IX

Latent variate interpretation and the deeper problem

1. *CA commitments regarding conceptual signification, meaning, and measurement*

The Central Account claims that the scores that comprise the distribution of random variate $\boldsymbol{\theta}$ to $\underline{\mathbf{X}}$ are measurements with respect a psychological property/attribute that is common to (that is a causal source of) the phenomena represented by the $\mathbf{X}_j$, j=1..p. It is claimed that, due to the problem of unobservability, the latent variable modeller must make an inference as to the identity of this property/attribute (causal source). But to imply that a set of scores "has an identity" is to imply that they are signified by some concept "φ". The task that is undertaken in latent variable interpretation is then, essentially, to infer the unkown concept doing the signifying. But latent variable models are not tools of detection of properties/attributes (causal sources), and, hence, there is nothing to the belief that the scores that comprise the distribution of the random variate $\boldsymbol{\theta}$ to $\underline{\mathbf{X}}$ are measurements with respect any such a property/attribute (causal source). It follows then that the practice of latent variate interpretation lacks a purpose.

It would, however, be mistaken to see the practice of latent variate interpretation as having arisen from the simple misportrayal of the kind of tool that is a latent variable model. The Central Account is an ürbild, and an ürbild is a complex of interwoven theses and commitments. Latent variable interpretation is equally the product of misconceptions about conceptual signification and measurement that are endemic to the social and behavioural sciences, and which are tied into the Central Account ürbild. Consider, once again, a research project in which a random p-vector $\underline{\mathbf{X}}$ is distributed in a population P_T, a sample is drawn from P_T, latent variable model lvm is fit to the sample data, and it is decided that (in P_T) $\underline{\Omega}_T \subset M_{lvm}$, i.e., that $\underline{\mathbf{X}}$ is described by lvm in P_T. The latent variable modeller mistakenly believes that he has detected a property/attribute (causal source) of the phenomena represented by the set of p variates, and commences to "interpret the latent variate" in an attempt to infer the identity of this property/attribute (causal source). McDonald has described the thinking involved:

> In using the common factor model, then, with rare exceptions the researcher makes the simple heuristic assumption that two tests are correlated because they in part measure the same trait, rather than because they are determined by a common cause, or linked together in a causal sequence, or related by some other theoretical mechanism. It is a consequence of this heuristic assumption that we

> interpret a common factor as a characteristic of the examinees that the tests measure in common, and, correlatively, regard the residual (the unique factor) as a characteristic of the examinees that each test measures uniquely (McDonald, 1981, p.107)

> ...I am describing the rule of correspondence which I both recommend as the normative rule of correspondence, and conjecture to be the rule as a matter of fact followed in most applications, namely: In an application, the common factor of a set of tests/items corresponds to their common property" (McDonald, 1996b, p.670).

> ..the interpretation of a common factor in terms of the common attribute of the tests that have high loadings on it... (McDonald & Mulaik, 1979, p.298);

Now, further assume that, by studying the estimated loadings from his analysis, a particular latent variable modeller interprets the latent variate to be "dominance". The commitments inherent to this product of latent variate interpretation can be unpacked as follows:

i) The scores that comprise the distribution of random variate $\boldsymbol{\theta}$ to $\underline{\mathbf{X}}$ are measurements of dominance, or, in slightly different words, they are "dominance scores".

ii) Because (i) is asserted, and, yet, the alleged measurements contained in the distribution of $\boldsymbol{\theta}$ to $\underline{\mathbf{X}}$ were not produced by the following of rules that govern a normative practice of measuring the dominances of individuals, it is being asserted that there can be such a thing as non-normative (non-rule guided), naturally occurring, measurement.

iii) Because the possibility is entertained that the scores that comprise the distribution of random variate $\boldsymbol{\theta}$ to $\underline{\mathbf{X}}$ might be measurements of dominance (recall that the modeller also allows for the possibility that his inference is incorrect), yet these scores were not produced by the following of rules that govern a normative practice of measuring the dominances of individuals, it is an implicit commitment of this portrayal that there can be such a thing as scores that are *intrinsically* measurements of dominance, i.e., that the meaning of these scores is a feature of natural reality. They are what they are, and the researcher cannot *know* (with certainty) what

they are, any more than he can *know* (with certainty) the causal determinants of cancer. Hence, the researcher might have been incorrect in his inference that the scores are measurements of dominance. They might, instead, have been measurements of aggression, assertiveness, leadership, or something else.

iv) Because, to make the claim that "*these* scores (the scores that comprise the distribution of random variate **θ** to **X**) are measurements of dominance" is to claim that they are signified by the concept *dominance* (i.e., that he is justified in applying the concept *dominance* to these scores), and, yet, the modeler does not argue that the rules of employment of the concept *dominance* warrant such an application, it is an implicit commitment of this portrayal that the notion of "naturally occurring conceptual signification" is coherent. That is, he is arguing that the justification for the application of a concept to a set of scores is an empirical matter, lying beyond the rules of language, and having, rather, to do with the properties of the scores themselves.[1]

The work of McDonald and Mulaik on factor indeterminacy and the variate domain response to this problem is representative of these commitments. For example, in discussing the issue of a domain of variates, they state that "If the developer of the tests does this without taking account of the empirical content of the additional items...he may end up with a test whose total score measures an attribute that is somewhat different from that measured by the original core items" (McDonald & Mulaik, 1979, p.306). This quote expresses the idea that a total score **t** calculated on some set of items, is signified by some concept "γ", but that no one can be sure *which* concept is doing the signifying. The scores that comprise the distribution of **t** are, as it were, intrinsically γ-scores, and, if the test constructor is not careful, "γ" might well turn out to be the wrong concept (e.g., *anxiety* rather than the desired *depression*). In a later commentary, McDonald declares that "...the factor loadings of the known indicators provide the grounds for settling disputes over the property that the indicators share..." (1996b, p.671). Mulaik (1986, p.29) states that "Just as factors have ambiguous meanings to be attached to them as we observe their loadings on a finite set of variables, so also do words have ambiguous meanings as we seek to infer how to use them from the finite number of instances in which we see others use them."[2] Once again, it is taken as a given that the scores that comprise the distribution of **θ** to **X** are intrinsically measurements of *some* property/attribute, but that one cannot be certain about *which* property/attribute they are measurements of. The researcher must resign himself to making an

---

1 This, of course, is a commitment shared by the construct validation approach to the support of measurement claims.

2 Note that Mulaik confusedly claims that the meanings of terms must be infered. It is precisely this type of category error, and, fundamentally, misunderstandings over the nature of the relationship between conceptual and empirical issues, that has nurtured the latent variate interpretation component of the CA. The roots of these misunderstandings will be discussed in Chapter XII.

inference as to the identity of the unknown concept through an examination of "the factor loadings of the known indicators."

However, because the scores that comprise the distribution of $\boldsymbol{\theta}$ to $\underline{\mathbf{X}}$ were not produced by following rules of measurement, they must, under McDonald's portrayal, be intrinsically meaningful. They are whatever they are, dominance scores, self-esteem scores, etc., as a matter of natural reality. On this account, conceptual signification is a naturally occuring phenomenon, established outside of the linguistic practices of the humans who employ the very concepts supposedly doing the signifying. This view is evident in texts such as that of Cureton and D'Agostino (1983, p.3), and filters down to researcher friendly accounts such as that of Cooper (1998, p.213), who proclaims that "Exploratory factor analysis essentially does two things...It shows how many distinct psychological constructs (factors) are measured by a set of variables. In the above example there are two factors (size and drunkenness)...It shows which variables measure which constructs."

Commitments (i)-(iv) are part of a general picture of concept meaning and signification, and measurement, that is bound up with the Central Account, and is typically augmented by the following theses[3]:

v) Whether or not a set of scores are measurements taken with respect an ordinary language concept "κ", and, similarly, whether a rule, r, produces measurements of κ, is an empirical issue, to wit, whether such scores are, *in fact*, intrinsically κ-scores. Hence, measurement claims, e.g., that following particular rule r_a produces measurements of anxiety, are adjudicated through the construction of scientific cases involving the usual combination of conceptual, theoretical, and empirical evidence. Of special relevance to such cases is the conformity of the statistical properties of the scores produced by application of rule r_a to extant theory about anxiety in the population under study. Instances of such conformity constitute support for the measurement claim;

vi) Whether measurements can be produced with respect some concept "κ" is, similarly, an empirical issue. In particular, one must investigate, through the carrying out of empirical research, whether a property ψ denoted by ordinary language concept "ψ" is measurable. In particular, it is perfectly reasonable to attempt to discover ways to measure psychological phenomena (hopes, desires, dispositions, intellectual capacities). It is conceivable, for example, that, within the context of a latent variable analysis, the scores that comprise the distribution of random variate $\boldsymbol{\theta}$ to $\underline{\mathbf{X}}$ might turn out to be measurements of dominance, and, hence, it be discovered that the dominances of people are measureable;

3 These, too, fundametal tenets of the construct validation view of measurement.

vii) The *true* meaning of a concept is an empirical issue. At best, a consideration of language yields primitive "hunches" about the true meaning of a concept. It thus follows that scientific, empirical investigation is the correct approach to take in revealing the true meaning of a concept.

2. *Latent variate interpretation is badly confused*

The picture of concept meaning, conceptual signification, and measurement, described by (i)-(vii) is profoundly confused. To sustain this verdict, one need only consider the following characteristic features of measurement practices.[4]

i. *Behavioural practice*. A practice of measuring objects in respect some particular property has an unmistakable look. It is a cluster of related skills and activities engaged in by a community of humans (see Ter Hark, 1990, for a detailed discussion). It includes methods for measuring, rules for expressing, in appropriate units, measurements taken of that which is measured, terminology for expressing comparisons of measurements, methods for varifying the correctness of measurements and measurement claims, methods for comparing objects to canonical samples (e.g., a book to a ruler), and procedures for the translation of units of one type into units of another. However, to correctly report that "the length of the book is 16cm" does not entail a specification as to *how* the measurement was taken. Nevertheless, both the use of units of length and statements involving concepts of length presuppose the existence of a practice in which the lengths of objects are, in fact, measured.

ii. *Rule guidedness (normativity)*. A measurement practice is a normative (rule guided) practice. Rules are human created standards of correctness. They define what constitutes correct and incorrect behaviour in a practice in which behaviour *can* be classified as either correct or incorrect. Rules are often, but not necessarily, codified. Measurement practices are paradigm cases of rule guided practices. In a given measurement practice, the employment of units of measurement, the translation of one set of units into another, and the use of measurement instruments to take measurements, among other behaviours, can all be done either correctly or incorrectly. A characteristic mark of a normative practice such as a measurement practice is that the skills of which it is comprised must be mastered, and, hence, are taught and learned. The

4 This section is heavily indepted to a discussion of measurement in Baker and Hacker (1980).

correction of incorrect measuring behaviour is a characteristic accompaniment to the learning of the rules that are constitutive for measuring.

iii. *Empirical background*. Measurement practices are founded and developed against a background of natural reality. If "things" (features of natural reality) were different, so too would be our measurement practices. If, for example, the gravitional forces exerted on objects were different than they currently are, so too would be our practices for measuring the weights of objects. That is, the rules that govern this practice would not be as they currently are. The rules which fix how to measure the length of an object with a wooden metre stick by simple methods of juxtaposition presuppose particular features of natural reality (e.g., that the objects that are to be measured do not all of a sudden loose their shapes). But given the way things are, given the empirical background conditions we do in fact face, humans lay down rules that govern practices of measurement, and these rules determine what constitutes measurement.

iv. *Grammatical ties*. One juxtaposes a ruler and a book, and reports that the length of the book is 10 centimetres. A measurement claim has the general form: The σ of δ_i is t_i υs, in which σ stands for the property measured, δ_i is member i of some class, δ, of objects each possessing a σ, t_i is a measurement of the σ of object δ_i, and υ is a particular unit of measurement. Note that a measurement claim is also, inherently, a claim of conceptual signification, to wit, that concept "σ" signifies (can legitimately be applied to) score t_i. This shows the internal connection between the measurement of a property σ that is measureable and the meaning of concept "σ" involved in resulting measurement claims. In particular, such applications of a concept, hence, such measurement claims, are legitimate only if they are warranted by the rules which govern the correct employment of concept "σ".

v. *Justification of a measurement claim*. Rules fix what constitutes correct (eo ipso incorrect) measuring behaviour. Measurement claims, e.g., that t_i is the σ of δ_i in υs, are, therefore, adjudicated for their correctness by comparing behaviour (e.g., how t_i was, in fact, produced) to rules (the steps one must, in fact, take in order to produce a measurement of the σ of δ_i in υs). To justify his claim that the width of his yard is 48'5", an individual will recite the *way* in which this number was produced. It is not, however, that the individual lacks a more sophisticated justification of his measurement claim (that, e.g., he does not understand physics), but that there exists no other justification. Human installed rules of measurement are constitutive for measuring. A number arrived at by twisting a tape measure around the cherry tree four times would not count as a measurement of the width of his yard, because these actions do not accord with the rules for measuring width. We would not say that such an individual has *taken* a

measurement of the width of his yard, even if the number he came up with by traipsing around his cherry tree turned out to coincide with the width of his yard.

vi. *Autonomies*. Measurement practices manifest a range of basic autonomies that are a charactistic of all normative (rule governed) practices.

Definitions are autonomous of facts/evidence. Definitions are rules of concept use. They settle the legitimate employments of certain concepts[5]. Rules, being standards of correctness, are autonomous of facts/evidence. Consider the role of a standard metre (a canonical sample) in the practice of measuring the lengths of objects. The standard metre defines what is *meant* by one metre. The measurement of the length of an object can be carried out by juxtaposing the standard metre and the object. The result of this operation can be formulated as a measurement sentence, e.g., that the object is 32.5 inches long. But to state that the length of *this*, the standard metre, is one metre is *not* to express the result of a measurement procedure, but rather to cite a definition of the concept *one metre*.

It is not, however, that the standard metre is *really* one metre, that it was *discovered* that the object called the standard metre just happened to be, as a matter of fact, one metre long, or, conversely, that it might turn out that the standard metre is not truly one metre long. Rules fix the meanings of concepts of measurement and the standard metre is enshrined within the practice as a canonical sample for the concept *one metre*. Thus, there is no *evidence* that could refute, or support, the claim that the standard metre is one metre long. While empirical propositions can, indeed, be adjudicated for their degree of truth in light of empirical evidence, definitions, and the concepts whose meanings they fix, are characterized by, e.g., their usefulness, their coherence of formulation, etc. Definitions in general, and the definitions of metric terminology in particular, are not empirical propositions, and, hence, are autonomous of empirical evidence.

This, of course, does not mean that one cannot point to a sequence of steps that, historically, led to the establishment of a particular standard metre as the definition of *one metre*, but simply that, whatever be these steps, once a particular standard metre is installed as the definition of *one metre*, it, henceforth, settles what is meant by the concept *one metre*. Anything else that might be said on the subject is, from this point forth, for the purposes of measuring length, irrelevant. To put this differently, once a standard metre is enshrined as the definition of *one metre*, the assertion "the standard metre is one metre long" is no longer an empirical proposition (it is not a statement about a property (the length) of the stick that happens to be called the standard metre). For there exists no standard of correctness in the determination of

5 But not all concepts. The employments of psychological concepts, for example, are not fixed by simple stipulative definitions given in terms of necessary and sufficient conditions (Baker and Hacker, 1982).

length *other* than the standard metre. It is the arbiter of judgments of length, never what is judged (Baker & Hacker, 1980).

Scientific (empirical) cases can have no direct bearing on the justification of measurement claims. Measurement claims, e.g., that "*these* {178,145,176,189} are measurements of the heights of Ted, Bill, Joe, and Bob in centimetres", that "property η can be measured", that "rule r produces measurements of anxiety (in some particular units)", are not supported by the building of scientific (empirical) cases. In the first place, rules establish what is correct and what is incorrect. To portray facts and theory, rather than rules, as the arbiters of a measurement claim, to portray support for a measurement claim as inductive and probabilistic, is to remove from a measurement practice the distinction between correct and incorrect measuring behaviour. Inductive cases are always approximations to the truth, for it is never possible to grasp fully the details of natural reality. Such cases are always provisional and revisable in light of new evidence. In contrast, rules of measurement *fix* what is correct and incorrect in a measurement practice, are created by humans, and, hence, are not elements of natural reality.

Because they are created by humans, rules of measurement are open to inspection by one and all. To specify the details of correct measuring behaviour is to do nothing more than to articulate well known, publicly available rules. If this were not the case, then the details of measurement practices could not be taught and learned as a matter of basic education. In the second place, measurement claims to the effect that "*these* {178,145,176,189} are measurements of the heights of Ted, Bill, Joe, and Bob in centimetres", that some property η can be measured, that rule r produces measurements of anxiety (in some particular units), are claims of concept application (e.g., that a particular concept can be applied to the members of a set of numbers), and, hence, are initimately tied to the grammars (rules of employment) of the concepts involved. Empirical facts and theory have no direct bearing on the grammars of concepts, and, hence, no direct bearing on when a concept's application is warranted.

Non-discoverability of measurement. It is certainly possible to discover that another individual has been engaged in, say, the measurement of radioactivity or to discover that more men than women know how to measure the mass of an object. But the notion that it is possible to discover whether a particular property is measureable, or whether a particular set of scores happen to be measurements of some property, or how to measure a characteristic denoted by an ordinary language concept, is incoherent. What it is to measure something, if in fact the something is measureable, is fixed by rules. This is why it makes sense to speak of correctly (and incorrectly) measuring. Rules, however, are not constituents of natural reality, and, hence, are not open to discovery. They are standards of correctness that are layed down, and referred to, by humans. There is no sense of correct measurement of a measureable property η outside of what the rules

for the measurement of η establish as correct, and, hence, nothing to be said in regard the details of the measurement of η outside of a reiteration of these rules. Nothing about a set of scores can reveal whether or not they are measurements. The status of a set of scores *as* measurements is guaranteed by taking them *as* measurements of some particular property for which there exists a normative practice of measurement (through the active following of rules of measurement). The rules of correct employment of a concept either warrant the appearance of the concept in measurement sentences (when the concept is embedded in a practice of measurement) or they do not. Nothing external to these rules is relevant to issues of the concept's employment, and, hence, nothing else is relevant to the issue as to whether the concept can legitimately appear in measurement sentences.

But this is not to say that empirical facts and theory have no place in measurement practices, but only that they have no *direct* bearing on what it is to measure something, whether or not something can be measured, whether or not the scores generated by a certain rule are measurements with respect a concept "σ". As will be discussed in the next section, sophisticated scientific theory and knowledge play a large role in measurement practices, notably in the development of more sophisticated measurement techniques. But such theory and discovery play the role of *input* into the active laying down of new rules of measurement by humans. Once new rules are layed down, it is they that are then constitutive for measuring, and how they came to be so is irrelevant for the *justification* of measurement claims (although most certainly not irrelevant to the historian of measurement practices).

vii. *Vertical structure of measurement practices*. It is a characteristic feature of measurement practices that they are built up vertically, over time, from less to more sophisticated. When a new level is added to a practice, e.g., a technique that yields a heretofore unachievable degree of measurement precision, or a new system of units for the expression of measurements, new rules are layed down, and these rules are squared with the rules that, to date, governed the practice. This squaring of old and new rules is required if what is delivered at the more advanced stage of a practice are still to be called measurements of that which was measured at the less advanced stage.

A key feature of the English system of units of length was the imperial standard yard, which was the distance between two lines on a bronze bar made in 1845 to replace an earlier standard bar that had been destroyed by fire in 1839. Because the imperial standard yard bar had been shrinking at the rate of roughly 1.5 millionths of an inch per year, the United States adopted a copy of the international prototype meter as its national standard of length in 1889. Until 1960, the U.S. system of units was based on a standard meter (meter prototype number 27), after which the standard meter was redefined in terms of wavelengths of light from a krypton-86 source. In 1983 it was again redefined as the length of the path traveled by light in a vacuum in 1/299,792,458 of a second. Notice that to make discoveries about what light does in a vacuum

presupposes a great deal of sophisticated theory and a great deal of knowledge. However, *humans* decide what is constitutive for measuring length, and, in particular, what constitutes one meter in length. Humans, employing theory and knowledge, came to decide that "the length of the path traveled by light in a vacuum in 1/299,792,458 of a second" should be installed as a canonical recipe for the concept *one meter*. Humans decide when re-definitions are necessary, what these re-definitions will be, and how they must be squared with previous conventions.

The use of canonical samples (and canonical recipes such as "the length of the path traveled by light in a vacuum in 1/299,792,458 of a second") is a characteristic of all measurement practices, as is the production of the copies of these samples that find their way into common use (e.g., the plastic meter rulers sold in countless stationary stores). The development of more advanced canonical samples and recipes (e.g., the shift from the imperial standard yard to "the length of the path traveled of light in a vacuum in 1/299,792,458 of a second") is certainly the product of scientific advance, but, *regardless of their origin*, the status of such samples and recipes *as* canonical samples and recipes is fixed by humans through the laying down of rules. Humans, not nature, enshrined (established as a standard of correctness in a practice of measurement) "the length of the path traveled of light in a vacuum in 1/299,792,458 of a second" as the definition of *one metre*. Humans laid down the rule "by *one metre*, what is meant is "the length of the path traveled of light in a vacuum in 1/299,792,458 of a second"". It was not the case, for example, that it was *discovered* that "the length of the path traveled of light in a vacuum in 1/299,792,458 of a second" was *really* what was meant by *one meter*. Rather, it was realized that, due to the unreliability and impermanence of the various standard meters that had, in the past, been employed, a new type of standard, a canoncical recipe, was required, and, as a matter of *fact*, the length of the wavelength of light from a krypton-86 source happened to have roughly the same length as the standard meters employed to date. Whatever the reasoning that led to the adoption of the length of the wavelength of light from a krypton-86 source as a canonical recipe, once it was fixed in grammar as the meaning of *one meter in length*, it became the standard of correctness in the normative practice of measuring the lengths of objects, and no longer a thing that was, in fact (happened to be), one meter long. Innovations "... may originate in experiments, but using them as conversion principles transforms the result of an experiment into a rule" (Baker & Hacker, 1980, p.174).

In regard the practice of latent variate interpretation and associated commitments (i)-(vii):

i) There is no such thing as naturally occuring measurement, nor naturally occuring conceptual signification. The idea that a score, e.g., a score belonging to the distribution of $\boldsymbol{\theta}$ to $\underline{\mathbf{X}}$, could be, as a matter of nature, a measurement of something (akin to a rock being, as a matter of nature, a chunk of feldspar), and, correlatively, that what it is a measurement of is an empirical issue, is nonsense. A set of numbers, {1.3,4.1,26,...}, are measurements of property γ if they have been *taken* in a particular way, to wit, in accord with rules for the production of

measurements of γ (supposing, of course, that concept "γ" is embedded in a normative measurement practice). Humans do not *find*, *discover*, *happen upon*, *catch*, *harvest*, *extract*, or *mine*, but, rather, *take* measurements. It is the uniquely human act of *taking* a measurement (of following particular rules to produce a particular desired result) that confers upon a number the status of being a measurement of some particular property. Such rule following behaviour, of course, presupposes the existence of a rule governed practice of measuring.

ii) Scientific theory, knowledge, and case building have no direct relevance to the justification of a measurement claim. It is badly mistaken to suppose, as does the discipline of psychometrics, that support for a claim to the effect that, e.g., "rule r_h yields measurements of aggression", is to be found in scientific case-building. Within normative measurement practices, measurement claims are adjudicated for their correctness by comparing behaviour to rules of measurement. Measurement claims made about properties for which there does not exist a normative measurement practice are not correct or incorrect (for there exist no standards against which to compare such claims), but, rather, nonsense (literally, expressions that lack a meaning).

It is likely that the mis-step that psychometrics has taken in regard this issue is a result of its misunderstanding the nature of the dramatic role played by scientific knowledge and theory in the evolution of practices of measuring quantities such as weight, mass, density, etc. It would, for example, be foolish to suggest that, in the absence of the sophisticated theory and knowledge generated by physicists, the concept *one metre* could have been defined in terms of "the length of the wavelength of light from a krypton-86 source". Both theory and knowledge provided humanity the opportunity to formulate this definition (and many others). But, once again, it remains for humans to take the products of science as input into the laying down of rules of measurement. Humans had to take the step of enshrining "the length of the wavelength of light from a krypton-86 source" as a canonical recipe for the concept *one metre*. And once such a rule is laid down by humans, it is constitutive for correct measuring, and, hence, is a standard of correctness to which behaviour must be compared in adjudicating measurement claims (e.g., that this fishing rod is one metre long).

iii) The notion that an ordinary language concept, "π", to date not embedded in a normative practice of measurement, might be *discovered* to signify scores produced in some fashion (i.e., that it might be *discovered* that it is possible to measure the π of objects), e.g., as output from a statistical technique, is incoherent. For the correct employment of an ordinary language concept is fixed by rules, and these rules are mastered by language users in the course of learning how to correctly employ the concept. There can be no *mystery* as to the employment of a concept (in the way that there *is* mystery as to the nature of the sub-atomic realm), no mystery as to whether or not it can legitmately appear in measurement sentences (for its

grammar either does or does not warrant such appearances), and, hence, no doubt as to whether or not it is embedded in a normative practice of measurement (because normative practices are created by humans, and, hence, are known to humans). The fact that a given concept is embedded in a normative practice of measurement is not a feature of natural reality, but a feature of the concept's employment, and its employment is fixed by linguistic rules that are layed down by humans. The meanings (employments) of concepts embedded in measurement practices must be taught along with associated rules for measuring and units in which to express the results of measurement operations, for the fact that such concepts signify measurements is part of their meanings. If the concept "π" (*weight*, *height*, *mass*, *length*, *density*, *magnetism*, etc.) is embedded in a practice of measurement, language users standardly learn to employ "π" along with the various normative techniques for measuring the π of things.

It is, therefore, non-controversial to state that ordinary language psychological concepts are *not* embedded in normative measurement practices, and, hence, that the phenomena they signify are not measureable. In learning, for example, the concept *hope* (its cognates and related terms), children are not taught how to measure their own, and others', hope. They are not taught the use of units in which measurements of hope are to be expressed, nor how to calculate the difference between their own, and another's, hope (although they are taught the grounds for ascriptions such as "she is much more hopeful than he that the cabin can be saved"). Joe can take his weight in July and compare it to his weight in April, and express the comparison in a sentence such as "I have gained 6 pounds since April". But a claim such as "my dominance has increased by 6 since April" lacks a sense. The linguistic rules that fix the correct employment of the concept *dominance* assign no meaning to such an expression.

The misconceptions of psychometricians (and social and behavioural scientists) in regard the possible measureability of phenomena signified by ordinary language psychological terms is likely attributable to two facts: i) The failure to comprehend that linguistic rules are *constitutive* for the meanings of psychological concepts. To correctly claim that a concept is applied to η is to correctly claim that the rules that fix the correct employment of the concept *warrant* application of the concept to η. Hence, to claim that η is an instance of ξ is necessarily a claim about the correct employment of concept "ξ". The lesson for the psychometrician is that he cannot employ ordinary language psychological concepts in ways that depart from normative usage (e.g., in measurement sentences) and yet legitimately claim to be speaking of the phenomena that these concepts signify; ii) The mistaken belief that to create a method for the measurement of, say, hope, rests on knowing, on the hand, what hope is, and, on the other, what it is to measure something (see Baker and Hacker, 1982, p.286). This misconception explains the existence of the many texts professing to explain to the social and behavioural scientist general techniques for the construction of measurement instruments applicable to the measurement of virtually any psychological phenomenon (as if one could provide a generalized recipe for the construction of measurement instruments that, with small modifications, could be used to measure temperature, weight, length, height, etc.). As was noted earlier, the meaning of

a concept (its rules of correct employment) that is embedded in a practice of measurement (e.g., *height*, *weight*, *mass*) is not independent of details of the measurement practice itself. The practice is tied up with the rules of correct employment of the concept. Whatever it is that a psychologist is doing when he claims, e.g., to be measuring dominance with instrument A, he is not measuring dominance. He is either very confused about what he can and cannot legitimately do with the ordinary language concept *dominance*, or, without acknowledging so, is measuring an enirely different phenomena denoted by a technical homynym of the concept.

Undoubtedly, some psychometricians will object to the claim that psychological concepts are not embedded in measurement practices. They will point to the countless inventories purporting to measure "constructs" such as *anxiety*, *depression*, and *agreeableness*, the multitude of published "test validations", and perhaps reiterate the basic principles of the CAM, or some other measurement technology (e.g., axiomatic measure theory). They might remark that a claim such as "my dominance has increased by 6 since April" can have a sense, as when one employs some particular dominance inventory. [6] But this is an illusion. It presupposes that the inventory yields measurements of dominance, while, in fact, the rules for the correct employment of the concept *dominance* do not warrant ascription of the concept to another individual on the basis of any dominance inventory, but, rather, on the basis of criterial behaviours that are linked to the concept in grammar. The claim represents either a misuse of the ordinary language concept *dominance* or an unacknowledged use of a technical homynym that denotes distinct phenomena. To put this another way, if the rules of concept "ψ" warrant its ascription to another on the basis of a dominance inventory, then concept "ψ" is not the ordinary language concept *dominance*, but some other technical concept, whose relation to the web of ordinary language psychological concepts, and the phenomena they denote, is in need of explanation. It would then be best to avoid confusion by giving this novel concept some *other* name.

A more sophisticated rebuttal would begin by noting that the employments of many psychological concepts are modified by terms such as "very", "somewhat", "extremely", etc. For example, we correctly speak of people as being *very dominant*, *extremely intelligent*, *quite creative*, and *somewhat hopeful*. It might then be claimed that this represents a primitive measurement practice, and that psychometrics has been engaged in the business of building this primitive practice into something akin to that which exists for the measurement of length, height, weight, etc. David Krantz, one of the originators of the axiomatic measurement approach, states, for example, that "If behavioral science measurement is modelled after examples drawn from the physical and biological sciences, it seems natural to move from the presupposition of an ordering

---

6 As a reviewer once remarked: "Maraun is incorrect when he asserts that: 'There is no public, *normative* status at all to assertions like "Tomorrow we are going to measure little Tommy's dominance"'. Indeed, one of the authors...participated years ago in a seminar on personality assessment that was designed so that students would learn the proper norms for how to administer and score a given psychological instrument in class one week and then set out the following week to administer the instrument to an individual research participant." But while there are norms that fix the correct administration and scoring of tests (one can make mistakes in this regard), this is not the same thing as there being norms for the measurement of phenomena denoted by psychological concepts.

to the goal of numerical measurement of the variable in question" (1991, p.3). And when, for example, a new version of the Jones Depression Inventory (JDI2) is correlated with the previous version, the JDI1, is this not analogous to what the chemist does in calibrating a new scale for the measurement of the weights of objects?

This line of reasoning is mistaken. Once again, what is required in moving from less to more advanced stages in a measurement practice is that the rules that are constitutive for measuring at the less sophisticated stage are squared with the proposed innovations. Linguistic rules fix the correct employments of ordinary language psychological concepts, and, in particular, the grounds of ascription of these concepts to another in the third-person present tense mode. The grounds of ascription of, say, the term *dominant* (*very dominant*, *passive*, etc.) to another individual rests on behavioural criteria tied to the concept in grammar. This is not equivalent to the existence of a primitive practice of measurement, for the ingredients of measurement, e.g., units of measurement, canonical samples and recipes, rules for the expression of the results of measuring, are absent. And even if this could rightly be said to constitute primitive measurement, there is no evidence, whatsoever, that the discipline of psychometrics has been engaged in the careful reconciliation of the rules of employment of ordinary language psychological concepts with the measurement innovations it proposes. On the contrary, far from attempting the arduous exercise of attempting to reconcile the existing grammars of psychological concepts with proposed measurement innovations, psychometricians, as has been noted, standardly *mischaracterize* the relationship between the conceptual and empirical features of their business. They routinely misportray the grammars of the psychological concepts that denote the phenomena they purport to measure as "primitive folk theories", hunches, empirical assertions or "mere common sense".[7] They mischaracterize the behavioural criteria that constitute grounds of ascription of psychological concepts in the third-person present tense mode as "indicators", and concepts as "unobservable". The very existence of the Central Account is indicative of a failure to grasp the difference between the grammatical considerations that bear on a concept's meaning, e.g., the internal relation between concept meaning and measurement, and the empirical issues ultimately of interest to the scientist. Not surprisingly, then, the discipline of Psychometrics has not managed to present a single case in which the grammar of a psychological concept has been clarified, a measurement innovation suggested, and an attempt made to reconcile the two.

Unlike the relatively simple grammars of the concepts found in the natural sciences, the grammars of ordinary language psychological concepts are "messy", having evolved in organic fashions, often through the "grafting of language onto natural ('animal') behaviour" (Baker & Hacker, 1982). Psychological concepts have diffuse and multifarious employments (many distinct senses). They are not instantiated in the third person mode by necessary and sufficient

---

7 In the above quote, for example, Krantz (1991) mistakes a characteristic mark of the employments of certain psychological concepts, a feature of the grammars of such concepts (the use of modifiers such as "very", "moderately", etc.) for a "presupposition of an ordering" (which, presumably, could be incorrect).

conditions, but rather by behavioural criteria (Baker & Hacker, 1982). Their third person ascriptions can always be defeated through the broadening of background conditions, a property known as the defeasibility of criteria (Baker & Hacker, 1982). The employments of certain concepts in the first person present tense mode are groundless avowals ("I am sad", "I believe that *p*") that do not rest on any evidence. Psychological concepts possess broad flexibility in their grounds of instantiation, a property that Baker and Hacker (1982) call an open-circumstance relativity. They are simply not organized around finite sets of behaviours, that could be listed in an inventory, and which jointly provide necessary and sufficient conditions for their third person ascription. Take, for example, the concept *dominance*. Given the appropriate circumstances, practically any "raw" behaviour could instantiate the concept. Hence, Joe's standing with his back to Sue could, in certain instances, play the role of instantiator for the concept *dominant*. On the other hand, ordering someone to get off the phone[8] would not instantiate the concept *dominant* if, for instance, a medical emergency necessitated an immediate call for an ambulance. Complicated contextual characteristics learned in the learning of language are components of the meanings of psychological concepts. The employments of psychological concepts are governed by a complicated web of conceptual linkages. Psychological concepts did not evolve within associated measurement practices, and they cannot be legitmately augmented by the measurement-like operations envisioned by the psychometrician.

Now, the psychometrician might choose to reply that "the scientist has no obligation to pay heed to the mere ordinary language employments of psychological concepts." This, of course, is absolutely true, but given the reality of the research conducted in the social and behavioural sciences, is also misleading. In the first place, the comprehensibility of the claims of the psychometrician (and the social and behavioural scientist, too) that, e.g., "summed responses to the Beck Depression Inventory are measurements of depression", "the items comprising this behaviour domain are indicators of anxiety", and "factor one was interpreted to be agreeableness", *presuppose* precisely the same ordinary language employments of psychological concepts for which disregard is professed. To study dominant behaviour is to study just those features of behaviour denoted by the ordinary language concept *dominance*. For the psychometrician does not lay down rules that fix the sense of technical homynyms of ordinary language concepts. His employments of psychological concepts betray adherence to the customary, ordinary language, web of conceptual implications. Thus, for example, when McDonald (1996a, p.598) talks of "...estimating the attitude of one freshly drawn examinee..." or expresses his view that "...a well conceptualized domain is at least implicit in all well conducted applications of FA/IRT to the estimation (with S.E.s) of abilities, attitudes, beliefs, etcetera" or when Cronbach and Meehl (1955, p.63) state that "Even low correlations, if consistent, would support the argument that people may be fruitfully described in terms of a generalized tendency to dominate or not dominate", the psychological terms contained in these expressions are being employed in their standard, ordinary language senses. And departures from standard usages, as

8 An item that appeared on the Act Frequency Approach Act Report C (Buss & Craik)

when such predicates are applied to the scores contained in the distributution of a random variate, result in nonsense.

If, truly, the psychometrician wished to break free of ordinary language employements then, of course, he is free to do so. He would merely need to follow the path taken by the natural sciences and invent a technical vocabularly, and provide accompanying stipulative definitions for the terms of which it is comprised. But the psychometrician has no desire to do this because the cachet of his trade is his claim that he is able to measure intelligence, dominance, agreeableness, depression, etc., the very phenomena signified by ordinary language concepts and of interest to the general public and the social and behavioural scientist, alike.

3. *What passes for measurement in the social and behavioral sciences*

The fact that the rules of employment of ordinary language psychological concepts do not warrant their presence in measurement sentences, i.e., the fact that psychological concepts are not embedded in practices of measurement, has not stopped the psychometrician from both desiring such measurement, and playing at having achieved it. However, what the psychometrician speaks of as "measurement" is a hopeful mislabeling of various brands of constructed numerical representation and various products of data reduction. This misrepresentation is unfortunate because the tools of numerical representation and data reduction invented by psychometricians are themselves impressive contributions to the toolbox of the scientist. The CAM is but one in a series of attempts to create a general framework within which measurements of various psychological phenomena can be produced. Several other of these attempts will now be briefly examined.

Operationism. An operational definition for concept "φ" is simply a set of rules for the production of φ-scores. The scores produced by following these rules are then, *by definition*, signified by "φ".[9] This approach to "measurement" came to psychology by way of the physicist Percy Bridgman, and is modelled after the general strategy for providing technical definitions as found in the natural sciences. Operationism was originally viewed as a saviour of a then young psychology: "Psychological measurement, understood in operational terms, is a *fait acompli*. It is physical measurement. It always has been. And the psychologist, now aware that he is using no mysteriously unique scientific instrument (the observer), can, secure in his new self-knowledge, proceed with his measurement, unimpeded by the hampering difficulties of the Cartesian dichotomy between mind and body" (McGregor, 1935). Indeed, operational

9 This, of course, does not establish anything about the properties of these scores, e.g., their "level of measurement".

definitions were invented for a great variety of psychological concepts, although, nowadays, the concept *operational definition* is most often used not in its original sense, but as a synonym for *stipulative definition*.

In his book *The philosophies of science*, Rom Harre (1985) criticized operationism on the grounds that, if a concept's meaning is truly to be taken as fixed through a statement of the details of whatever measurement setup yields scores with respect it, then each particular setup yields a distinct concept. He reasoned that, eventually, the operationist program would yield an unhelpful proliferation of concepts within the social and behavioural sciences. While true, this misses the mark. The crime committed by the social and behavioural scientist in his use of operationism was to misunderstand the true nature of his problem. If he had truly been in the business of creating a new technical vocabulary, then operationism might indeed have served his purposes. But, in fact, he was in the business of attempting to scientifically study phenomena denoted by ordinary language psychological concepts, and, hence, his task should not usually have been concept formation, but concept *clarification*. For he needed to come to terms with the rules of correct employment of the existing, ordinary language psychological concepts that signified the phenomena (intellectual capacities, attitudes, beliefs, etc.) of interest to both he and society.

The practice of operationally defining existing ordinary language psychological concepts has two possible outcomes: i) The creation of new concepts, technical homynyms of ordinary language concepts, these concepts employed thereafter to organize and denote phenomena that is distinct from the phenomena denoted by ordinary language concepts. There is nothing objectionable about this outcome, as long as it is realized that *new* phenomena, not the phenomena denoted by ordinary language psychological concepts, are now the object of study. Care must, obviously, be exercised to avoid confusing the distinct technical (operationally defined) and ordinary senses of a given term; ii) The *appearance* of the creation of technical homynyms of ordinary language concepts through the statement of operational definitions, accompanied by the belief that these homynyms settle what is meant by the ordinary language senses, and followed thereafter by the adherence to the ordinary language senses to denote and organize phenomena of interest.

Regardless of the impression given by the social and behavioural scientist in his operationism, he was, in fact, interested in explaining, not the phenomena denoted by his operational definitions, but the phenomena that were of relevance to society, and those phenomena were precisely the phenomena denoted by the ordinary language psychological concepts of English. Thus, outcome (ii) was the predictable consequence of his brief infatuation with operationism. The charge that must, then, be levied against the operationist is that of conceptual fraud, and the consequence of this midemeanour was an impenetrable murkiness in regard the concepts that were to be understood as signifying the phenomena under study (unexplicated ordinary language sense mixed with technical, operational senses), and,

consequently, an impenetrable murkiness in regard the phenomena that were the objects of his studies.

The construct validation approach. As discussed previously, the construct validation approach to the support of measurement claims borrows heavily from empirical realist thinking, was popularized by a series of papers by Cronbach and Meehl, notably their 1955 Psychological Bulletin entry, was refined and extended by various individuals, perhaps notably Messick (e.g., 1981), and, finally, was enshrined as received account in the *Standards for Educational and Psychological Test and Manuals* (1985) of the American Psychological Association. The core logic of the approach can be unpacked as follows.

(A) Natural reality and psychology

(i) Natural reality is comprised of *observable* behavioural phenomena, and the *unobservable* or *latent* causes of these phenomena. Unobservable causes are concealed from perception (Rozeboom, 1984, p.212). They exist independently of human ability to detect their presence. Their existence "does not depend on any mind (cf. Tuomela, 1973, p.7).

(ii) Unobservable causes of interest to the psychologist include *traits* and *situational factors*. Traits (e.g., *dominance*, *cheerfulness*) are attributes of people. Traits that are quantitative in nature (e.g., *extroversion*) are possessed by people according to degree, while those that are discrete in nature (e.g., *amnesia*; Cronbach & Meehl, 1955, p.60) characterize people as types or kinds. Situational factors are true, unobservable, causal forces, possessed by situational contexts, and exerting their influences on humans.

(B) Observational and theoretical terms

(i) Scientific concepts can, in general, be categorized as either *theoretical terms* or *observational terms* (Tuomela, 1973). Observational terms are sematically unproblematic, as they designate observable phenomena. They are *defined* in data-language terms (explicitly in terms of observables; e.g., via ostensive definition). Theoretical terms (e.g., *electron*, *neutrino*, *anxiety*, *intelligence*) semantically designate unobservable entities, and, notably, unobservable causes of observable phenomena. As a result, these terms cannot be defined on the basis of observational terms. They are introduced by scientific theories as a means of explaining and predicting established empirical laws, and as a means of providing economy, simplicity, and

coherence of description of empirical phenomena. The presumptive hypotheses of truly explanatory theories often include theoretical terms.

(ii) The term *construct* as employed in the theory of construct validity is a synonym for *theoretical term*[10].

(C) Psychological tests

(i) A test, T, is a collection of *stimulus materials* and *response options*, plus a *scoring rule*. The scoring rule converts an individuals responses to the stimulus materials, as recorded by the response options, into a test score. When a population, P, of individuals is given a test, the scores thus produced may be represented by a random variate ω.

(ii) The responding behaviour of the individuals in P to T, called *test behaviour*, is an example of observable behavioural phenomena. In short, ω is an observable variate. From the perspective of the test analyst, all other observables are classified as *non-test* behaviours.

(iii) Test and non-test behaviours are caused by a range of unobservable causal traits, situational factors, and method factors (causal sources associated with a given test, rather than with the individuals tested). Be they causal or not, the web of relationships between observable test and non-test behaviours, and unobservable causal sources are (at least potentially) characterizable by laws.

(D) Constructs and their defining theories

(i) Let there be a theory, TH, that explains the relationships posited to exist between a set of test and non-test observables, $\{\omega,nt_{(i)}\}$, i=1..p, and a set of causal unobservables, $\{\phi_{T(j)},\phi_{S(k)},\phi_{M(l)}\}$, j=1..q, k=1..r, and l=1..s, in which $\phi_{T(.)}$ represents a trait, $\phi_{S(.)}$ a situational factor, and $\phi_{M(.)}$ a method factor. The theory is: 1) Expressed in terms of theoretical terms, $\{C_{\phi(Tj)},C_{\phi(Sk)},C_{\phi(Ml)}\}$, which refer to the unobservables and observational terms, $\{'o'_{(\omega)},'o'_{(nt(i))}\}$, which refer to the observables; 2) Will typically include presumptive hypotheses, and propositions taken provisionally to be laws, regarding $\{\phi_{T(j)},\phi_{S(k)},\phi_{M(l)}\}$and their relations to $\{\omega,nt_{(i)}\}$; 3) Is the conjoinment of all sentences taken to be true of $\{\phi_{T(j)},\phi_{S(k)},\phi_{M(l)}\}$, and, in

10 See, e.g., Hempel, 1958; Sellars, 1963; Tuomela, 1973; Rozeboom, 1984.

particular, those stating how these unobservables are related to observational entities $\{\omega,nt_{(i)}\}$ (Rozeboom, 1984).

(ii) The term *nomological network* as employed in the theory of construct validation is a synonym for TH.

(iii) The meanings of $\{C_{\phi(Tj)},C_{\phi(Sk)},C_{\phi(Ml)}\}$ are given by the properties of the unobservable entities they designate, if, in fact, such unobservables actually exist. However, because these entities are unobservable, their existences and properties are never directly knowable. As a result, the theoretical terms $\{C_{\phi(Tj)},C_{\phi(Sk)},C_{\phi(Ml)}\}$ are *implicitly* defined by TH. Put another way, they are given *partial interpretation* through correspondence rules or interpretive systems stated in terms of the observational terms $\{'o'_{(\omega)},'o'_{(nt(i))}\}$ (Tuomela, 1973). The greater the number of *nomological strands*, i.e., laws and presumptive hypotheses concerning $\{\phi_{T(j)},\phi_{S(k)},\phi_{M(l)}\}$ and contained within TH, the clearer do the meanings of $\{C_{\phi(Tj)},C_{\phi(Sk)},C_{\phi(Ml)}\}$become.

(iv) While the theoretical terms $\{C_{\phi(Tj)},C_{\phi(Sk)},C_{\phi(Ml)}\}$ are given partial interpretation by TH, if TH is, in fact, true, then $\{C_{\phi(Tj)},C_{\phi(Sk)},C_{\phi(Ml)}\}$do, in fact, semantically designate unobservable causes $\{\phi_{T(j)},\phi_{S(k)},\phi_{M(l)}\}$. That is, while observables give $\{C_{\phi(Tj)},C_{\phi(Sk)},C_{\phi(Ml)}\}$ partial meaning, the $\{C_{\phi(Tj)},C_{\phi(Sk)},C_{\phi(Ml)}\}$ are not *about* these observables, but rather the hypothesized causal sources to which they make reference. On the other hand, there may not exist $\{\phi_{T(j)},\phi_{S(k)},\phi_{M(l)}\}$, in which case TH is false: "...phlogiston theory expired not from the discovery that phlogiston's properties were other than as originally conjectured but from the conclusion that phlogiston did not exist" (Rozeboom, 1984, p.212). Hence, the theoretical term *phlogiston*, while not lacking meaning, failed at reference, i.e., turned out not to designate an unobservable substance with the properties described by phlogiston theory.

(E) The meaning of *construct valid test*

(i) The construct validity of a test is of concern when the test is designed to measure a construct which designates a quantitative or discrete trait possessed by the members of a population (Cronbach & Meehl, 1955, p.59).

(ii) A test T is a construct valid measure of construct $C_{\phi(T^*)}$ only if responding to T is causally determined by $\phi_{T(*)}$. That is, only if the variation in ω in population P is solely attributable to variation in $\phi_{T(*)}$ in P.

(iii) In all applied settings, variation in ω will be attributable not only to $\phi_{T(*)}$, but also to a range of other traits, situational and method factors. Hence, the proportion of variance in ω attributable to $\phi_{T(*)}$, $V(\omega,\phi_{T(*)})$, will be less than unity.

(iv) The focal presumptive hypothesis that "T is a construct valid measure of construct $C_{\phi(T^*)}$" (but, in practice, that the value of $V(\omega,\phi_{T(*)})$ is "large enough") may be incorrect. T may, instead, measure primarily constructs other than $C_{\phi(T^*)}$. That is, ω may be causally determined primarily by unobservable causes other than $\phi_{T(*)}$.

(F) The program of construct validation

(i) Because $\phi_{T(*)}$ is unobservable, $V(\omega,\phi_{T(*)})$ is not directly estimable. The test analyst must, therefore, make an inference as to the value of $V(\omega,\phi_{T(*)})$. In particular, the focal presumptive hypothesis that "T is a construct valid measure of construct $C_{\phi(T^*)}$" must be transformed into derived hypotheses stated in terms of the observables $\{\omega,nt_{(i)}\}$. These derived hypotheses are then tested directly.

(ii) Given an appropriate transformation, confirmation of a derived hypothesis is a necessary, but not sufficient, condition for the truth of the focal presumptive hypothesis. Confirmation implies that the focal presumptive hypothis remains *plausible*. Disconfirmation means that some aspect of TH is false, but which aspect is false is indeterminate with regards a single experiment (a consequence of the Quine-Duhem thesis). In particular, TH may be incorrect, or T may not be a valid measure of $C_{\phi(T^*)}$. The test analyst cannot know with certainty which is the case.

(G) The progressive, ongoing nature of investigation

(i) The attempt to assess whether T is a valid measure of $C_{\phi(T^*)}$ is a progressive, ongoing enterprise. It never properly comes to an end. The investigator must learn more about $\{\phi_{(T^*)},\phi_{T(j)},\phi_{S(k)},\phi_{M(l)}\}$ via a program of research, and the yield of this program, realized through an ongoing recursion of explanatory inductions, will: 1) Gradually add strands to the nomological network (i.e., make TH progressively more highly articulated); 2) Increasingly pin down the meanings of $\{C_{\phi(Tj)},C_{\phi(Sk)},C_{\phi(Ml)}\}$; 3) Provide the basis for successive modifications to TH; and 4) allow for progressively more precise inferences regarding the value of $V(\omega,\phi_{T(*)})$.

(ii) The use of T in research generates empirical propositions regarding ω and its relationships to other observables and unobservables. To the extent that T is a construct valid measure of $C_{\phi(T^*)}$ these are also statements regarding the relationships between $T_{(*)}$ and other observables and unobservables.

(iii) Exactly how the research program described in (i) is carried out will depend, at any point in time, on current opinion regarding (1), (2), (3) and (4) in G(i). Hence, (1), (2), (3) and (4) also play a role in the accumulation of knowledge regarding $\{\phi_{(T^*)},\phi_{T(j)},\phi_{S(k)},\phi_{M(l)}\}$. It may be concluded then that there is, in the construct validation program, a constant interplay between knowledge accumulation, theory, meaning, and the gaining of understanding regarding what T measures.

(H) The irremediable uncertainty attendant to construct validity claims

(i) Knowledge of both observables and unobservables: 1) is only partial; 2) is gained through successive approximations; 3) is attained through the interplay of experience and reason; 4) is hypothetical and corrigible rather than apodictic and final; 5) is distorted, symbolic, and, hence, indirect (Tuomela, 1973, p.7).

(ii) Because $\{\phi_{(T^*)},\phi_{T(j)},\phi_{S(k)},\phi_{M(l)}\}$ are unobservable, the healthy scientific scepticism expressed in (i) is, in a construct validation program, amplified. Both the existence, and, given their existence, the properties of $\{\phi_{(T^*)},\phi_{T(j)},\phi_{S(k)},\phi_{M(l)}\}$, including their causal impacts on $\{\omega,nt_{(i)}\}$, are not even in principle verifiable. If they exist, unobservable causes$\{\phi_{(T^*)},\phi_{T(j)},\phi_{S(k)},\phi_{M(l)}\}$ are detectable only through their "data consequences" (Rozeboom, 1984, p.212), i.e., their deduced impacts on observable entities $\{\omega,nt_{(i)}\}$. This implies the existence of several distinct sources of uncertainty attendant to a construct validation program. Obviously, TH will always be incomplete, and, hence, so too the implicitly defined meanings of $\{C_{\phi(Tj)},C_{\phi(Sk)},C_{\phi(Ml)}\}$. This will, in turn, introduce uncertainty as to the soundness of the derived hypotheses as transformations of the focal presumptive hypotheses of TH. Hence, since decisions regarding the place of test T in TH, and, notably, decisions regarding its degree of construct validity, are based on such transformations, such decisions will always be open to question.

But this characterization of measurement is badly mistaken. It is, in the first place, a misreading of the logic of measurement in the natural sciences, and, in particular, a misreading of the place of theory and knowledge in regard the adjudication of measurement claims. The correct claims that empirical reality is a backdrop against which measurement practices are founded, and that scientific theory and knowledge have been essential to the striking innovations humans have made in regard their practices of measuring, are mutated into the faulty claim that theory and empirical fact are jointly the *adjudicators* of measurement claims. This is akin to claiming that the theory of weather forecasting partly determines what it is to make a correct claim about the current air temperature. In particular, the key conceptual feature of measurement, that the rules of correct employment of certain concepts warrant their ascriptions

to numbers produced in certain particular ways, is overlooked in favour of the wholly irrelevant search for the *causes* of responding to psychological tests.

To claim that the scores produced by application of rule r to members of population P_T are measurements of extroversion is to claim that the rules that fix the correct employment of the concept *extroversion* warrant the application of the concept to scores produced in accord with rule r (i.e., that the scores are signified by concept *extroversion*). Hence, once again, a measurement case is adjudicated by comparing behaviour (in this case, the production of scores in accordance with rule r) to rules (in this case, the rules that fix the correct employment of concept *extroversion*). The sought after conceptual signification, that *these* scores are measurements of extroversion, will hold only if the concept *extroversion* is, in fact, embedded in a normative practice of measurement, and rule r is, in fact, a rule that fixes what it is to take measurements of extroversion. Empirical cases are provisional, potentially disconfirmable, and, indeed, adjudicated on the balance of evidence. But a measurement case is not an empirical case. For, given that σ is, in fact, measurable, to correctly follow the rules for the measurement of σ imparts to the resulting scores certainty that they are, indeed, measurements of σ.

Construct validation theory's misreading of the place of theory and knowledge in regard the adjudication of measurement claims is hardly surprising, given that a reading of the extant literature on the subject reveals not only rampant equivocation over what is intended by the term construct (as noted earlier in the book), but also profound confusion in regard the components of scientific investigation, and, in particular, the relationship between the meaning of a concept that denotes and its class of referents. The lack of sensitivity to these issues is reflected in Cronbach and Meehl's assertions that "Scientifically speaking, to 'make clear what something *is*' means to set forth laws in which it occurs" (1955, p.20) and that "We currently don't know what *anxiety* means. We are engaged in scientific work to answer this question" (1955), which blur the distinction between a concept's meaning (fixed by its rules of correct employment) and the empirical natures of the phenomena it denotes (if it does, in fact denote). And the "background notions" (1955, p.67) considered by Cronbach and Meehl, are very often not primitive *empirical* hunches at all, but, rather, the shadow of the ever present reticulation of linguistic rules that fix the correct employments of psychological concepts. Having wholly misplaced the language that allows for the conceptualization of psychological phenomena, it then resurfaces as some strange new beast, which Cronbach and Meehl proceed to describe as constituting primitive hunches.

Let the correct employment of concept "σ" be governed by rules $r_{"\sigma"}$. Also assume that concept "σ" denotes a class of constituents of natural reality, say, σ-things. What this means is that $r_{"\sigma"}$ warrant application of "σ" to certain constituents of natural reality, σ-things, but not to others. Now, on the one hand, to make clear what a σ-thing *is*, i.e., to clarify to which constituents of natural reality the concept "σ" is rightly applied, is to clarify the rules of correct employment of concept "σ" (the meaning of concept "σ"). For only by grasping the rules of correct employment of concept "σ" can one grasp whether the concept denotes, and, if it does denote, which features of natural reality it denotes. On the other hand, to reveal the properties of

σ-things is to reveal the properties of the referents of concept "σ" (given that concept "σ" does, in fact, denote), e.g., the causes of these referents. To reveal properties of these referents is to make empirical discoveries and such discoveries will, indeed, be the product of empirical investigation (and, possibly, the setting forth of the (empirical) laws into which σ-things enter). *But the latter issue, the collection of facts about σ-things and the formulation of laws involving σ-things, has no bearing on the former issue, i.e., which constituents of natural reality concept "σ" can rightly be applied to*. On the contrary, to collect facts about, and formulate laws involving, σ-things *presupposes* that the scientist can specify what is, and is not, a σ-thing, the possession of this capacity is equivalent to grasping the rules of correct employment of concept "σ" (of knowing to which constituents of reality "σ" is legitimately applied).

The blurring of conceptual and empirical issues within the construct validation view of measurement means that incoherence is never far away. For, example, in the APA *Standards for educational and psychological tests* (1966), it is claimed that "Construct validity is evaluated by investigating what psychological qualities a test measures; i.e., by determining the degree to which certain explanatory concepts or constructs account for performance on a test" (pp.12-13). In the first place, concepts do not "account for performance on a test" or anything else, for they are not entities which possess causal powers, although certain concepts denote constituents of natural reality that possess such powers. Nor do constructs account for test performance, at least if the term *construct* is used as a synonym for *theoretical term*. In the second place, the quote betrays a presumption that a given test measures *something*, but that researchers can't be sure what this something is. It is as if a given test is a bucket that is dipped into the well of (unobservable) psychological "qualities", and that, even though the researcher can't be sure *what* he has come up with, he knows that he has come up with something. This is the Central Account speaking, and it is nonsense. Psychological characteristics are not, in distinction to, say, viruses, entities floating around in nature. Psychological characteristics are simply those characteristics named by the psychological concepts that are contained in language, and these concepts are ascribed to others on the basis of behaviour. Finally, it is deeply confused to suggest that what is measured by a measurement instrument *accounts* for the measurements yielded by the instrument. This is akin to claiming that the height of an individual determines or accounts for measurements of the height of this individual. What determines the height of an individual are a range of causal and background factors, many biological in nature (e.g., heredity, diet), others currently unknown. To investigate the factors that determine the heights of individuals *presupposes* an independent means of measuring their heights, for, without such a means of measuring the heights of individuals, there would be nothing to explain. The practice of measuring the heights of individuals is a rule guided practice, and is, thus, autonomous of facts *about* the actual heights of individuals, including the causal determiners of these heights.

Nevertheless, modern construct validators continue to believe that measurement cases are intimately related to causal cases. The following account of measurement was taken from a structural equation modeling discussion group:

5. We wish to develop an indicator of how much "?" or "Self Esteem" is being "generated" or "is present" as a causal process for any individual at any time.

9. Then, we decide to create a questionnaire or behavioural rating measure, using a numbered scale, of the amount of "?" or "Self-Esteem". We call this scale "a measure of Self-Esteem".

10. Unless we choose a readily reproducible maximum and minimum value for the end-points of our scale - we generate a scale of measurement that is not linked to any "standard" limits.

11. We then impose (or create via additive conjoint theory) equal-interval units, called "?", with the reproducible limits now postulated as the minumum and maximum possible score of our particular scale of questionnaire/rating items. These "units" indicate the effect of our cause "?" or "Self-Esteem". At this point, the functional relation between the "amount of latent cause" and the unit of the scale of "?/Self-Esteem behaviours" is unknown - because we do not know what this "latent" is – or how it causes our phenomena to occur - merely that it is hypothesised to do so because of the systematic changes that we can observe in our phenomena.

12. We state - The cause "?" or "Self-Esteem" is observed via our measurement scale of "?/Self Esteem" which is applied to individuals who change in some reliable way with the "influence" of "?" or "Self Esteem".

The idea is that *self-esteem* is a name given to the, currently unknown cause, c, of a set of behaviours {a,b,c,…,g} that appear to co-occur. The aim is to construct a rule, r, that accurately scales people with regard the amount of c they possess. The {a,b,c,…,g} are seen as "indicators" of c, and, hence, are relevant to the construction of r. If such a rule can be constructed (preferably possessing various attractive properties), then it is a measure of self-esteem.

This account is badly mistaken. Even if the cause, c, of behaviours {a,b,c,…,g} were to be discovered through *real* scientific investigation (a search for a "latent cause" would be akin to launching an expedition to Narnia), and a rule, r, constructed that provided information about the statuses of people with respect c, r would not be a measure of self-esteem, but rather an indicator of c. To claim that r yields *measurements* of self-esteem is to claim that the concept *self-esteem* can legitimately be said to signify the scores r yields, and such signification is only possible if the concept is embedded in a normative practice of measuring. But the concept *self-esteem* is not embedded in such a practice. If it were, language users would learn how to measure self-esteem as part of learning how to correctly employ the concept *self-esteem*. Children would learn the standard uses of expressions akin to "her self-esteem is 15 gergls on the r scale." The fact that the concept *self-esteem* is not embedded in a normative practice of measuring is why the linguistic rules that fix its correct employment do not warrant expressions akin to these. Such expressions are merely nonsense.

The linguistic rules that fix the correct employment of the concept *self-esteem* do not warrant ascription of the concept to another in the third-person, present-tense mode, on the basis of his status with respect the cause, known or otherwise, of the behavioural criteria of the concept. The cause, known or otherwise, of these behaviours is of no relevance to the normative employment of the concept. The teaching and learning of how to correctly employ the concept *self-esteem* goes on without reference to any putative causes. The ascription of (high-, low-, etc.) self-esteem to another is, instead, justified by the individual's having behaved in a manner criterial for the ascription of the concept, and behavioural criteria are internally (grammatically) related to the concept. One is warranted by the linguistic rules that fix the correct employment of the concept *self-esteem* to ascribe it to another given that the other has behaved in certain ways, in certain contexts. It should be noted that there would not even be the possibility of discovering the cause c of self-esteem behaviours unless one could antecedently *identify* self-esteem behaviours. One would have to know *which* behaviours were criteria of, say, high self-esteem, and how to distinguish these behaviours from leadership behaviours, dominant behaviours, etc. But since self-esteem behaviours are just those behaviours that are criteria for the concept *self-esteem*, and, hence, are internally related to the concept, to be able to antecedently identify self-esteem behaviours is to grasp certain features of the concept's normative employment. The concept of *self-esteem*, its behavioural criteria, and their possible cause(s) must be carefully distinguished, and their logical relationships kept in clear view. The relationship between concept and criteria is internal or conceptual, between cause (if such a thing exists) and behavioural criteria, contingent and empirical, and between concept and cause, nonexistent.

Consider some of the claims made by Jost and Gustafson (1998) in their account of the construct validation program:

"A goal of empirical investigation is to determine how each theoretical term interacts with others (their roles in nomological nets, in the ideal case), with external conditions, with whatever other variables can be studied in relation to what we initially and more-or-less pre-theoretically take as our target of investigation" (Jost & Gustafson, 1998, p.474)

Here, Jost and Gustafson conflate *theoretical term* with *referent of theoretical term*. Theoretical terms, not being constituents of natural reality, can neither be said to "interact with others", nor "with external conditions", and certainly do not occupy positions in nomological nets (networks representing nomological laws describing constituents of natural reality). Hence, empirical science, whose task it is to address questions about constituents of natural reality, cannot hope to resolve questions about theoretical terms, but only the referents of theoretical terms (these, indeed, constituents of natural reality).

"If an instrument is being used to predict outcomes that are too close in conceptual space (e.g., if there is a tautological connection between the independent and dependent variables)..." (Jost & Gustafson, 1998, p.471)

The constituents of a predictive case are natural phenomena, or, at least, variates representing these phenomena. Neither the constituents of natural reality, nor variates, can be said to be elements of "conceptual space". Not being elements of conceptual space, a pair of predictive outcomes can neither be "too close", nor "too far", from each other within "conceptual space". Concepts are, figuratively speaking, elements of conceptual space, but, then, not being natural phenomena, they cannot be coherently said to be constituents of a predictive case.

"In these contexts there is no similar idiocy in comparing the hopefulness (or hopelessness) of persons distinguished by gender or age or socio-economic status or whatever else. Here we have a trait which, as with others, comes in more or less, in degrees or extents. Then we can devise a measure- a device, an instrument-the statistical properties of which can be explored, and which can be compared to a range of other indicators of hopefulness. In other words, arguments for its reliability and validity can be put forth on the basis of empirical investigations...Once a given psychological instrument is found to be reliable (i.e., stable across time), it might then be found useful in assessing hypotheses, say, about the relation of hopefulness measurements and voting behavior. There are no conceptual problems here..." (Jost & Gustafson, 1998, p.476)

Now, here, one finds a grain of truth. First, the grammars of many psychological concepts do warrant the making of comparisons of individuals, although not via the employments of psychological inventories, for such inventories are not a part of the normative employments of any ordinary language psychological concept. Second, it is true that one could invent a rule, say r_h, by which scores are assigned to individuals belonging to a population *P* under study, these scores idealized as a distribution, in population *P*, of a random variate **X**, and the statistical properties of **X**, explored. One could, for example, compare properties of the distribution of **X**, conditional on particular values of a second variate ("persons distinguished by gender or age or socio-economic status or whatever else"). Or, by further idealizing each individual's score on **X** as $X_i=\tau_{Xi}+\varepsilon_{Xi}$, and asserting that the τ_i and ε_i have a joint distribution in *P* possessing the properties $\rho(\boldsymbol{\tau}_X,\boldsymbol{\varepsilon}_X)=0$ and $E(\boldsymbol{\varepsilon}_X)=0$, it could be deduced that:

i. $E(\mathbf{X})=E(\boldsymbol{\tau}_X)$

ii. $V(\mathbf{X})=V(\boldsymbol{\tau}_{\mathbf{X}})+V(\boldsymbol{\varepsilon}_{\mathbf{X}})$

The reliability of the scores produced by rule r_h could, then, be defined as

iii. $\rho_{hh'} = \dfrac{V(\boldsymbol{\tau}_{\mathbf{X}})}{V(\boldsymbol{\tau}_{\mathbf{X}}) + V(\boldsymbol{\varepsilon}_{\mathbf{X}})} = \dfrac{V(\boldsymbol{\tau}_{\mathbf{X}})}{V(\mathbf{X})}$.

Without further restrictions being imposed, $\rho_{hh'}$ would not be calculable, because the τ_{Xi} and ε_{Xi} scores can not be produced, and equation (ii) is indeterminate. The traditional next step is to consider a second rule, r_{h2}, which, when applied to members of *P*, yields scores that comprise the distribution of a second random variate, **Y**. Once again, each individual's score on **Y** is idealized as $Y_i=\tau_{Yi}+\varepsilon_{Yi}$, and it is asserted that the τ_{Xi}, τ_{Yi}, ε_{Xi}, and ε_{Yi} have a joint distribution in *P* with the properties $\rho(\boldsymbol{\tau}_{\mathbf{X}},\boldsymbol{\varepsilon}_{\mathbf{X}})=\rho(\boldsymbol{\tau}_{\mathbf{X}},\boldsymbol{\varepsilon}_{\mathbf{Y}})=\rho(\boldsymbol{\tau}_{\mathbf{Y}},\boldsymbol{\varepsilon}_{\mathbf{X}})=\rho(\boldsymbol{\tau}_{\mathbf{Y}},\boldsymbol{\varepsilon}_{\mathbf{Y}})=0$, and $E(\boldsymbol{\varepsilon}_{\mathbf{X}})=E(\boldsymbol{\varepsilon}_{\mathbf{Y}})=0$. Random variates **X** and **Y** are, by definition, "parallel variates" if and only if:

iv. $\tau_{Yi}=\tau_{Xi} \quad \forall i \in P$

v. $V(\boldsymbol{\varepsilon}_{\mathbf{X}})=V(\boldsymbol{\varepsilon}_{\mathbf{Y}})$.

If **X** and **Y** are parallel variates, it follows that $C(\mathbf{X},\mathbf{Y})=C(\boldsymbol{\tau}_{\mathbf{X}}+\boldsymbol{\varepsilon}_{\mathbf{X}},\boldsymbol{\tau}_{\mathbf{Y}}+\boldsymbol{\varepsilon}_{\mathbf{Y}})=C(\boldsymbol{\tau}_{\mathbf{X}},\boldsymbol{\tau}_{\mathbf{Y}})=V(\boldsymbol{\tau}_{\mathbf{X}})=V(\boldsymbol{\tau}_{\mathbf{Y}})$, and $V(\mathbf{X})=V(\mathbf{Y})$, so that $\rho(\mathbf{X},\mathbf{Y})=\dfrac{C(\mathbf{X},\mathbf{Y})}{\sqrt{V(\mathbf{X})V(\mathbf{Y})}} = \dfrac{V(\boldsymbol{\tau}_{\mathbf{X}})}{V(\mathbf{X})}\dfrac{V(\boldsymbol{\tau}_{\mathbf{X}})}{V(\boldsymbol{\tau}_{\mathbf{X}}) + V(\boldsymbol{\varepsilon}_{\mathbf{X}})} = \rho_{hh'}$. Hence, if a researcher could somehow invent two parallel variates, he could estimate the reliability of each variate by estimating their Pearson Product Moment correlation in *P*. The attempt to produce variates that are, at least approximately, parallel, yielded a plethora of strategies including alternative forms (and, eventually, the various internal consistency methods in which individual items are taken as alternative forms), split-half, and test-retest.

As long as a well defined sense of the term *reliability* is given, the reliability of a variate **X** can be studied, and, because reliability is a property of univariate, bivariate, or multivariate distributions, such a study is empirical in nature. The program of empirical investigation to which Jost and Gustafson allude might well lead to the development of a rule which yields scores on a variate **X** that possess some very nice properties, and, importantly, a variate that can be put to effective use in research. *But notice that this program can be carried out without reference to*

*the meaning of the scores contained in the distribution of **X***. Once a rule r_h of score production is created, there is no end to the statistical products that can be generated. The kinds of analyses Jost and Gustafson envision could literally be carried out forever. *However, a program devoted to the investigation of statistical facts about the distribution of **X**, i.e., the way X-scores (scores produced by rule r_h) distribute in population P, has no bearing on the issue as to whether scores produced by rule r_h happen to be signified by some concept, e.g., hopefulness*. The only consideration relevant to a question of signification (whether one can legitimately apply a particular concept to the scores produced in accord with a given rule), *a measurement question*, is the consideration of *how* the scores were produced, and, in particular, whether they were produced in accord with rules for the production of measurements of some property φ (presuming that concept "φ" is embedded in a normative practice of measurement).

"Jost and Thomson (1997) have argued, for example, that the existing measure of SDO confounds two distinct response tendencies, one of which captures a desire for ingroup superiority, and the other of which captures a desire to preserve existing hierarchical relationships. This conceptual critique led to a set of empirical predictions, for instance that a "two-factor" would provide a better comparative fit of the data than a "one-factor" solution and that the two factors would be more highly inter-correlated among European Americans than among African Americans...Empirical support for these hypotheses then strengthened the conceptual and theoretical argument that prior definitions of SDO conflated two different variables. Conceptual and empirical claims go together here they are not at odds" (Jost & Gustafson, 1998, p.473)

This is the archetypal mess induced by the confusions inherent to construct validitation theory. Let us unpack it:

i. There is in play in psychology a technical concept, *social dominance orientation* (Pratto, Sidanius, Stallworth, and Malle, 1994), whose correct employment is fixed by (more or less well specified) rules r_{SDO}.

ii. This being a technical concept, "the existing measure of SDO", say I_{SDO}, produces SDO scores. Since ordinary language psychological concepts are not embedded in normative practices of measurement, these scores are not signified by an ordinary language psychological concept. Let the distribution of random variate **X** contain the scores produced by application of I_{SDO} to the members of some population *P*.

iii. Jost and Thomson (1997) hypothesized that the response of each member of population *P* to the I_{SDO} is determined (at least probabilistically) by two behavioural tendencies, one of which is the desire for ingroup superiority, and the other, the desire to preserve existing hierarchical relationships. This is an hypothesis about the causes of responding to the I_{SDO}. This hypothesis certainly did not result from a "conceptual critique", because it did not result from a critique of concept employment. It resulted, instead, from a set of reasoned conjectures about the causal determinants of responding to a psychological inventory. The belief of Jost and Gustafson that such a set of reasoned conjectures constitutes a "conceptual critique" is an example of the misportrayal of a conceptual issue, an issue pertaining to the rules of correct employment of a concept, as having to do with opinions, hunches, conjectures, or hypotheses about constituents of natural reality.

iv. Jost and Thomson paraphrased their hypothesis in terms of the CAC. That is, they interpreted the hypothesis that "there are two particular causal determinants of responding to the I_{SDO}, "desire for ingroup superiority" and "desire to preserve existing hierarchical relationships"", as implying that the I_{SDO} should be two-dimensional in the linear factor analytic sense of dimensionality. So here one has a prime example of the misportrayal of linear factor analysis as a tool of detection. Whereas what was needed were definitions of the concepts *desire for ingroup superiority* and *desire to preserve existing hierarchical relationships*, so that a research program could have been launched to examine the impact of the *referents* of these concepts on responding to the I_{SDO}, the course of action taken was the lazy, wholly ambiguous, course of action in which (latent) random variates are introduced. The linear factor analytic hypothesis that a set of items is two-dimensional is borne out in the population if $\Sigma=\Lambda_2\Lambda_2+\Psi$, in which Ψ is a diagonal, positive definite, matrix, and is equivalent to the existence of at least one pair of random variates $[\boldsymbol{\theta}_1,\boldsymbol{\theta}_2]$ with the properties that the conditional distribution of $\underline{\mathbf{X}}$ given $[\boldsymbol{\theta}_1,\boldsymbol{\theta}_2]=[\theta_1,\theta_2]$ has mean vector $\Lambda_2\begin{bmatrix}\theta_1\\ \theta_2\end{bmatrix}$ and covariance matrix Ψ. In other words, Jost and Thomson's hypothesis has implications for the joint distribution of the items of the I_{SDO} in the population *P*, and, in particular, the covariance structure of these items.

v. Certainly, even if the CAC *were* a reasonable paraphrase of the hypothesis of two causal determinants, and, of course, it is not, confirmation of this hypothesis could not possibly imply that "prior definitions of SDO conflated two different variables", for this sentence is incoherent. In the first place, definitions, being as they are necessary and sufficient conditions for a concept's application, have nothing to do with variables, which are simply functions. Secondly, it makes no sense whatsoever to speak of two variables as being conflated. Concepts can be conflated,

and this occurs when the rules of correct employment of one concept are wrongly believed to fix the correct employment of another. But variates are simply functions, and two functions cannot be conflated. At best, what this program of research might have hoped to establish is that responding to the items of the I_{SDO} has two causal determinants. But certainly, neither concepts, nor their definitions, could ever be rightly said to conflate causal sources! Concepts *can* denote constituents of natural reality (responding to the items of the I_{SDO}) that happen to have multiple causal sources. It is, of course, entirely possible for researchers to misunderstand, or be unclear about, the causal determinants at work in a given context.

vi. By the end of their discussion, Jost and Gustafson lie sprawled at the bottom of the slippery slope they began to descend when they failed to grasp the difference between the linguistic rules that fix the correct employment of a concept (its meaning) and the "platitudes of common sense" ("We then refine our conception and note that some or all of the platitudes of common sense (Maraun's 'common-or-garden concept' domain)", Jost & Gustafson, 1998, p.475). Their next step was to confuse features of the correct employment of the concept *social dominance orientation* (e.g., its definition) with conceptions (hunches, hypotheses) about the phenomenon of social dominance orientation (that which the concept denotes), and before they were aware of it they were lying at the bottom of the slope mouthing such nonsense as "prior definitions of SDO conflate two different variables". It is absolutely true that "conceptual and empirical investigations go together in science", but the relationship between these two constituents of empirical investigation is anything but as described in construct validation theory. Conceptual clarifications are grammatical clarifications, and must be carried out prior to (fruitful) empirical investigation because they settle *which* constituents of natural reality are to be studied in a study of, e.g., θ-things. One can only know what it is to study θ-things if one grasps the rules of employment of concept "θ", and, in particular, the features of natural reality it denotes, for its referents, if, in fact, it does refer, are precisely those constituents of natural reality that are θ-things.

Case Study: Lubinski, D. (2000). Intelligence: success and fitness. In *The Nature of Intelligence*, 6-36, Chichester: Wiley.

In his 2000 paper, *Intelligence: success and fitness*, Professor David Lubinski discusses research on general intelligence (g), and exemplifies the construct validity approach with an analysis of horsepower. The paper is a shining example of all that is wrong with construct validity, and, in particle, of how it sends the social scientist down the path to pseudoscience. The fatal confusion at the root of construct validity is over the distinction between conceptual and empirical issues. While conceptual/empirical confusions do occur in science, no true science has

ever *enshrined* this confusion as a fundamental postulate of its operations. The adherence of the social sciences to construct validity is the first and only attempt to do so. Someday, historians will look back on the social sciences and view construct validity as a dark ages for social science.

In explaining the construct validity program, Professor Lubinski has this to say about horsepower:

"Construct validity seeks to validate measures of a postulated attribute. "Horsepower" is a postulated attribute, you can't 'see' horsepower, but you can construct indicators that co-vary with meaningful criteria that reflect our concept of horsepower and make it a conceptually powerful and useful concept. Just as horsepower is an abstract property of complex combustion systems, g is an abstract property of complex biological systems."

To begin, as pointed out earlier in this section, construct validators vacillate unpredictably between two distinct senses of the term *construct*. At times, they talk as if the term stands for some feature of natural reality (as when they speak of constructs as having indicators), and at other times as if *construct* is to be taken as synonomous with *theoretical term* (a technical term that denotes an hypothesized unobservable constituent of natural reality). Not a single natural science has felt the need to invoke the notion of *construct*, this despite the fact that natural scientists frequently must deal with cases in which the phenomena of interest to them are perceptually unobservable. Chemistry, biology, physics, and every other science, get by perfectly well by speaking of terms/concepts, on the one hand, and the referents of terms/concepts, on the other. The distinction between term and referent (the entities the term denotes if, in fact, it does denote) is the fundamental conceptual/empirical distinction of science, and construct validity's failure to respect this distinction is the first of many tragic steps it takes.

The correct employments of a given term (*neutrino*, *acidity*, *radioactivity*, *homosapien*) are fixed by linguistic rules. That they are so is clear from the fact that one can be *correct* or *incorrect* in one's employments of these terms (there can only be correct and incorrect behaviour in domains of activity in which humans have layed down rules that fix what it *means* to be correct and incorrect). If, for example, one ascribes the concept *bachelor* to a married man, then one has misemployed the concept. It is a rule of language (a logical feature of the employment of the term) that one cannot rightly ascribe the term *bachelor* to a man who has a wife. The correct employments of the vast majority of the technical terms of science (e.g., *neutrino*) are fixed by linguistic rules of the necessary and sufficient condition variety. These rules, expressed

as definitions (in philosophy, "definition" usually means a rule of concept employment of the necessary and sufficient condition type), are ubiquitous in texts on science. Thus, the rule for the correct employment of the concept *alpha particle* is of the form: the concept *alpha particle* is ascribed to subatomic particle θ if and only if θ consists of two protons bound to two neutrons.

Fundamentally, the doing of science can be divided into conceptual and empirical tasks.

Conceptual task:

Imagine that our aim is to study alpha particles, their properties, their behaviour under particular conditions, etc. The first task is then to settle the issue of *what* it is that we are going to be studying? That is, under what grounds can it be correctly claimed that a fact about constituent of natural reality θ is a fact about an alpha particle?

Answer.

i. Alpha particles are precisely and only those constituents of natural reality to which the concept *alpha particle* is correctly ascribed;

ii. The grounds for correctly ascribing the concept *alpha particle* to constituents of natural reality are fixed by linguistic rules. Thus, to know *what* we are to study in a study of alpha particles is to grasp the definition of the concept *alpha particle*.

Empirical task:

What are the empirical natures of alpha particles. That is, what are the natures of the referents of the concept *alpha particle* (those constituents of natural reality to which the concept *alpha particle* is correctly ascribed).

The fruitful investigation of empirical questions *presupposes* clarity with respect the concepts in terms of which they are stated. No empirical study could succeed in yielding facts about, e.g.,

bachelors, unless it involved the study of those men who do not have wives. And to grasp that the focus of a study of bachelors are men who do not have wives is to grasp a linguistic rule: "the concept *bachelor* is ascribed to men who do not have wives."

The fruitful study of horsepower involves the resolution of both conceptual and empirical issues. But Professor Lubinski, under the influence of construct validity, willfully confuses these issues to such an extent that his account no longer has any relation to science. He notes that "verbal definitions...are always problematic because they lack consensus..." (p.19) and, as have so many social scientists before him, takes this as grounds for the scientist to sidestep the difficult conceptual work whose resolution is a precursor to fruitful science (*he* sloughs off the job to construct validity while latent variable modelers pass it off to the Central Account). Scientists struggle long and hard to clarify the correct employments of the concepts that will denote the phenomena of interest to them. This, once again, is because to know what one must study in a study of ϖ-things is nothing other than to grasp which constituents of natural reality are denoted by concept "ϖ". Thus, Mach spends forty pages analyzing the coherence of Newton's definition of the concept *mass*, while one of Einstein's greatest contributions was to clarify certain ambiguities in the correct employment of the concept *simultaneous events*.

Note: The correct employments of the ordinary language psychological concepts that denote psychological phenomena do not lack for "consensus." If they did, we would neither be able to communicate with each other, nor *teach* the correct employments of these concepts to language learners, nor *correct* others' misemployments. The problem with these concepts is that they have incredibly complicated rules of correct employment, rules that take many years to master, and then, once overlearned, are rarely actively reflected upon. A great deal of skill is, therefore, required to clarify the rules that fix their correct employments.

What does the scientist have to say about horsepower? Obviously, his aim is to study empirical issues centering on horsepower. But what feature of natural reality must then be his target? The answer to this question is a conceptual matter, as every scientist knows. Here is what Bueche (1972, p.83, Principles of Physics) has to say:

"Does a baseball player work when he is playing baseball? Many people would say that since he is playing a game he is not working. But what if he were being paid to play baseball? Is the ground underneath a house doing work? It is holding the house. Is it, therefore, basically different in its function from a pillar holding the roof over the porch of the house? Yet some

would insist that the pillar was doing work. Clearly, if we are to use the term *work* in physics, we need to define it in a precise way."

Thus, Bueche points out the fact that the physicist cannot usefully proceed to his empirical investigations of work unless the term *work* is precisely defined. To define the term precisely is to settle what will be investigated in investigations of work. Bueche also implies that the definition to be provided will be a technical homynym of ordinary language senses of the term *work* (the latter, but not the former, senses prohibiting us from coherently stating that, e.g., "the ground underneath the house is doing work"). In short order, Bueche defines the term *work* followed by the term *power* (def'n: power is the amount of work done in a unit of time). There are available various units in which measurements of power may be expressed, examples being the Watt (Joules/Sec), foot*pounds per second, and Horsepower (1hp is equal to both 746 Watts and 550 ft*lb/sec).

Note: Empirical findings cannot establish 1hp to be equal to 746 Watts. This equivalence is established by the rules that fix the correct employments of *Watt* and *horsepower*. It is not factually incorrect but *nonsensical* to assert, e.g., that 1hp is equal to 500 Watts. One either grasps that 746 watts is equal to 1hp, or one has failed to grasp a rule of correct employment of these units of power.

Having settled the rules that fix the correct employment of the term *horsepower*, the scientist is then free to conduct empirical investigations into "horsepower phenomena." Horsepower is a unit for expressing the power generated by any constituent of natural reality that can, in fact, generate power (i.e., produce work in a particular unit of time). Presumably, then, empirical investigations into horsepower will be investigations into the horspower produced, under particular conditions, by such work producing entities (e.g., machines, people, structures, etc.). Such investigations might centre on the causes of an entity's capacity to generate high horsepower under particular conditions, or, alternatively, its inability to produce an expected level of horsepower. Empirical laws involving horsepower are generalized descriptions of the capacities of work producing entities to generate horsepower. Such laws, *pace* the construct validators, can have no direct implications for the meaning of the term *horsepower*. On the contrary, fruitful empirical investigations centering on the horsepower generated by various constituents of natural reality, and the laws that arise from such investigations, *presuppose* clarity with respect the correct employment of the term *horsepower*.

Professor Lubinski follows the confused lead of Cronbach and Meehl in running roughshod over science's most essential division of labour (that between the conceptual and empirical facets of investigation) and the result is a predictable sequence of incoherent statements, some of which are the following:

i. Professor Lubinski claims that "...you can't 'see' horsepower." However, this is a category error arising from inattention to the conceptual work that is a precursor to legitimate science. Not everything involved in empirical investigation can be legitimately placed on the dimension that runs from perceptually unobservable to perceptually observable. Material and other entities that are extended in space may legitimately be characterized as observable or unobservable in relation to particular observational setups. But, e.g., hopes, dreams, desires, intelligence, concepts, and units of measurement such as horsepower can neither be said to be perceptually unobservable, nor observable. *They are not the right kind of thing to be so*. To observe "big horsepower" is not to observe indicators of an unobservable entity called "big horsepower", but, rather, to observe what a work producing entity can do as a result of its capacity to generate big horsepower (as when one observes a speedboat with a 300hp engine racing over the water. One is here observing what the *boat* can do as a result of it's engine's 300hp rating).

ii. On page 10 of his paper, Professor Lubinski states that "The discipline of psychometrics has developed instruments for dealing with psychological phenomena remote from personal experience. Psychological constructs are 'removed' from experience because they co-occur with other phenomena. Multiple behavioural episodes are necessary to detect them." (p.10)

Psychological phenomena are just those phenomena that are denoted by psychological concepts. This phenomena (e.g., anxiety, anger, intelligence) is not "removed" or "remote" from experience, but rather is precisely what we *do* experience when we experience others' psychologies. On the other hand, the *causes* of these phenomena are very often "removed from experience" in the sense that we do not typically experience, e.g., neurochemical processes. Finally, it makes absolutely no sense to claim that psychological phenomena are removed from experience *because* they co-occur. This view is based on an unacknowledged (and indefensible) premise that is tied into the Central Account, namely that "true" psychological phenomena (constructs) are the unobservable nodes of covarying observables. The *causes* of co-occuring psychological phenomena might well be responsible for these co-occurrences, but the causes of psychological phenomena are not the "true" psychological phenomena, but, rather, the causes of this phenomena.

iii. "To be clear, the g construct is not a 'thing'; it's an abstraction like horsepower. There are different components to horsepower, such as carburetors and cylinders, but still there is a general property. The overall functioning of this property can be increased by tinkering with the components individually, tinkering with the whole system or tinkering with fuel: there are a variety of different variables underlying horsepower as there undoubtedly are with g" (p.28)

What is to be made of this? One of the commentators on Professor Lubinski's article was of the impression that g was being taken to be a 'thing' (a charge Arthur Jensen attempted to rebuff), and this is understandable in light of the fact that Professor Lubinski repeatedly conflates issues pertaining to the *concept* "general intelligence", on the one hand, and the empirical characteristics of its alleged referents (whatever that be), on the other. So, let me try to unpack this tangle.

a. To say that there are different components of horsepower might mean that there are different components to the *concept* "horsepower". Indeed, these components are the concepts *work* and *unit of time*.

b. On the other hand, it might mean that multiple components are part of a satisfactory causal account of the amount of horsepower that particular entities can generate under particular conditions.

c. On no account are carburetors, cylinders, and fuel "components of horsepower." They are components of combustion engines, and properly functioning combustion engines have the capacity to generate horsepower (i.e., the work that they produce per unit of time can be measured and expressed in horsepower). As a consequence, carburetors, cylinders, and fuel must be part of a proper causal account of the amounts of horsepower that combustion engines can generate under particular conditions. Thus, it may be true that switching from fuel B to fuel A results in engine B's being able to deliver an additional 5hp, or, less controversially, that a combustion engine without a carburetor will be unable to generate any horsepower (i.e., it will be unable to do any work).

d. Horsepower is not an abstraction of any sort. It is a unit in which measurements of power may legitimately be expressed. The g construct, on the other hand, is a muddled mystery (which is why scientists often have no idea what it is about). It gets its wings from a tradition of false

identifications with the ordinary language concept *intelligence* (and related terms). In fact, it is striking that Professor Lubinski claims that the conceptual underpinnings of g "...were embryonically embedded in differential psychology's origin" (p.7). Construct validators, having failed to grasp the place of language in science, time and again fail to notice that the "constructs" of interest to them are precisely the concepts of ordinary language. And this is why they come to feel that they have always had "hunches" about their constructs. The concept *intelligence* (related terms and cognates) has been a part of the language for a very long time. The OED records the sense of the term as denoting understanding in degree as dating back to 1430 (Grafton (1568): "That some learned Englisheman of good intelligence would...confute such errors"). The correct employments of psychological concepts are widely ramifying, a property that, when misunderstood, can produce in competent speakers the feeling that they possess certain primitive *empirica*l hunches. But it is a grammatical feature (a feature of the rule goverened employments of concepts) that, e.g., a methodical thinker reasons. It violates correct usage to judge someone a methodical thinker but deny that they reason.

If g were a legitimate candidate for scientific investigation, then a definition could be laid down which would *settle* the reference of the term "g". The "definition" provided in Professor Lubinski's paper simply borrows certain components from the ordinary language concept *intelligence*, thereby leaving wholly unsettled the role that factors, components and other variates are to play in a scientific study of g.

e. After observing the fact that combustion engines are comprised of multiple parts that jointly play a role in the capacity of the engine to generate power, Professor Lubinski concludes, "...but still there is a general property" (p.28)

Of course there is! That there is, is ensured by the fact that language contains a property-term *power* that can legitimately be ascribed to entities on the basis of the work that they produce per unit of time (i.e., in accord with the linguistic rules that fix the correct employments of the term *power*). The number of distinct elements inherent to a satisfactory causal account of the power outputs of particular entities has no bearing on this. Linguistic rules are autonomous of facts. If it weren't for the fact that humans possessed the concept *power* (measurements of power expressible in horsepower), there certainly could be no investigations into the causal story of the power productions of, say, combustion engines (this is why Bueche *begins* his chapter on work with definitions of *power* and *horsepower*, before moving on to facts that physicists have learned about the laws governing the power outputs of various types of entity)

iv. In the first quote given in this case study, Professor Lubinski referred to horsepower as a "postulated attribute." This is confused. Horsepower is an honest to god unit of measurement for the expression of the power that work producing entities can generate. This is all fixed in language and there is nothing postulated about it. Similarly, it is wholly unclear what Professor Lubinski means by the expression "postulated general cognitive ability." What is being postulated? People can correctly be said to possess abiltities, capacities, skills, and the like, cognitive and otherwise, because the language contains ability, capacity, and skill terms, and these terms can be correctly (and incorrectly) ascribed to people on the basis of their behaviour. A brilliant student is a student who can *do* certain things (reason, problem solve, present highly articulate arguments, etc.). We teach our children the grounds for correctly ascribing such terms to others. Perhaps what is being postulated (hypothesized?) is the existence of a single *cause* of the multitude of cognitive capacities in terms of which people can be characterized. Or perhaps what is meant is that a person's possession of some particular cognitive capacity is sufficient for their acquiring a multitude of other cognitive capacities. I do not know, and, based on his paper, I do not think that Professor Lubinski knows. And scientific endeavours cannot flourish in this garden of confusions.

v. Professor Lubinki expresses the view that a high Spearman-Brown coefficient indicates that "...a reliable source of individual differences has been established...attention naturally turns to its psychological nature" (p.11). This is, of course, an expression of the Central Account. The idea is that the psychometrician casts his net into the sea of psychological objects and, under certain conditions, comes up with something. He must then make out the name of his unobservable haul. But, once again, psychometric tools cannot be used as detectors. They are not built to play this role. To coherently claim that, e.g., we have detected a strong magnetic field, presupposes that we are able to antecedently *identify* magnetic fields when they are present. And to be able to do so is equivalent to grasping the rules that fix the correct employments of the concept *magnetic field*. In advance of constructing a tool for the detection of magnetic fields one must grasp which constituents of natural reality the concept *magnetic field* is correctly ascribed to. Without the capacity to do so, there can be no way of showing that the detector behaves in some particular manner in the *presence* of magnetic fields.

Representational (axiomatic) measurement. Representational or axiomatic measurement was once seen as the key to placing the social and behavioural sciences on the same firm ground on which sit the natural sciences. Developed by mathematical psychologists such as Luce, Kahneman, and Krantz, the idea behind representational measurement is quite simple. Essentially, the question is whether isomorphisms can be established between empirical and numerical relational systems, and, if so, what are their properties. Lord and Novick make mention of the construction of isomorphisms, as do McDonald and Mulaik (1979, p.301): "Measurement constitutes the assignment of numbers to objects in such a way as to represent empirical relationships by numerical relationships." Joel Michell (1990) has recently written extensively on the topic, claiming that this approach is the only approach capable of yielding measurement within the social and behavioural sciences.

The belief that the techniques of representational measurement deliver measurement is, herein, disputed. It is not, however, suggested that the insights of the axiomatic/representational approach are not profound, and of great import to the social and behavioural sciences, but only that application of these insights cannot deliver measurement. The axiomatic/representational specialist seems not to realize that to claim that a set of scores are measurements of the σs of δs is to claim that the scores are signified by concept "σ", and, hence, that such a claim is necessarily about the grounds of correct employment of concept "σ". When one claims that one has produced measurements of *anxiety*, one is making a claim that the scores thus produced are signified by the concept *anxiety*. To support such a claim, one must provide grammatical evidence that the concept *anxiety* can legitimately be applied to these scores. And such a case will turn on a comparison of the rules that fix the correct employment of the concept *anxiety* with the rules that were, in fact, followed to produce the scores. The case does not come down to particulars of data, as *does* a representional case. What the axiomatic/representational approach is capable of delivering are high quality numerical *representations* of features of natural reality. A representation of something in terms of something by something else requires the laying down of a set of bridging rules to establish the relation of representation to that represented. A sophisticated understanding of such rules is what has been provided by the inventors of the axiomatic approach. But to construct representations *presupposes* the capacity to antecedently identify the relata of such representational relationships. That is, if one is to legitimately claim that B is a representation of A, one must already grasp the rules that fix the correct employment of concept "A" that denotes A. Such conceptual knowledge resides at the same level as measurement, for rules are constitutive for both. The construction of representational cases, on the other hand, *presupposes* such knowledge. Mere representation is not measurement.