X

The Myth that Latent Variable Models are Models

> How does this bear on the canal theory of traits? I do not know. Mathematically the theory of resolution into general and specifics gives no indication of canals...The canals or crossroads have been put in without any indication, so far as I can see, that they have been put in any other way than we put in roads across our midwestern plains, namely, to meet the convenience or suit the fancy of the pioneers (Wilson on Kelley, 1929, p.164).

1. *The myth that latent variable models are models*

To model is, nowadays, a very popular pastime within the social and behavioural sciences. Researchers appear to model prolifically with programs such as LISREL, fit regression models of various sorts to their data, and estimate model parameters as a routine matter. The prestige of modeling likely comes from justified admiration of the mathematical models of the physical sciences. However, there exist multiple senses to term model, with some new, and, arguably, trivial, contributions to this collection having been made by the psychometrician. Care must be excercised in respect the use of the term *model*, for different senses are rightly attached to different kinds of modeling exercises, speak to different aims, and imply the right to make different kinds of claims. There are, to name but a few, scale models (e.g., architectural models), analogical models (electricity understood on the model of water pressure), and mathematical models (e.g., Kepler's laws of planetary motions), of which stochastic models (e.g., Mendelian genetic models) are a special case. And of course, there are the statistical models of the social and behavioural sciences in general, and, what concerns us here, that particular sub-species, the latent variable model. The physical sciences have employed to great success analogical, scale, and mathematical models (see Freedman, 1985, for examples, and contrasts between the models of the physical sciences, on the one hand, and the social sciences, on the other).

The classical mathematical models of the natural sciences are representers of actual or possible states of affairs, and, in the former case, often express empirical laws. Such representational models function representationally through the laying down of bridging rules, rules of correspondence, that link features of that which is modeled to the terms of the model. For example, it is known that the intensity of a magnetic field, B, moving with speed, v, is related to the intensity of the electrical field it gives rise to, E, by the equation $E=vB$. The symbols in this equation represent because they are *assigned* particular roles, namely, to stand for the intensity of the electrical field produced by a particular magnetic field, the speed at which the magnetic field is moving, and the intensity of the magnetic field, respectively. And these expressions have a sense because there exist the concepts *magnetic field intensity*, *electrical field intensity*, *speed of magnetic field*, whose correct employments are fixed by rules. The fact that these concepts have normative employments enabled the empirical study of magnetic field intensity, i.e., the phenomenon denoted by the concept *magnetic field intensity*, and, as a result of this study, the formulation of the mathematical model $E=vB$, which expresses a law *about* the relationship between the phenomena denoted by these concepts.

To create a model, T, that represents features of some state of affairs, A, involves the laying down of rules of correspondence between certain of the constituents of A, $\{a_1,a_2,...,a_p\}$, to be represented, and constituents of T, $\{t_1,t_2,...,t_p\}$. And to do this presupposes the capacity to antecedently identify the constituents of A and T involved in the correspondence relationship. This, in turn, presupooses antecedent knowledge of the rules that fix the correct employments of the concepts $\{"a"_1,"a"_2,...,"a"_p\}$ that signify the $\{a_1,a_2,...,a_p\}$ that are to be represented, and those, $\{"t_1","t_2",...,"t_p"\}$, that signify the constituents of the model. One must actually be able to single out and discuss the features of natural reality that one would like to represent. Imagine attempting to create a three-dimensional model of an amoeba without grasping what was meant by the concept *amoeba*, hence, without an understanding of the entities to which the concept is applied, hence, without possessing knowledge as to the parts of which an amoeba is comprised (to be unable to see them *as* different parts). Under these circumstances, one would have no clue as to *what* was to be modeled. Moreover, if an individual operating under these conditions offered up T_1 as a model of an amoeba, neither he, nor his audience, would have any grounds for judging whether or not it was, in fact, a model of an amoeba, let alone whether it was a satisfactory model.

Now, the existence of different senses of the term *model* and different types of modeling exercises within the various scientific practices, is, in and of itself, unproblematic. What *is* perilous is to fail to *notice* that the concept *model* has multiple, distinct senses, and, as a result, come to take the contributions possible in the employment of one type of model as deliverable through the employment of another, distinct, type of model. The Central Account portrays the latent variable model as a model in the classical sense of representer of a state of affairs. In particular, it claims that when a latent variable model describes a set of manifest variates, the model represents a particular state of affairs with respect constituents of natural reality, namely, the relationship between a detected, unobservable property/attribute (causal source) and the phenomena represented by the manifest variates that are input to the analysis. Because the latent variable model is portrayed in the Central Account as representing a state of affairs with respect constituents of natural reality, it, then, follows naturally to believe that its terms *must* denote particular constituents of natural reality. In particular, it is taken as a given that the symbol $\boldsymbol{\theta}$, the latent variate, which appears in the equations of a given model, *must* represent something. In the absence of antecedently specified rules of correspondence linking the symbol $\boldsymbol{\theta}$ with a constituent of natural reality to be represented, the CA explains that this something is an "unobservable", detected property/attribute (cause) of the phenomena represented by the manifest variates. All of this leads to the conclusion that the scores that comprise the distribution of $\boldsymbol{\theta}$ must be measurements with respect this unobservable property/attribute (causal source), and represents a misportrayal of the sense of *model* actually in play in the expression *latent variable model*.

Now, it can be argued, with varying degrees of success, depending on the details of a particular application, that the manifest variate terms of a latent variable model, $\{\mathbf{X}_1,\mathbf{X}_2,...,\mathbf{X}_p\}$, stand in correspondence to phenomena of interest, $\{a_1,a_2,...,a_p\}$, these phenomena signified by concepts $\{"a"_1,"a"_2,...,"a"_p\}$, by virtue of the fact that the scores that comprise the distributions of the $\mathbf{X}_j$ are produced by following rules, $\{r_1,r_2,...,r_p\}$, which quantify features of the a_j, j=1..p. The fact that the scores that comprise the distributions of the $\mathbf{X}_j$ are produced in accord with rules that quantify features of the a_j, establishes the set $\{\mathbf{X}_1,\mathbf{X}_2,...,\mathbf{X}_p\}$ as standing for, or representing, the $\{a_1,a_2,...,a_p\}$. However, the same argument cannot be made for the latent variate term, $\boldsymbol{\theta}$, that appears in the equations of a latent variable model. For no antecedently stated rule of score production links the symbol **θ** in the equations of a latent variable model to some consituent of natural reality of interest to the latent variable modeller, and whose representation is desired.

Latent variable modellers have simply failed to note the absence, in their employments of latent variable models, of key ingredients necessary for these statistical tools to play the role of the representational tool described in the Central Account. They call these particular statistical tools models, but, yet, have failed to note that, at the least, the sense in which they are models is markedly different from that in which the classical (representational) mathematical models of the natural sciences are models, a point Freedman (1985) has previously made. Once again, various false identifications have served to occlude these points. As was noted earlier, rather than speak directly about particular phenomena under study, and, hence, particular phenomena to be represented, latent variable modellers speak of the study of (unobservable) variates or constructs. This way of talking serves to excuse the absence of key ingredients of a true modeling exercise, and, in particular, the laying down of the rules that would be required to establish correspondence relationships involving the symbol **θ**. Bentler (1996, p.434), for example, attempts to equate the "constructs" (whatever he means by this term) of interest to the psychologist with the latent variates of latent variable models, this, as was seen earlier in this chapter, a common ploy. But why, in Bentler's view, should the psychologist be interested in unobservable constructs, a convenient mythology, rather than particular natural, behavioural phenomena? Well, because to express interest in constructs is sufficiently ambiguous to allow him to bypass the careful definitional and measurement work that is a characteristic of science and to bypass the tortorous formulation of correspondence relations that are the hallmark of true modeling exercises. It is as if by placing the symbol **θ** in the equations of a latent variable model and uttering the name of the "construct" of interest, e.g., *disordered eating*, the researcher establishes a correspondence between the two.

McDonald (1996b, p.668) has attempted to argue that latent variable models are, in fact, classically representational. He states, "Now we need a few remarks about (mathematical or nonmathematical) propositions that are offered as a *model*. Such propositions are necessarily offered as a model for...-they do not reflexively model themselves, or remain self-contained. There will be rules of correspondence between (some) features of the model, and (some) features of the modeled. In the early version of the kinetic theory of gases- gases as rigid spherical bodies in motion-the correspondence yielding the Boyle-Charles law (pressure x volume = constant x temperature) is partial, yet sufficient. (There is no requirement to account for the motions of particles, as in a Wilson cloud chamber.) In the nearest parallel in psychology, Thomson's classic derivation of Spearman's *g* from the sampling of neural bonds, the correspondence is, again, partial yet sufficient"; "But once he gives the equations as a model, the questions follow, to what do they supply a model, by what principle of correspondence, and which features of the model are in correspondence with features of the modelled?" (p.669).

McDonald's reasoning is flawed. Latent variable models are not models in the sense he desires, and "giving equations and calling them a model" hardly assures that the symbols contained within the equations correspond to anything. It is true that to construct a model *of* something one must supply rules of correspondence that link the terms of the model to the something they are to represent, but McDonald's argument affirms the consequent: He argues that because linear factor analysis is *called* a model, there *must* exist rules of correspondence linking its terms to phenomena to be represented! He suggests that the equations are layed down, and *then* "…the questions follow, to what do they supply a model, by what principle of correspondence, and which features of the model are in correspondence with features of the modelled?" (p.669). In fact, the scientist gives the terms contained in a set of equations meaning *as* representers of phenomena through the antecedent laying down of rules of correspondence. It is preposterous to imagine him guessing post-hoc what he might mean by the symbols contained in his equations. Certainly, Boyle and Charles were not the least bit perplexed by what the symbols of their law were to represent. In contrast, neither Spearman, nor any latent variable modeller since, has laid down rules of correspondence between the symbol $\boldsymbol{\theta}$ (the latent variate), which appears in the equations of latent variable models, and particular constituents of natural reality. And why would they have? They believed in the Central Account, and the Central Account tells them that, in employing a latent variable model, they are in the business of attempting to detect unobservable properties (causal sources), whose identities had to be inferred.

McDonald seems to be playing upon his readership's fervent desire to believe that latent variable modeling is proceeding along the same road of progress as taken by the natural sciences. But his likening of early factor analysis to the Boyle-Charles law has to inspire a guffaw. The symbols contained in the quantitative statement of the Boyle-Charles law represent "pressure", "volume", and "temperature", and these forces and properties are signified by the concepts *pressure*, *volume*, and *temperature*, whose correct employments are fixed by carefully formulated rules. Hence, the law does, indeed, feature carefully detailed correspondences between the symbols in terms of which it is stated, and that which the symbols are to represent. Early factor models (and later ones, too), on the other hand, were not models of anything. Spearman, for example, did not even define the concept *general intelligence*, let alone lay down a rule that fixed the employment of this concept as a denoter of some constituent of natural reality. He was thus in no position to lay down rules of correspondence between any such constituent and the symbol *g* that appeared in his factor analytic equations. Instead, he created the linear factor "model" in a badly misguided attempt to define *general intelligence*. It is then not surprising that, rather than representing some modelled feature of natural reality, the symbol *g* could only play the role of a place-holder for any *variate* that satisfied the requirements for common factor-hood as stipulated by the model equations. And, *pace* McDonald (1996), Thomson did not assign to the concept *neural bond* a sense. Thus, even if he had wanted to, he could not have layed down rules of correspondence linking, on the one hand, some phenomenon signified by the concept *neural bond* and, on the other, the symbol *g*.

McDonald's suggestion that the "...correspondence yielding the Boyle-Charles law (pressure x volume = constant x temperature) is partial, yet sufficient. (There is no requirement to account for the motions of particles, as in a Wilson cloud chamber.)" is merely misleading. McDonald portrays his example as an instance in which there exist only partial correspondences between the terms of a model and features of natural reality, while, in reality, it is simply a case in which a particular model does not represent features of natural reality that it was not *designed* to represent (which can hardly be seen as a criticism of the model). It is not that latent variable models only represent a narrow cross-section of empirical reality, as does the Boyle-Charles law, but that certain of the terms inherent to these models do not represent at all. McDonald's analyses fail because he does not give due attention to how correspondence relationships are established. The key to a correspondence relationship is that there must actually have been layed down bridging rules that establish the ways in which a symbol, t, appearing in a set of equations, is to be seen as representing some particular constituent of natural reality, φ. But to be able to designate that φ, and not some other constituent, is the constituent to be represented requires a grasp of the rules of correct employment of the concept "φ" that denotes φ. For one can establish a correspondence relationship only if one can stipulate what it is that is to be represented. It is not enough to write a symbol *g* and mouth to oneself "general intelligence".

If his linear factor analytic equations were to have played a role analogous to the Boyle-Charles law, then Spearman would have had to specify *what* was to be represented by these equations. In terms of random variates, what Spearman required was something along the following lines:

Model equations:

$\mathbf{X}_j=\lambda_j\mathbf{g}+\boldsymbol{\varepsilon}_j$, $C(\mathbf{g},\boldsymbol{\varepsilon}_j)=0$, $j=1..p$, $C(\underline{\boldsymbol{\varepsilon}},\underline{\boldsymbol{\varepsilon}}')=I$.

Terms of model:

$\mathbf{X}_j$, $j=1..p$: random variates, scores on each of which are produced by following rule r_j of score production, each r_j quantifying a feature of intellectual behaviour;

λ_j, $j=1..p$: parameters to be estimated from data in a particular application of the model;

g: a random variate scores on which are produced by following rule r_g:{ }, which produces scores denoted by the technical concept *general intelligence*;

$\boldsymbol{\varepsilon}_j$, $j=1..p$: random variates representing deviations of reality from the functional law $\mathbf{X}_j=\lambda_j\mathbf{g}$, the parameters of the distributions of the $\boldsymbol{\varepsilon}_j$ to be estimated.

Note: random variates are employed because it is claimed that the law holds in population *P* of humans. The data employed to support the claim that the model accurately describes reality are $(N\times(p+1))$ scores matrices whose columns contain scores produced by application of $\{r_1,r_2,..,r_p,r_g\}$ to the members of samples of N drawn from *P*.

Lord and Novick (1968) are also at sea on these points. They declare that "A mathematical model differs from a purely verbal model in several respects. First, it is identified with an exact mathematical system, usually of a very high order (an algebra or calculus), by which the elementary constructs may be manipulated to facilitate deductions from the model. Also, since the mathematical model is more precise, its use avoids the confusion that results from the imprecise statement of a purely verbal model....Finally, the mathematical model usually abstracts from and portrays only very limited aspects of the behavioral domain. The latent trait models, which are presented in the later chapters of this book, have in part been offered as a means of linking the more precise but limited mathematical models of behavior with the more broadly conceived psychological theories." Sadly, this is but wishful thinking. If the aim in creating latent variable models had been as described by Lord and Novick (and the history of latent variable modeling and, in particular, the evolution of the Central Account, suggests the evolution of these models is rooted in confusion, rather than carefully reasoned science), then this aim has not, and cannot be, realized. For it has never been part of the employments of latent variable models that rules are layed down to establish correspondences between the latent variate symbols that appear in these models, and particular features of the "behavioural domain." If it had been, then an entirely different brand of analysis would have been the result, and there would have been no need for the invention of the Central Account.

The great majority of the statistical models of the social and behavioural sciences are not models in the classical sense of the term, i.e., they are not representers of known states of affairs, such knowledge the product of brilliant insights about, and experiments on, natural phenomena. This should be clear from Bartholomew's (1996) definition of a statistical model as "...a statement about the joint distribution of a set of random variables. A *mathematical model*...is a set of equations relating real numbers." Such a trivial sense of the term *model* (i.e., assigning to the term a role as a designator of a mere bunch of random variates, along with some statements as to their stochastic properties) implies a brand of modeling in which representation is not the issue, thus sparing the social and behavioural scientist wishing to "model" the need to engage in the arduous tasks that are the stuff of true modeling exercises. The "models" of the behavioural and social sciences are, rather, quantitative criteria (definitions of properties of importance to a particular discipline, as when the ulcf is offered up as a definition of what is meant by "the items are unidimensional"), data smooths (processors that take data as input and present as output essentially the same information in a more manageable form), and other techniques of data reduction (processors that take data as input and produce as output a lower dimensional representation of this data).

What, then, is the latent variable model if it is not a model in the classical sense of the term? As will be argued in detail in Part III, the answer is that it is a replacement variate generator, a set of quantitative requirements imposed upon a set of input random variates that, if satisfied, allow for the replacement, in some particular sense, of the input variates by a set of constructed random variates. The symbols that stand for the latent variates are placeholders for any random variates that replace the input variates in the sense required by the equations and distributional specifications of a given generator. This fact explains Freedman's (1985) comment regarding the place of stochastics in regression models:

> There are great stochastic models in the natural sciences, and a lot of attention goes into their basic assumptions. After all, it is the assumptions which define the model. Regression models too make quite strict assumptions, explicitly or implicitly, about the stochastic nature of the world. In most social-science applications, these assumptions do not hold water. Neither do the resulting models (p.345).

Assumptions do indeed define the model. But, in regard a latent variate model, they do not constitute an attempted description of reality, or even a set of tacitly held beliefs about natural reality, but, rather, *requirements* that must be met in order to license a particular type of optimal replacement. Bartholomew's definition of statistical model betrays this brand of employment. For, if, as he claims, a statistical model is really just a statement regarding the joint distribution of a set of (input) random variables, then to create a statistical model is not to create a representation of some constituent of natural reality, but, rather, a recipe for the construction of random variates that satisfy precisely these distributional specifications. McArdle (1990, p.81) claims that "The one strength of the CFA model comes from the fact that the assumption... of uncorrelated specific factors is a falsifiable hypothesis." But this makes it sound as if there are such things as specific factors in natural reality whose properties can be checked for uncorrelatedness. In fact, what is testable is the hypothesis that a covariance matrix is decomposable as in $\Sigma = \Lambda\Phi\Lambda' + \Psi$, in which Ψ is diagonal and positive definite. Given that this hypothesis is correct, it follows that random variates called common and specific factors can be constructed. To construct such random variates, one employs the equations (4.4)-(4.6).

Consider Joreskog and Sorbom's (1985, p.386) response to Freedman's criticisms of regression models in general, and covariance structure models in particular:

> It seems that Freedman misses the main idea of latent variable modeling. The latent variables are supposed to account for the intercorrelations among the variables, in the sense that, when the latent variables are partialed out of the observed variables, there are no more intercorrelations left (see, e.g., Joreskog, 1979b). Hence, by definition, the ε are uncorrelated among themselves and also uncorrelated with the latent variables.

Freedman, however, would have been fully justified in parrying that Joreskog and Sorbom have placed the cart before the horse: If a LISREL model is to be taken seriously as a *model* in the classical sense of the term, then *it* must answer to reality, and not reality to it. But a LISREL model cannot answer to reality, for this is not its purpose. If, in a given application, a LISREL model describes a set of random variates, this simply means that the input variates can be replaced by each of a set of constructed random vectors, meaning that, to paraphrase Joreskog and Sorbom, "each constructed random vector accounts for the intercorrelations among the variates, in the sense that, when it is partialed out of the observed variates, there are no more intercorrelations left." The LISREL "model" is a generator of random vectors that possess precisely this property.