Conceptual issues in psychological measurement

Borsboom, D.

Publication date
2003

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.
5. RELATIONS BETWEEN THE MODELS

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

5.1 Introduction

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

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

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

5.2 Levels of connection

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

5.2.1 Syntax

Latent variables and true scores Syntactically, the true score model and the latent variable model are closely connected. In fact, they are so closely connected that the distinction between true scores and latent variables may get blurred in certain situations. It is suggested by Schmidt & Hunter (1999, p. 185), for example, that the relation between true scores and latent variables is ‘usually close enough to linear’ so that the latent variables approach has no conceptual or practical advantage. This is not the case, because whether there is any relation in the first place depends on the dimensionality of the latent variable model, which is not tested in the classical test model. The mistake made by Schmidt & Hunter (1999) is understandable,
5.2 Levels of connection

however, because if a unidimensional model holds then it will often be possible to construct a simple sumscore that can reasonably be used as a proxy for the latent variable in question.

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

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

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

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

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

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

where $P(U_{ij})$ is the probability of a correct or affirmative answer and $\beta_j$ is the location of item $j$, conceptualized as the point on the $\theta$ scale where $P(U_{ij}) = 0.5$.

Now, a monotonic transformation of the item response probabilities will yield a simple additive representation. Specifically, the model can be rewritten as

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

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

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

Second, the independence condition holds. That is, item difficulty and person ability are seen as the two independent variables, and items and subjects are independently ordered on ability and difficulty, respectively, by the dependent variable $P(U_{ij})$. Rasch (1960) actually derived the model from the requirement of parameter separation, i.e., it should be possible to estimate the ordering of items and subjects independently, which basically comes down to the same type of requirement as posed by the independence axiom in the additive conjoint model. Rasch called this property specific objectivity. Statistically, this implies that the item and person parameters can be estimated independently, because the sumscore is a minimally sufficient statistic for the person parameter, which enables parameter estimation by Conditional Maximum Likelihood (Andersen, 1973).
Third, if the Rasch model is true, then the double cancellation condition is satisfied. If, for any three levels of ability and any three levels of item difficulty, if it is true that
\[ \theta_2 + \beta_1 \geq \theta_1 + \beta_2 \] (5.4)
and it is also true that
\[ \theta_3 + \beta_2 \geq \theta_2 + \beta_3 \] (5.5)
then
\[ \theta_2 + \beta_1 + \theta_3 + \beta_2 \geq \theta_1 + \beta_2 + \theta_2 + \beta_3, \] (5.6)
so that
\[ \theta_3 + \beta_1 \geq \theta_1 + \beta_3 \] (5.7)
and double cancellation holds. Thus, the structure of the Rasch model sustains representational measurement theory. As soon as the model is extended with a discrimination parameter, as in Birnbaum’s (1968) model, this resemblance vanishes because the independence condition will no longer hold.

**Scales and true scores** The fact that the latent variable model can be constructed from the imposition of restrictions on the relations between true scores, and the fact that additive latent variable models are special cases of representational measurement theory, suggests that appropriately constructed versions of the classical model can be written in representational form too. For instance, the true score model for tau-equivalent tests assumes that for any two true scores of person \( i \) on tests \( j \) and \( k \), denoted \( t_{ij} \) and \( t_{ik} \), it is true that \( t_{ij} = c + t_{ik} \), where \( c \) is constant over persons. The structure of the model can be written in terms of a common factor model (Jöreskog, 1971):
\[ \mathcal{E}(X_{ij}) = \nu_j + \lambda \theta_i \] (5.8)
where the \( \nu_j \) parameter is a test-specific intercept term that absorbs the effect of the constant \( c \) in the definition of tau-equivalence, \( \lambda \) is the factor loading, and \( \theta_i \) is subject \( i \)’s position on the latent variable. Because, by the definition of tau-equivalence, \( \lambda \) is constant over tests, it has no test subscript as in the congeneric model. We may set it to unity without loss of generality. This gives the additive representation
\[ \mathcal{E}(X_{ij}) = \nu_j + \theta_i. \] (5.9)
The axioms of conjoint measurement then hold for the so constructed model. The instrumental comparison is made through the true scores on the tests, as it is made through the item response probabilities in the Rasch model. The true scores induce an ordering because, like probabilities, true scores are numbers and numbers are ordered. The condition of independence holds because the item and person effects do not interact (this would occur if the factor loadings differed across items): Persons can be stochastically ordered by true scores, regardless of which test is used for this purpose, and tests can be stochastically ordered by true scores, regardless of which person is used for this purpose. That the double cancellation axiom holds is obvious, because the additive decomposition of the observed scores into a test
and person specific part guarantees this to be the case; one may follow the line of reasoning as discussed above for the Rasch model and substitute $\nu_j$ for $\beta_j$.

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

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

### 5.2.2 Semantics and ontology

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

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

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

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

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

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

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

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

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

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

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

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

For fundamental measurement theory, denying the validity of probability assignments at the individual level has no theoretical consequences. Since the model is most naturally stated in deterministic terms in the first place, the theory does not have to be modified or reinterpreted when responses are stripped of randomness at the level of the individual. Such a view does lead to the conclusion that the axioms of fundamental measurement are usually not satisfied by empirical data, either in psychology or elsewhere. This observation, of course, is hardly surprising, given the strict demands of the theory. What is interesting, however, is that the connection between probabilistic latent variable models and fundamental measurement breaks down if the propensity interpretation is denied. For instance, the stochastic dominance relations as discussed by Scheiblechner (1999) no longer apply, because they are defined on true scores, which are no longer admitted in the present interpretation. Thus, the only item response model that is properly admitted as a case of fundamental measurement in this interpretation, is the deterministic Guttman model. It is thus clear that the popular view, which holds that the Rasch model 'is' a fundamental measurement model (Perline, Wright, & Wainer, 1978; Scheiblechner, 1999; Bond & Fox, 2001), is parasitic on the stochastic subject interpretation of the item response model. Once that interpretation is denied, the Rasch model has little to do with fundamental measurement. In fact, the only thing that conjoint measurement and Rasch models have in common, in this interpretation, is additivity.
Thus, there are at least two ways of looking at the relations between the different theories of measurement discussed in this book. The similarities and dissimilarities between these models depend on a rather high level philosophical assumption, namely on whether one wants to admit propensities into psychological measurement or not. Admitting propensities gives a consistent and unified picture, in which the different approaches focus on different parts of the measurement process, but are not necessarily at odds. Denying the existence of propensities immediately destroys the classical test model, and leaves one with two models that have relatively little to do with each other.

5.3 Discussion

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

5.3.1 Theoretical status

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

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

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

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

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

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

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

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

\(^1\) This paper is included in this dissertation as Appendix B.
whether this would be generally appropriate, but that nothing in the formulation of the latent variable model precludes it. The representationalist does not seem to be in a position to take such a course of action. The qualitative relations mapped into the numerical domain in population A are different from those in population B. Because measurement is representation, it seems to me that the representationalist must say that something different is being measured in each population, not that the same attribute is being measured differently. The representationalist cannot therefore assume the kind of meaning invariance that the latent variable theorist can.

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

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

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

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

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

### 5.3.2 The interpretation of probability

The interpretation of the theoretical status of the discussed models, the theoretical terms figuring therein, and the relations between these models, were seen to depend
Relations between the models

crucially on the interpretation of probability. Obviously, neither the stochastic subject nor the repeated sampling interpretation of probability is logically imposed upon us. Can we nevertheless force a choice between these interpretations? For example, could such a choice be defended on more general metatheoretical principles?

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

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

well be that no substantive interpretation is available for that explanation, so that it remains a purely statistical oddity. Thus, although the introduction of propensity undoubtedly introduces a metaphysical element in a psychological theory, one cannot say that it should therefore be considered inadmissible, unless one holds an unduly narrow view of what is admissible in scientific research.

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

5.3.3 Experimental control and local homogeneity

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

Consider, for instance, the fundamental measurement account of length. The adequacy of this account hinges on the possibility to concatenate objects. It is paradigmatic for fundamental measurement that, if one has two objects \( a \) and \( b \) of unequal length, say \( a \leq b \), then it must always be possible to find a third object \( c \) so that the concatenation of \( a \) and \( c \) is not noticeably different (i.e., both \( \leq \) and \( \geq \) to) from \( b \). This is a prediction of what would happen upon executing a special kind of experiment. Concatenation can thus be considered an experimental manipulation of the variable length, in which the additivity of length is tested. Of course, nobody would ever carry the experiment out, because it is obvious from the outset that length supports such experimental manipulations, and most people will have an overwhelmingly strong intuition that this experimental 'hypothesis' is true (although one could, strictly speaking, doubt it; Batitsky, 1998; Rozeboom,
Similar considerations are invoked in conjoint measurement. The idea of conjoint measurement is that one can experimentally vary the levels of both independent variables and assess their effect on the dependent variable. For a trade-off to be represented, it must not only be theoretically, but experimentally possible to find a change of levels in the first factor that can undo the effect (on the dependent variable) of a change levels level in the second factor. This means it is essential for the possibility of conjoint measurement that one has experimental control over the independent variables.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

So, with respect to the relation of validity, we must conclude the following. Classical test theory does not formulate a serious account of measurement, and therefore is inadequate to deal with the question of validity. In fact, if it begins to formulate such an account, it invokes a kind of embryonic latent variable model. Latent variable theory is able, by its very conceptualization, to hold that measurement is a causal relation between the latent variable and its indicators. In fact, this is a natural interpretation of the theory, because it is virtually equivalent to a common cause model (Glymour, 2001). Representationalism works on observable relations between objects, and therefore has no place for the relation of validity: the very fact that we are supposed to be able to judge relations like ‘not noticeably longer than’ with the unaided eye, means that validity is presupposed in the model. However, upon closer inspection, representational measurement is strongly related to the requirement of experimental control; and this requirement cannot be considered to demand anything less than the possibility to intervene in a causal system. If the representationalist now drops the condition that relations between objects be ‘noticeable’, which is unrealistic in the first place, he turns out to have been hiding a latent variable model under the cloak of noticeability all this time. And if he reduces the demand for experimental control to the weaker demand that a causal relation between the attribute and its indicators hold, then he turns out to formulate virtually the same semantics of measurement as the latent variable theorist.
So, when we look upon these models in the context of validity, they converge to a surprising extent. As a prelude to the following chapter, I will now abstract what I think are good ideas from the different models. In my opinion, true score theory is wholly inadequate insofar as we are talking about measurement. It is a purely statistical theory on the behavior of (composite) random variables and, in the case of psychological testing, not a very plausible one. I think that latent variable theory has a reasonable philosophy of measurement. However, it places too much emphasis on technical requirements, such as unidimensionality. Unidimensionality is a good idea in itself because it has clear statistical implications, but I think that, in latent variable theory, it has come to occupy an unreasonably strong position. Strictly taken, unidimensionality is not a very realistic assumption to make when dealing with psychological test scores. The assumption appears moreover to be motivated by an, in itself understandable, desire to measure one thing at a time, rather than from a psychological theory that says why we should expect unidimensionality to hold in a particular testing situation. But the psychometrician’s desires would not seem to be sufficient as a motivation for an assumption as strong as unidimensionality. The unidimensionality assumption could be motivated, however, by invoking a causal relation between variation on the latent variable and variation on its indicators. In this case, one says that unidimensionality will hold if the causal relation between the latent variable and its indicators holds, if this relation is correctly specified, and if the latent variable is the only attribute that causes variation on the indicators. Unidimensionality can then be considered a specific instantiation of the common cause idea, and local independence is one of its testable consequences. The kind of causal relation I am envisioning does not require local homogeneity, for I am taking the position that one can reasonably say that variation on an attribute causes variation on the observed scores, without the attribute being a causally efficient entity at the individual level.

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

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