attitudes is produced by a singular mechanism, which would produce a bipolar attitude structure (with positive and negative affect as two ends of a single continuum), or should be conceptualized as consisting of two relatively independent mechanisms (one for positive, and one for negative affect). This debate is characterized by a strong uniformity assumption: It either is a singular dimension (for everyone), or we have two relatively independent subsystems (for everyone). It is, however, not obvious that the affect system should be the same for all individuals; for it may turn out that the affective component in attitudes is unidimensional for some people but not for others. It must be emphasized that such a finding would not render the concept of attitude obsolete; but clearly, a construct governed by different latent variable models within different individuals will have to play a different role in psychological theories than a locally homogeneous construct. I propose to call such constructs locally heterogeneous. Locally heterogeneous constructs may have a clear dimensional structure between subjects, but they pertain to different structures at the level of individuals. Thus, we now have a distinction between two types of constructs: locally homogeneous constructs, for which the latent dimension is the same within and between subjects, and locally heterogeneous constructs, for which this is not the case. Locally homogeneous constructs allow for testing hypotheses concerning individual processes, modules, and subsystems, through the analysis of between-subjects variability, while locally heterogeneous constructs do not. In applications, it is imperative that we find out about which of the two we are talking, especially when we are testing hypotheses concerning processes at the individual level with between-subjects models.

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

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

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

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

3.4 Discussion

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

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

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

What are the consequences of this problem for the conception of psychological measurement that latent variable theory offers? It depends on how you look at it. If one accepts the possibility that a causal account can apply to characteristics of populations, without applying to each element of these populations, the problems
Latent variables

are relatively small. Such causal accounts are not uncommon: Variation in the
variable ‘smoking’ causes variation in the variable ‘cancer’, but it does not do so
for each person. Still, I think that causality can be meaningfully applied in this
case, be it with the understanding that its validity at the population level does not
imply that the causal relation holds for each individual. Upon such a view, one
does have to settle for a measurement relation that is solely expressed in terms of
variation: Variation on the latent variable causes variation on the observables, but
for a single person the latent variable does not have to play a role in this respect.
One could argue against this view by saying that, if a causal model is invalid for
each individual, then it cannot be valid in the population. Upon this view, a causal
account of the measurement process is impossible in the locally heterogeneous and
locally irrelevant cases. I think such a view is too restrictive, because it would imply
that it is impossible to measure between-subjects differences in attributes, if these
attributes are inherently stable within-subjects. This would mean, for instance, that
genotypic differences cannot be measured through phenotypic effects. However, if
the purpose of a measurement procedure is to measure differences between subjects,
then one cannot hold it against the procedure that its results do not apply to
differences within subjects. It does seem that these are radically different levels of
explanation, and therefore they should not be mixed up.

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

For now, I contend that latent variable theory can offer a quite elegant account
of the measurement process. The theory has several notable benefits. First, it
places the attribute in the measurement model in a way that seems very plausible:
Differences in the attribute (either within or between subjects) lead to differences
in the observations. It is clear that such a view requires both realism about the at­
tribute and a causal interpretation of the measurement process. Second, although
this view introduces some heavy metaphysics, the metaphysics are clearly neces­sary, serve a clear purpose, and in fact lead to some interesting research questions.
This is a substantial improvement over the classical test theory model, which has
metaphysics wandering all over the place for no clear purpose except to be able to
construct mathematically simple equations. Third, the latent variable view seems
to align closely with the way many working researchers think about measurement.
This property cannot be ascribed to the classical test model, and neither to the fundamental measurement model, as will be argued in the next chapter. The latent variable model is, of course, in danger of misinterpretation. However, if the fact that a technique is easily misinterpreted were to be held against it, methodology and statistics would probably be empty within a day. At present, latent variable theory must be considered to formulate a plausible philosophy of measurement.