Measurement as a Normative Practice

Implications of Wittgenstein's Philosophy for Measurement in Psychology

Michael D. Maraun
SIMON FRASER UNIVERSITY

ABSTRACT. Recently, a number of prominent measurement specialists (e.g. Cliff, 1992; Schonemann, 1994) have pondered the lack of progress in the development of convincing solutions to the measurement problems of psychology, and have attempted to identify the factors responsible for this lack of progress. They suggest a number of possibilities, including a basic lack of talent in the ranks of the social sciences. It is argued here, however, that the philosophy of Wittgenstein provides an interesting alternative explanation. Specifically, despite their apparent differences, current approaches to the support of psychological measurement claims are unanimous in viewing measurement as chiefly an empirical matter. On Wittgenstein's account, however, this is a mischaracterization of measurement, for, as he argued in elaborate detail, measurement is a normative, rule-guided practice. Hence, empirically based argument is not relevant to the support of measurement claims. If this verdict is correct, it explains not only the failure of measurement theory in psychology, but the much discussed success of measurement in the physical sciences. In this paper, Wittgenstein's characterization of measurement, and its implications for psychology, are discussed.

KEY WORDS: measurement, normative practice, psychometrics, rules, Wittgenstein

In a recent article titled 'Measurement: The Reasonable Ineffectiveness of Mathematics in the Social Sciences', the psychometrician Peter Schonemann (1994) considers why, after years of effort, there still do not exist convincing solutions to the measurement problems of psychology. He wonders why in psychology there have been few examples of 'The miracle of the appropriateness of the language of mathematics for the formulation of the laws of physics' (Schönemann, 1994, p. 157). And he shares Norman Cliff's (1992) lack of optimism regarding the possibility of measurement in psychology: 'we are not just dealing with another "paradigm shift", but quite possibly

THEORY & PSYCHOLOGY Copyright © 1998 SAGE Publications. VOL. 8(4): 435–461
[0959-3543(199808)8:4;435–461;004903]

with the prospect that psychology will never make much progress towards becoming a quantitative science' (Schonemann, 1994, p. 150). He provides several explanations for the failure of measurement theory, including the possibility that there is a lack of talent in the ranks of psychology, and that social science has become big business. Even earlier, there were accounts that suggested that not all was well with psychological measurement (see, e.g., Campbell, 1921; Guttman, 1971). While the explanations Schonemann provides are undoubtedly contributing factors to the malaise, a familiarity with Wittgenstein's commentary on measurement and psychology suggests an altogether more serious problem. For Wittgenstein, both directly and indirectly, argues the following:

1. Measurement is a normative, rule-guided practice.
2. The logical basis for correct measurement, since it is constituted in rules, exhibits the standard autonomy of grammar (rules) with respect to empirical calculations, claims and discoveries.
3. The justification of measurement claims cannot therefore be a matter of empirical case-building, but instead rests on establishing that actions are in accord with the rules that determine what constitutes correct measurement.
4. It is incoherent to claim that whether, and how, something can be measured is a matter of *empirical discovery*.
5. Measurements of 'σ' are, tautologically, denoted by σ, and denotational relations are established in grammar. Hence, the grammar of concept σ should be the focus in debating measurement claims involving σ.

If these theses are correct, the explanation for the difficulties surrounding psychological measurement is clear:

A. Measurement practice in psychology misdiagnoses the nature of measurement, since it is uniformly formulated under the assumption that measurement claims are justified in large part through empirical case-building.
B. The difficulty faced by psychologists in measuring is not mathematical or empirical in nature, but is instead that the concepts they wish to have enter into their measurement operations are typically of the common-or-garden variety. These concepts have notoriously complicated grammars. In light of (5), this generates serious difficulties.
C. The relative lack of success of measurement in the social sciences as compared to the physical sciences is attributable to their sharply different *conceptual* foundations. In particular, the physical sciences rest on a bedrock of technical concepts, while psychology rests on a web of common-or-garden psychological concepts. The fact of (B) completes the explanation.

Measurement as a Normative Practice

On Wittgenstein's account, measurement is fundamentally normative, or rule-governed (Wittgenstein, 1953, 1976). He compares it, directly and indirectly, to the logical structure of language in general, with its constitutive rules, and to specific philosophico-grammatical issues (Wittgenstein, 1953, 1967a, 1976, 1993). He states, for example, that:

> I measure the table in inches and go over to centimetres *on the ruler*.—And of course there is such a thing as right and wrong in passing from one measure to the other; but what is the reality that 'right' accords with here? Presumably a *convention*, or a *use*, and perhaps our practical requirements. (Wittgenstein, 1967a, p. 6)

> So I must make a general remark about grammar and reality. Roughly speaking, the relation of the grammar of expressions to the facts which they are used to describe is that between the description of methods and units of measurement and the measures of objects measured by those methods and units. (Wittgenstein, 1993, p. 448)

When individuals claim that 'These are measurements of σ', or 'The way to measure the "σ" of "δ" is...', their claims are not supported through empirical investigation. Instead, the justification for claims such as these comes from: (1) the existence of a rule-guided practice (i.e. a set of conventions) for measuring σ; and (2) having correctly followed the rules that govern the practice.

Generally speaking, a *practice of measurement* is a cluster of related skills and activities engaged in by a community of individuals (see Ter Hark, 1990, for a detailed discussion). It typically includes techniques for measuring σ using particular instruments, the making of measurement claims about σ, standard methods for verifying the correctness of measurements and measurement claims, methods for comparing objects to canonical samples (e.g. a book to a ruler), rules for the translation of units of one type into units of another, and the teaching and learning of aspects of the practice. It is rule-guided in that there are in fact standards of correctness (i.e. rules) that determine what it is to perform correctly the various activities that define the practice. To teach aspects of the practice *means* to teach the rules that are standards of correct behaviour, while to have learned to measure is to understand the rules that define correct behaviour. Wittgenstein (1990a) states that 'As children we acquire concepts and what one can do with them simultaneously.' Rules are *constitutive* for the practice: They determine what it means to measure σ, translate from one unit of measurement into another, make a correct measurement claim, and so on.

There are a number of characteristic features of a measurement practice, two of which will be mentioned. First, if challenged to defend the claim 'These are measurements of σ', an individual will point, as it were, to the rules that govern what it *means* to measure σ correctly: 'the measurements were taken as follows . . .'. For example, an individual called on to justify the claim that the width of his yard is 48'5" will recite, as it were, the *way* in which his number was arrived at. 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 the yard, just because the action of twisting a tape measure around object A does not accord with the rules for measuring the width of object B. It is not, however, that the individual lacks a more sophisticated justification, but that there exists no other justification for his claim. To put it another way, justification runs out at the level of rules (Ter Hark, 1990; Wittgenstein, 1953):

> The rules of grammar are arbitrary in the same sense as the choice of a unit of measurement. . . . The rules of grammar cannot be justified by shewing that their application makes a representation agree with reality. The analogy between grammar and games. (Wittgenstein, 1974, p. 29)

Second, measurement claims about σ make sense to individuals who understand the rules that govern the practice, and make little sense to those for whom the practice is foreign. The rules are constitutive for the practice, and so without an understanding of the rules, measurement claims lack meaning:

> Now a series can be said to be of infinite length if there is a method of measurement. The sense of this statement, like that of 'This rod is three yards long,' depends on how we determine its length, and differs according to the method of measurement. A method of measurement must be given before the statement can make sense. (Wittgenstein, 1979, p. 209)

Furthermore, if John has not followed the rules for the measurement of σ, then he has not measured σ. This is because a measurement of σ is something taken in accord with the rules and nothing more. There is no mystery as to what it is to measure σ, and correlatively no empirical discovery that could reveal a separate *true* approach to the measurement of σ, since rules are necessarily public (see Hacker, 1988, for further clarification). Any claim John makes to the effect that 'These *are* measurements of σ' can therefore be justifiably contradicted by another individual, and the contradiction will involve a comparison of John's behaviour (the behaviour that generated the numbers) to the rules. The point, then, is that an individual cannot just do any old thing and legitimately call it the measurement of σ, any more than he or she can do any old thing and legitimately call it stopping at a stop sign, playing chess or multiplying 10 and 2.

Additional detail could be provided on many fronts. For the present work, however, what is required is further clarification of the logical character of a measurement claim. Standardly, a measurement claim has the form: The σ of δ is t_i, in which t_i, the measurement, is given in particular units. The practice of measuring the σ of δ in particular units is the basis for a *grammatical* denotational relation between σ and t_i. That is:

> *M1*: Numbers t_i that are measurements of the σ of δ are *denoted* by σ. But denotational relations are established in grammar. In this case, the denotational link between σ and the t_i is established by taking the t_i in accordance with the rules for the measurement of the σ of δ. Hence, measurement claims involving σ are bound inextricably to the grammar of σ.

To put this another way, inquiring as to how one justifies the claim that a set of numbers are measurements of some particular property σ is, inter alia, equivalent to inquiring how one justifies a denotational claim involving concept σ. But a denotational claim is justified by having correctly followed rules for the application of a concept, and rules of application are the substance of grammar. Hence, if one can legitimately take measurements of the σ of δ, then the grammar of σ will reflect the grounding of σ in a practice of measurement. For example, the grammar of *height* contains rules that specify legitimate units of measurement (e.g. inches, feet), specify the correct translation of one type of measurement unit into another (e.g. 12 inches is a foot), disqualify phrases such as '20 lbs high', and so on. Because of this grammatical basis there can be no mystery about whether something can be measured. On the other hand, because grammar determines what legitimately can be done with concepts, whether something can legitimately be measured is indeed a relevant question.

Measurement and the Logical Character of Rules

In a practice of measurement, rules determine what constitutes correct measuring behaviour. Certain features of the logic of measurement are therefore dependent upon the characteristics of rules themselves. For the present work it is necessary to review a number of these characteristics. This review depends heavily on a paper by Hacker (1988).

R1: Rules are human creations. They are formulated in a language, and are standards of correctness for various species of human behaviour.

R2: Rules are not descriptions of human behaviour. The rule 'Stop at the intersection', usually formulated as a concrete symbol (i.e. a 'stop sign'), is a standard of correctness for driving behaviour (i.e. a driver is correct if she stops at the intersection), but is not a description of what actually occurs (e.g. she may decide to drive through the intersection without stopping). In fact, while the rules of language do establish the meanings of concepts that denote objects and events, they are nevertheless autonomous of the particular empirical relations into which these objects and events enter (Baker & Hacker, 1982). A standard of correctness does not determine anything empirical, and, conversely, empirical happenings do not have direct implications for a standard of correctness.

R3: Rule-governed behaviour is regular, but the mere regularity of behaviour is not a criterion for rule-governed behaviour: Individuals must *intend* their behaviour to be in accord with a rule for it to be rule-governed (Hacker, 1988). Not that they must always be aware of the rule while engaging in the activity that is governed by the rule, but when necessary they must be able to cite the rule as a justification for their behaviour. The citing of a rule as a justification for behaviour is a component of a normative practice, as are the teaching and learning of the rules of the practice, the correction of mistakes with reference to the rules, and so on.

R4: A practice is not a mere correlate of rule-guided behaviour, but instead is a *pre-condition* for the existence of a rule. A rule presupposes a practice because *it* cannot interpret whether an action is or is not in accord with *it*; only humans engaged in the practice can make this determination (Hacker, 1988). Yet without a means of determining whether a rule has been correctly followed, it is not possible to distinguish between *correct* and *incorrect* behaviour, and so the rule lacks meaning. Hence, there must exist a standard practice in which the rule is actually employed by humans as a standard of correctness (Hacker, 1988).

R5: To understand a rule is to understand what actions accord with it (Hacker, 1988). For instance, it would not be possible to understand the rule 'Stop at the intersection' without at the same time understanding that 'stopped at intersection behaviour' accords with it. *A rule and the behaviours that accord with it are thus internally, or grammatically (as opposed to empirically or contingently), related* (see Ter Hark, 1990, for further clarification). R5 further explains M1, which can now be seen to assert that a concept that is embedded in a practice of measurement is *internally* related to the measurements it denotes.

R6: A rule or criterion cannot be learned as an object of knowledge (e.g. through empirical investigation). This follows as a result of the internal relation between a rule or criterion and the actions that accord with it. A rule is, in other words, autonomous with respect to empirical phenomena. Wittgenstein (1990b) states, for example, that

> Whether a phenomenon is a symptom of rain, experience teaches; what counts as a criterion of rain is a *matter of agreement*, of our determination. (Definition)

A rule ('criterion') cannot be discovered, established via empirical *facts* ('symptoms') or detected from a careful study of the empirical. This is because, in a given context, the only thing that could establish the claim that 'There is in play a rule ϕ' as correct or incorrect is a comparison of the claim to the rule itself (if there is indeed a rule in play). It is obvious, however, that such a comparison would *presuppose* an understanding of the rule in play itself. Hence, there is no such thing as empirical evidence establishing the nature or existence of a rule. Rules are autonomous and non-discoverable (Ter Hark, 1990). It is tempting to counter that in certain instances, given *these* facts, one could conclude that there was at least a high probability that 'There is in play a rule ϕ'. Once again, however, such a claim would be empty unless it could be compared to a standard of correctness, and in this context there exists no standard of correctness to judge probabilistic support. Belief that 'ϕ is the case' is not the same as 'ϕ being the case'.

As an example of the foregoing discussion, consider the rule 'Drivers must stop their vehicles when the light is red'. This rule could never be discovered or established by empirical investigation since any empirical case that could be made could only be judged as good or bad, correct or incorrect, relevant or irrelevant, and so on, by comparison to the rule itself. That is, knowledge of the rule itself would be required. As a second example, assume that, based on my observations of traffic, I claim that there is a rule to ρ. What would make my claim correct or incorrect? How could I verify whether my claim was right or wrong? Certainly any empirical results I provide could say nothing on this matter since they too would be in need of a standard, in this case a standard of relevance. The only standard of relevance, however, is the rule itself (i.e. that the traffic is in fact running in conformity with ρ). That is, I am correct if there is, in fact, a rule governing the practice, and if my hypothesis identifies it. But I can't establish whether this is the case unless I somehow *understand* the rules that govern driving behaviour. These rules are *encoded* in books and are known by those with a mastery of driving. But if I consult these sources then I am *learning* about a standard of correctness, not verifying anything by empirical means.

The Autonomy of Measurement

To this point, Wittgenstein's views may be summarized as follows. Measurement is a species of rule-guided behaviour. It is dependent on the logical contours of rules. Rules are autonomous with respect to empirical facts. They are not discoverable, and cannot be revealed through empirical investigation. From these points, we now arrive at a basic insight of Wittgenstein's: There are a number of senses in which measurement itself is autonomous with respect to empirical facts, cases, events and happenings. In fact, Baker and Hacker (1980) claim that Wittgenstein takes measurement to be a paradigm case of the autonomy of grammar with respect to the empirical. In what ways does measurement exhibit this autonomy? There are three of particular importance to psychology:

M2: No empirical finding can refute or support a measurement claim. For example, the claim that '*These* are measurements of depression' cannot be established as correct or incorrect by considering or examining the actual numbers recorded, nor the correlation of these numbers with other sets of numbers (e.g. measurements of self-esteem). This is because the correctness of the claim is established by comparison to rules, and, by R6, rules are autonomous with respect to empirical evidence. To put this another way, rules are constitutive for the meanings of the concepts that organize empirical 'evidence': *These* empirical findings are not about, for example, *depression* at all unless one can *already* make the case that they are based on measurements of depression. The correlation between the sets of numbers x and y is not the correlation between, for example, depression and self-esteem, unless x and y are *already* justifiable *as* measurements of depression and self-esteem, respectively. Similarly, no set of empirical facts about a set of numbers x, no matter how idiosyncratic, could justify a refutation of the claim that the 'x are measurements of depression' if, in fact, the x were taken correctly *as* measurements of depression. Instead, the autonomous correctness of the measurements would necessitate the explanation of some possibly idiosyncratic facts about depression.

One might counter that if a set of numbers were presented as, for example, 'measurements of the heights of human beings in feet', and included the number 500, this would be an example of how empirical *facts* (in this case that humans don't grow to 500 ft) can legislate on the felicity of measurement claims. However, appearances aside, 'facts' do not legislate on the correctness of the measurements. For to know that humans do not grow to 500 ft *presupposes* a practice of measuring the heights of objects, and this practice is grounded in rules for measuring. Knowledge about the heights of objects is accumulated by measuring correctly (i.e. by following the rules). If challenged to justify the claim that humans do not grow to 500 ft, one would ultimately have to cite the rules for measuring the heights of objects. In other words, what makes *this* a *fact* is that it was generated according to correct measurement practices. Thus, while empirical facts can be used to cast doubt on the accuracy of particular measurements, they themselves are dependent on the web of rules that provide *justification* for measurement claims, and thus are equally in need of justification.

M3: For similar reasons, nothing intrinsic to a set of numbers, nor their relations to other numbers, can reveal whether they are in fact measurements of *depression*, *self-esteem*, or anything else. What would make t_i a measurement of the σ of δ is the fact that there does exist a practice for the measurement of the σ of δ, and that t_i was taken in conformity with the rules that govern this practice.

M4: From R5 and R6, there is no such thing as *discovering* which actions constitute the legitimate measurement of σ, nor whether σ can, in fact, be measured. The answers to such questions are manifest in the rules that govern practices of measuring, and rules are public. They are clarified in conceptual investigations. An understanding of the concept σ (e.g.

intelligence) makes the question of measurement transparent for, if σ is indeed measurable, it is measured in particular units, and is embedded in a *practice* of measurement, aspects of which are taught and learned. There is no *mystery* about which actions result in measurements of σ. One may forget how to measure, be temporarily confused, or require a lesson in the measurement of σ, but there is never any *mystery* about whether or how a σ that can be measured is to be measured. Confusion over what role σ plays in measurement is, inter alia, confusion over its meaning. The grounds for supporting measurement claims must be public, and normative, or else there is no sense to the notion of correct and incorrect measurement.

As an illustration of the previous points, consider the measurement of the heights reached by a projectile thrown into the air by a number of individuals. The numbers derived in this experiment are measurements of *height* by virtue of the fact that they were in fact taken *as* measurements of *height*; that is, in conformity with the rules that determine what it is to measure height. But while this grammatical relation (i.e. the rule and the actions that accord with it) gives the outcomes meaning *as* measurements of height, it implies nothing about the empirical properties of these measurements, nor their relationships with other empirical outcomes, for example that the average height achieved was 15 ft, that Joe recorded a throw of 10 ft. Conversely, nothing about the properties of a set of numbers, for example {6.1, 5.6, 4.8, 6.2}, could enable a determination of whether or not they were in fact measurements of the heights reached by a projectile. Nor could an empirical case be built that would justify the claim that {6.1, 5.6, 4.8, 6.2} were such measurements, for example that the correlation of these numbers with a second set is 'as it should be'. The one relevant issue is whether these numbers were *taken* as measurements of the heights reached by projectiles.

The Place of the Empirical

Wittgenstein's claim that measurement is rule-based, and thus autonomous with respect to empirical discoveries, cases and phenomena in general, may cause concern to the psychologist about the place of empirical considerations. Is it not, for instance, the case that empirical results, for example from psychological experiments, may motivate refinements to techniques of measuring, or the wholesale adoption of new measurement procedures? Wittgenstein's answer is in the affirmative. In a discussion of this issue Wittgenstein (1993) states:

> This is the way in which grammar depends on facts. If so-and-so were not the case—if it were not constant—then we should not be inclined to do with it what we do.

> Similarly, for example, with weighing. Suppose one said 'If the weight varied erratically, then no one would speak of the weight of a body.' This seems stupid. Like saying 'If the foot constantly changed its length, it would have no sense to speak of one foot.' What is *meant* is, If the *rod* changed its length, we would not be inclined to *measure* with it. (p. 311)

He also remarks that

> . . . I could describe the shape and size of this room by giving its length, breadth and height in feet and just as well by giving them in meters. I could also give them in microns. In a way, therefore, you might say that the choice of units is arbitrary. But in a most important sense it is not. It has a most important reason lying both in the size and in the irregularity of shape and in the use we make of a room that we don't measure its dimensions in [microns][2] or even millimeters. (Wittgenstein, 1993, p. 449)

In other words, if 'things' were different, so too would be our measurement practices. If, for example, the gravitational forces acting on earth were different, so too would be our procedures for the measurement of weight. But we also see here the importance of Wittgenstein's distinction between *reasons* and *justifications* in understanding the relationship between measurement and empirical considerations. Measurement is rule-based, and so is arbitrary in the sense that the *justification* (that which establishes correctness) of measurement claims is only a matter of comparing actions to rules. An explanation of the grounds for justifying measurement claims is a product of a grammatical investigation. On the other hand, the *reasons* why particular rules were laid down in the first place often do answer to empirical considerations. Once a rule is laid down, however, it is a part of grammar and no longer answers to empirical phenomena, for it is constitutive for empirical phenomena. The anthropology of measurement, for example why it is that in 1995 we can in fact measure weight with an atomic scale, is external to the *justification* of measurement practices and claims. As Baker and Hacker (1980) state:

> Empirical connections may motivate adopting a rule, and the fact that cessation of certain regularities would undermine the point of having a rule may be conflated with the idea that the sentence expressing the rule asserts that these regularities hold. (p. 174)

They also comment that innovations

> . . . may originate in experiments, but using them as conversion principles transforms the result of an experiment into a rule. (p. 174)

Hence, it is not Wittgenstein's contention that empirical considerations are irrelevant to measurement, but that they are external to the fabric of rules that constitute the basis for the *justification* of measurement claims.

The Enduring Relevence to Psychology of Wittgenstein's Remarks on Measurement

The foregoing discussion raises the possibility that the measurement difficulties of psychology may be a consequence of its having misdiagnosed the grounds for the justification of measurement claims. In particular, despite their differences, the measurement justifications provided within construct validation theory (e.g. Cronbach & Meehl, 1955), psychometric test theory (e.g. Bohrnstedt, 1983; McDonald, 1981) and axiomatic measurement theory (e.g. Krantz, Luce, Suppes, & Tversky, 1971) are all organized around the view that measurement is largely an empirical issue. But this view is in direct conflict with M1 to M4. In fact, on Wittgenstein's account, the proper input to a consideration of psychological measurement claims is the grammar of psychological concepts. Furthermore, it is certainly not the case that a given psychological concept is *necessarily* grounded in a measurement practice. If, however, it was, this fact would be manifest in its grammar. Specifically, its grammar would contain rules for the use of units of measurement, and society would standardly teach rules to children for the measurement of the property it denotes. An understanding of its meaning would, inter alia, be an understanding of these publicly recognized rules, which, if followed, would result in measurements of the property the concept represents. The issue of whether this is the case for psychological concepts will later be discussed. At present, several examples of psychological practice will be discussed. The aim is not to exhaustively characterize psychological practice, but only to briefly examine several of the more prominent approaches.

Construct Validation Theory

On the construct validation account, one attempts to justify claims like 'This scale measures *intelligence*' and 'The meaning of *dominance* is "..." ' by engaging in empirical case-building. The justification for measurement claims is taken to rest on empirical findings, derived in the same manner as in any scientific investigation. Thus, an investigator who wishes to establish that his or her scale measures *dominance* might well factor analyse the items of the scale, hoping that they will turn out to be unidimensional:

> If a trait such as *dominance* is hypothesized, and the items inquire about behaviours subsumed under this label, then the hypothesis appears to require that these items be generally intercorrelated. (Cronbach & Meehl, 1955, p. 63)

The general view that meaning and measurement depend on empirical facts is exemplified by an early statement of Cronbach and Meehl (1955):

> We will be able to say what anxiety is when we know all of the laws involving it; meanwhile, since we are in the process of discovering these laws, we do not yet know precisely what anxiety is. (p. 294)

One can, however, detect the influence of construct validation theory throughout the social sciences. Bohrnstedt (1983), for example, provides a sophisticated account of the means to support measurement claims via construct validation, while, in its *Standards for Educational and Psychological Tests and Manuals*, the American Psychological Association (1985) enshrines construct validation as state-of-the-art.

As seen in a Wittgensteinian light, the construct validation approach is a conflation of conceptual and empirical issues. It is an example of the failure to grasp the normative, rule-based character of measurement, and the autonomy of rules. Specifically, by M2 measurement cannot be justified by empirical results, and by M3 and M4 the attempt to *discover* or establish empirically whether, or how, something can be measured is incoherent. Consider Figure 1, which depicts the correlations among five variables that describe the population of British Columbia over a period of 44 quarters, from 1980 to 1990.[1] This example, while not drawn from psychology, is useful because it features correlations of a magnitude seldom seen in psychology. Four of the variables are explicitly denoted: average number of kilometres driven, number of citizens employed, average insurance premium, and number of vehicles registered. The fifth, variable *X*, is a mystery. All that is known about *X* are its correlations with the other variables. Thus, the characterization of *X* is wholly empirical, resting on its web of empirical relationships (i.e. its nomological net) with other variables. The magnitudes of these correlations

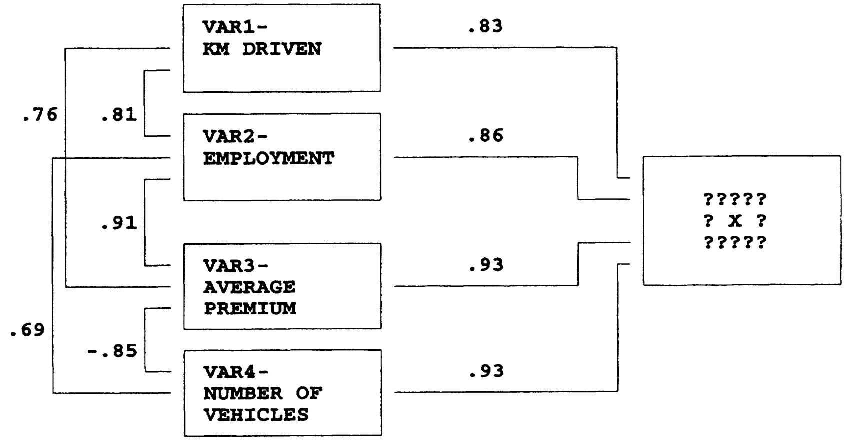

FIGURE 1. Correlations of variables with mystery variable *X*

are, by social science standards, very large. The question, of course, is 'What concept denotes *X*?' or 'What is the meaning of *X*?'

What can be done to answer this question? On the construct validation account, the answer rests on the building of an empirical case. One would attempt, through the interface of theory and empirical results, to stitch together support for a candidate concept. Note, however, that there may not be a correct answer. The 'results' provided may not pertain to anything, and yet there is nothing about their appearance that would mark them as such: Results cannot be judged as good, bad, meaningful, meaningless, relevant or irrelevant from within. On the other hand, it does not a bit of good to *hypothesize* that *X* is, for example, denoted by δ, and then ask for more evidence. A hypothesis is supported by evidence, but here, the standard of correctness is conceptual: *X* either *is* or *is not* denoted by δ. Whether δ can be correctly applied to *X* rests on the grammar of δ. The hypothesizer is left in the lurch as to the adequacy of his or her hypothesis since adequacy in this context is assessed by comparison to a conceptual standard of correctness. Thus, there is no way of knowing when to conclude the investigation, or even whether progress is being made (i.e. 'Are we now closer to the answer than when we started our empirical work?'). Finally, if the claim is that typically we enter into such an analysis with a definition, or, more generally, an understanding of the meanings of the concepts that denote the correlations we compute, and that this is the key to the riddle, then the idea that empirical results are relevant to questions of meaning and measurement is rightly contradicted (i.e. all that was required was initial conceptual clarity). (For a similar point see Guttman, 1971.) For, once again, claims pertaining to denotation rest on standards of correctness for the application of concepts, and understanding the grounds for the application of a concept is equivalent to understanding its meaning. This line of reasoning, in other words, exposes the correlational analysis as irrelevant to the original question.

It is important to once again emphasize that empirical evidence is not downplayed as unimportant by Wittgenstein, but that he considers it to address *different* questions than the logical questions that arise in the consideration of measurement. Psychological investigation is (typically) empirical investigation, and, it could be argued, is predicated on correct measurement practices. However, what can be measured, and how to measure, is not itself an empirical issue, but a logico-grammatical issue. The category errors that are manifest in attempts to solve conceptual questions through empirical investigation are a theme Wittgenstein lectured on extensively. As Baker and Hacker (1982) paraphrase:

> The endemic sin of the experimental psychologist, the sin that explains and justifies Wittgenstein's remarks that 'problem and method pass one another by', is to neglect the conceptual investigations which are preconditions for fruitful, intelligible experiments. (p. 228)

The proponent of construct validation theory is likely to reason that the more one knows about something, the greater is one's understanding of that thing. Knowing about something involves evidence, and so this is akin to stating that our understanding of the moon will increase as more about the moon is discovered. Certainly there is nothing wrong with *this* idea. The problem is that in construct validation theory, *knowing* about something is confused with an understanding of the *meaning* of the concept that denotes that something:

> Scientifically speaking, to 'make clear what something *is*' means to set forth the laws in which it occurs. (Cronbach & Meehl, 1955, p. 290)

This is mistaken. One may know more or less about *it*, build a correct or incorrect case about *it*, articulate to a greater or lesser extent the laws into which *it* enters, discover much, or very little, about *it*. However, these activities all presuppose rules for the application of the concept that denotes *it* (e.g. *intelligence*, *dominance*). Furthermore, one must be prepared to cite these standards as justification for the claim that *these* empirical facts are about *it*. One cannot even begin the exercise of coherent empirical investigation, hypothesis construction, theory-building, and so on, without an understanding of the meaning (i.e. grounds for application) of the concept that will organize this work. A study of what the moon is made of could not begin without a criterion for *moon*. Similarly, one has not begun an empirical study of *memory* unless *these* hypotheses, theories, data, and so on, bear on *memory*: And what is *that*?

The point for the psychologist to consider is that the existence of denotational relations is central to the justification of measurement claims, and these relations are laid down in grammar. Empirical results that are not denoted by a concept are, tautologically, meaningless. One can illustrate the logical difficulties inherent to the construct validation account by drawing two implications from its basic tenets.

(1) Discovery/validation conflation. Consider the construct validation tenet that a test's 'validity' rests on its conforming to certain empirical conditions. Assume that for test T to be a valid measure of θ, T must have a large positive correlation with variable X. Assume further that initially $r(T,X)$ is in fact large and positive, and hence that it is concluded that test T looks reasonable. The aim then is to employ the test in research, or, in other words, in the accumulation of empirical facts. The test is employed in a number of projects, and over the course of a year, it is noticed that $r(T,X)$ moves toward zero. The question is: Is this an important new discovery about the relationship between θ and X (e.g. that their relationship is changing over time), or evidence that T is no longer a valid measure of θ? According to

construct validation theory, the result implies T's invalidity. Yet surely it should be possible to make an empirical discovery about θ and X using T. Wasn't the potential for empirical discovery precisely the reason that T was constructed in the first place? Further, since any result is potentially a part of the nomological net of θ and T, this discovery/validity ambiguity applies to *all* empirical results involving T. And it is not merely a case of uncertainty as to whether a given result is, in fact, about validity or discovery, for in construct validation theory there *exists* no criterion to distinguish the two. Construct validation bars the usual criteria, the rules of language, for distinguishing between empirical discovery and meaning. Hence, in a strange sense they are rendered the same thing in construct validation theory.

(2) Nomological paradox. Consider the construct validation tenet that empirical discoveries can progressively clarify the meaning of a concept, or, in other words, that progressive approximations can be made in the quest for a concept's meaning. Assume that discoveries about θ are made with measurement instrument T. Now, for T to provide relevant empirical findings, that is, findings *about* θ, it must be in line with the meaning of θ (metaphorically, it must have been created in the image of θ). Assume that it is found that $r(T,Y) = 0$, and that we agree that this result changes our understanding of the meaning of θ. This implies that originally we did not have a correct conception of the meaning of θ. But since T was constructed in accord with this earlier understanding of the meaning of θ, T is now seen to be an improper measure of θ. It follows then that $r(T,Y) = 0$ was not really about θ, that is, it was not denoted by θ. Hence, the paradox is that the 'result' $r(T,Y) = 0$ undermines itself: The fact that $r(T,Y) = 0$ implies that this result has no meaning (i.e. it is not denoted by anything)!

Without care these might be taken as examples of the Duhem–Quine thesis (e.g. Quine, 1980), and nothing more. This, however, would be a mistake, and it is worth considering why. The Duhem–Quine thesis speaks to the issue of the support of theoretical postulates by empirical evidence. The previous points, however, are not about empirical underdetermination, or falsifiability at all. Neither are they about the so-called 'theory/observation split'. Instead, they centre on the distinction between conceptual and empirical issues. Logical difficulties like these arise in a system that mischaracterizes measurement, and in so doing fails to maintain the distinction between empirical evidence and the *conceptual* standards of correctness (i.e. rules) that are constitutive for this evidence. Interestingly, it seems that the only post-Wittgensteinian theorists who have recognized this are the operationists, who accused the construct validators of conflating meaning with 'significance' (e.g. Bechtoldt, 1959). The operationist 'solution' to the problem was, of course, different than Wittgenstein's.

An Example from Psychometrics

Krantz (1991) distinguishes the psychometric tradition of measurement from the axiomatic tradition that has its roots in psychophysics. As he describes, while psychometrics plays a role in many facets of psychological investigation (including construct validation), it, in addition, has its own set of guiding principles for the support of measurement claims. Assume that there is data for N individuals on p items, the items denoted by the concept *dominance*. A data analysis is carried out and the items are found to be unidimensional in a factor analytic sense (see, e.g., McDonald, 1981), with equal factor loadings. The loadings are large and so the items jointly form a high reliability composite (see, e.g., Bohrnstedt, 1983; Thissen, Steinberg, Pyszczynski, & Greenberg, 1983). The unweighted sum of the items, T, is then proportional to the maximum likelihood estimate of θ, the latent variate that 'underlies' the items. Does this make the N total scores T_i measurements of *dominance*? I believe that many psychometricians (perhaps with minor qualifications), and practising psychologists, would answer in the affirmative to this question (e.g. Bohrnstedt, 1983; McDonald, 1981). In many ways, this is a paradigm case for the psychometric characterization of measurement, factor analysis having been invented by Spearman (1927) in an attempt to measure intelligence. The answer provided here, however, is in the negative. This is because for the T_i to be measurements of *dominance* they would, by M1, have to be *denoted* by the concept of *dominance*. Denotation, however, is a grammatical matter, resting on the rules of application of *dominance* and not on the covariance structure of the items. The issue is whether any such rules of measurement *are* in fact tied up with the grammar of the concept. If the answer was in the affirmative, then measurements could be *taken* by following these rules. On the other hand, the covariance structure described here was *discovered* (i.e. it was an empirical finding). One follows rules to *get* measurements, but here the one-dimensional structure *occurred*. To put this more firmly, the psychometric account provides no explication of what the technique for measuring dominance involves. Given the psychometric rationale, how would one instruct another to go about taking measurements of dominance? What inspires the psychometric claim is a covariance structure, and no set of instructions can bring about a particular covariance structure.

Further insight is gained when one considers Wittgenstein's comments on the autonomy of measurement, that is, M2 to M4. If one has measurements of height for 10 people, and drops 8 people from the set, the remaining two numbers are still measurements of height. This is so by virtue of the autonomous (of any empirical happening) grounds for judging the correctness of measurement claims. However, in the psychometric scenario, no such autonomy holds. For example, if a percentage of the subjects were dropped from the set, the covariance structure of the data might well change.

The structure might itself then change in such a way that, by the psychometric justification given, the remaining T_i would no longer be measurements. It follows that, on the psychometric account, whether the T_i are measurements depends on the sample chosen, the items chosen, the idiosyncratic behaviour of the individuals responding to the items, the time at which the data were collected, and so on. The T_i may well be 'measurements' one day and not the next.

If the T_i are not measurements, then what are they? The foregoing discussion should make the issue transparent. The T_i are sufficient *statistics* for the joint distribution of persons and items. They are a summary of the data, that is, empirical *findings*. They replace the items in an optimal sense (Maraun, 1996), or, in other words, describe the sample in an especially compact form. The compactness of description of the empirical distribution of responses is the *gift* of a unidimensional structure, and this compactness of description in no way answers to the grammatical requirements of denotation which are central to the making of measurement claims concerning *dominance*.

An Example from Axiomatic Measurement Theory

Roughly the same line of argument applies to axiomatic measurement theory. Consider a simple axiomatic treatment. There are p scenarios, each of which contains a description of a 'dominant behaviour'. Each of N individuals is presented with each of the possible pairs of these scenarios, and is asked to judge, for each pair, which scenario contains the behaviour that manifests the greatest amount of *dominance*. Such an experiment yields a set of proportions, $P(a,b)$, read as 'the proportion of individuals who judged scenario a to be more dominant than b'. Define a binary relation $\gtrsim$ on the set of pairs such that

$$(a,b) \gtrsim (c,d) \leftrightarrow P(a,b) \geq P(c,d).$$

This relation is read as 'the amount of dominance by which a exceeds b is greater than the amount of dominance by which c exceeds d' if and only if $P(a,b) \geq P(c,d)$. Now, the aim is to *represent* the qualitative judgements by way of a numerical mapping (see, e.g., Krantz, 1991). That is, one seeks to establish a homomorphism between the qualitative structure of judgements and a numerical structure. Axiomatic measurement theory involves, among other things, an analysis of the conditions under which such a homomorphism exists. In the present case, the homomorphism, if it exists, is usually taken to be of the following form:

$$P(a,b) \geq P(c,d) \leftrightarrow F(u(a) - u(b))$$

in which u is a function mapping the scenarios onto the real line, and F is a strictly increasing function. If such a representation is possible, the $u(x)$ are considered to be measurements of the dominance manifest by the behaviours contained in each scenario.

Now, it would be difficult to argue with the *representationist* claims of the axiomatic approach. In the previous example, for instance, a set of numbers $u(x)$ reproduces a set of empirical judgements, and, in this sense, represents them. But the issue is whether the $u(x)$ are *measurements* of dominance. And to this question the answer is in the negative. For the mere *existence* of a homomorphism involving $u(x)$ does not even imply the identity (the meaning) of the $u(x)$. But there can be no such thing as a measurement whose identity is in question, for measurements are, inter alia, measurements *of* something. From M1, the identity of t_i as measurements is established when the t_i are taken according to rules of measurement, and rules of measurement are tied up with the grammar of the concept that is to denote the measurements (if, in fact, it *is* embedded in a practice of measurement). But the grammatical rules for the application of *dominance* say nothing about any $u(x)$. Note also that there are a limitless number of homomorphisms that might be considered, each of which would represent a particular feature of the data. We may choose these as we please (as they suit our purposes), for they are each merely a part of the 'results'. We might have different homomorphisms for different populations, points in time, sets of items, purposes, and so on. Measurements are, on the other hand, special in that they are internally related to that which is measured, and hence are not properly empirical at all.

Constraints on Measurement in Psychology

If Wittgenstein's views are correct, then difficulties in psychological measurement may be explained by the fact that psychology has mischaracterized measurement. However, what if one did consider the grammars of psychological concepts. What would grammar say about how to measure psychological concepts? I believe that a grammatical investigation of psychological concepts (e.g. as in Ter Hark, 1990) reveals (a) the obvious fact that, as it stands, common-or-garden psychological concepts are not measurable, and (b) the existence of grammatical constraints on the *possibility* of measurement involving common-or-garden psychological concepts. These constraints may, in part, explain the enduring difficulty faced by psychology in attempting to measure, and, in particular, as compared to the physical sciences (for related discussions of the latter point, see Campbell, 1921; Schonemann, 1994). While (a) and (b) may have the tone of overstatement, I believe that it is nevertheless interesting to briefly consider each. To do so requires that two types of concept be distinguished: (1) technical concepts and (2) common-or-garden concepts.

A technical concept is a concept defined by a specialized or expert community, and employed within a narrow, technical field of application. A common-or-garden concept, on the other hand, is a concept with a common employment in everyday life (Baker & Hacker, 1982). Common-or-garden concepts are taught, learned and understood by the *person on the street*, and have meanings that are manifest in broad, normative linguistic practices. They, in addition, are more apt to be the source of confusion in the context of scientific investigation. Wittgenstein (1976) had already indicated as much, and had discussed a number of particular cases:

> Puzzles may arise out of words not ordinary and everyday-technical mathematical terms. These misunderstandings don't concern me. They don't have the characteristic we are particularly interested in. They are not so tenacious, or difficult to get rid of. (p. 4)

Psychology and the physical sciences differ greatly in their use of technical and common-or-garden concepts (Baker & Hacker, 1982). The physical sciences rest on a bedrock of technical concepts, for example mass, force, gravity, hydrogen, ganglia and neutrino. Importantly, whether a technical concept plays a role in measurement is up to the inventor of the concept, since the meaning of the concept is as stated by the inventor. Mach's concept of *mass*, for example, is measurable *by definition*. Recall that in objecting to Newton's definition of mass on the grounds that it was circular, Mach *proposed* a new *definition* based on the ratio of accelerations induced in a pair of particles:

$$m(A) = \left\{ \frac{-a_{A/C_o}}{a_{C_o/A}} \right\}$$

with a_{A/C_o} the acceleration of particle A induced by a reference particle C_o, and $a_{C_o/A}$ the acceleration of the reference particle induced by particle A (Mach, 1960; see also Falmagne, 1992). It should be emphasized that his claim was not that he had made a startling new empirical *discovery* about mass, or that his empirical investigations had enabled him to establish the *true* nature of mass. Instead, he attempted to persuade the community of physicists to adopt a new standard of correctness, or set of rules, for the measurement of mass. As befitting a definition, as opposed to a discovery, Mach painted the issue as one of usefulness: Overcoming the circularity inherent in the old definition would provide for a coherent conceptual bedrock upon which to base theoretical and empirical work (Mach, 1960, p. 216). If the definition was accepted by the community of physicists, the justification for the claim '*These* are measurements of mass' would be that the measurements were taken as $m(A)$. Mach's definition would be the new standard of correctness for this claim. Note that Mach can provide no *proof* that this definition is 'correct', because definitions are not provable, nor correct or incorrect. Instead, if Mach's definition is accepted by the community of physicists, it is *constitutive* for the practice of measuring mass. That is, it would establish what it is to measure mass correctly, and, a fortiori, what it is to be incorrect in measuring mass.

Technical concepts also arise in psychology. Parmia and premsia (Cattell, 1965), self-monitoring (Snyder, 1974), *g* (Spearman, 1927) and residential school syndrome are technical concepts. Once again, the inventors of these concepts are free to lay down any definition whatever. Furthermore, the measurement characteristics of such concepts follow directly from their definitions. If I want to measure an individual's *grek*, defined as {. . .}, I must simply provide the rules for the measurement of *grek*. Moreover, I cannot be incorrect about this, since I am providing the rules that *establish* what it is to measure *grek*. In the majority of cases, however, the conceptual bedrock of psychology is made of common-or-garden concepts. The reason for this has much to do with the aims of the discipline. Psychology arose from a need to understand the very same phenomena that are of interest to authors, poets and the person on the street, phenomena denoted by common-or-garden concepts. Depression, dominance, intelligence, happiness, fear, motivation, personality, memory, and so on, are of interest to the psychologist, as well as to Tolstoy and Dickens. However, in marked contrast to technical concepts, common-or-garden concepts are not developed, laid down or modified at the outset of empirical investigation. This is because these concepts already have meanings, as manifest in their everyday use, use being governed by grammar. Hence, there exist grammatical restrictions on what one may legitimately do with them. One may define a technical notion, but only *clarify* the meaning of a common-or-garden concept such as depression. Correlatively, one can be shown to be incorrect in claiming that *these* results are about, for example, depression. To put this in more Wittgensteinian terms, when one proclaims interest in studying empirical correlates of depression, there are already in play standards of correctness for the use of the concept. Any claims made that *these* results are about depression are justified only if the concept does in fact denote the result, and this depends on whether the investigator has employed the term correctly. On the other hand, it is not the case that common-or-garden concepts *must* provide the conceptual foundation for empirical work in psychology, but merely that if the phenomena they denote are to be the focus of investigation, coherent empirical work necessitates that they be employed correctly. For when the meaning of a concept is subverted, the link between the phenomena and the concept that was supposed to denote them is severed: The denotational link is not established. According to Wittgenstein:

> Grammar tells us what makes sense and what does not—It lets us do some things with language and not others; it fixes the degree of freedom. (Wittgenstein, 1961, p. 113)

The science of psychology is founded on a lattice-work of common-or-garden concepts. Consequently, the psychologist must very often attempt to support measurement claims involving, not just any old concept, but a common-or-garden psychological concept. But, by M1, to support the claim that, for example, $\{1.2, 3.4, \ldots\}$ are measurements of a common-or-garden concept σ, the numbers must be *denoted* by σ, and denotation is a grammatical matter. Hence, the support of a measurement claim involving σ is inextricably tied to its grammar. *This is the difficulty faced by the practising psychologist, that which sets his or her measurement problem apart from that of the physical sciences. For common-or-garden psychological concepts have notoriously complicated, unsystematic grammars* (see, e.g., Ter Hark, 1990, for a detailed survey). It is indeed an interesting question as to whether such concepts can be brought to heel by the ambitions of psychology.

Now, the claim that common-or-garden psychological concepts are not measurable follows from the simple observation that common-or-garden psychological concepts as they stand are *not* embedded in normative practices of measurement. This is a simple observation, because it rests on whether there exist normative, rule-governed techniques for taking measurements of *dominance*, *intelligence*, *creativity*, *tension*, and so on. The normativity of rules mean that, if they exist, they are *public*, *surveyable* standards of correctness for behaviour. Rules of measurement, if they existed for common-or-garden concepts, would be manifest in descriptions of the correct employments of these concepts (see, e.g., Ter Hark, 1990). But they do not exist. There is no public, *normative* status at all to assertions like 'Tomorrow we are going to measure little Tommy's dominance'. What does this mean? In contrast to the teaching of the use of concepts such as *weight* and *height*, the teaching of the use of concepts such as *dominance* and *intelligence* does not involve the teaching of rules for measuring. There is no common language standard of correctness for a claim like 'I measured Sue's leadership this morning'. In other words, there is no public, standardly taught notion of what it is to be correct in making such an assertion; instead, it sounds merely curious. Furthermore, the grammars of common-or-garden concepts do not contain rules for determining what can count as a unit of measurement, nor for the conversion of one type of measurement unit into another. Wittgenstein would therefore argue that common-or-garden psychological concepts cannot be legitimately measured, at least not if they are 'allowed their proper use'. As he states:

> There are gradations of expecting, but it is nonsense to speak of a measurement of hoping, if one allows the word 'hope' its use. (Wittgenstein, 1990c)

Hence, on Wittgenstein's account, when it is claimed by a psychologist that a common-or-garden psychological concept has been measured, the concept is being misused. More accurately, the psychologist is mixing technical and common-or-garden senses of the term. *Whatever* the Beck Depression Inventory measures it is not depression. And this is not because it has been established that the scale has poor psychometric properties, but instead because the grammar of the concept *depression* doesn't square with *any* measurement operation. It is the mismatch of the goal of psychological measurement with the grammars of psychological concepts that causes the difficulty.

Might it not be countered that many common psychological concepts *are* measurable, and the grammatical evidence is that these concepts are accompanied by modifiers such as 'very', 'somewhat', 'extremely', and so on? For instance, we correctly speak of individuals as being *very dominant*, *extremely intelligent* and *quite creative*. Krantz (1991), for example, states 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 to the goal of numerical measurement of the variable in question. (p. 3)

Despite its current appeal, this line of reasoning may be a bit fast, for a concept's 'measurement-like' grammar is not the same thing as its being grounded in a practice of measurement. A practice of measurement implies normative measurement techniques, instruments and units of measurement, while a measurement-like grammar does not.

On the other hand, it might be argued that there are indeed examples of the valid measurement of common-or-garden psychological concepts, as can be seen in the use of the electroencephalogram to produce measures of *arousal* (Posner, 1975), *g* as a measure of *intelligence* (e.g. Jensen, 1979), and the amount of deprivation as a measurement of *drive* (e.g. Weiner, 1980). On Wittgenstein's account, this argument rests on a conflation of conceptual criteria and empirical symptoms (see Ter Hark, 1990, p. 31). A criterion instantiates a concept, and thus is related to the concept in grammar. It provides grounds for the application of the concept, and so *manifests* what is meant by the concept. An empirical symptom, on the other hand, is dependent on, and so is external to, meaning. The amount of deprivation is perhaps an empirical symptom or correlate of a drive, in the same way that success in life is a correlate of intelligence, but the amount of deprivation is no more a *criterion* for a drive than is life success a criterion for intelligence. Hence, taking the variable 'amount of deprivation' *for* the level of a drive is a misuse of the concept of drive. When we speak of a *drive*, the amount of deprivation suffered is not what is meant (i.e. it is not a part of its grammar), and so is irrelevant to the measurement of a drive.

Let us briefly consider the claim that there are grammatical constraints on the *possibility* of measurement involving common-or-garden psychological concepts. Is it not possible that innovations will occur so that at some future point in time it *will* be possible to measure *intelligence*, *dominance*, *leadership*, and so on? Might it not be the case that this kind of refinement and innovation is precisely the currency of modern psychological measurement practice? Certainly, Wittgenstein himself recognized that refinement and innovation are fundamental to measurement. However, because measurement is a normative practice, real innovation must occur 'naturally', in the sense that the community of speakers employing the concept would have to come to recognize the measurement innovation as a constituent of its grammar. This would (at the very least) require the rules representing the innovation to be squared with the rules that currently govern the use of the concept, as when, for example, metric units are squared with imperial units. Wittgenstein (1967b) remarks that:

> A measurable phenomenon occupies the place previously occupied by a non-measurable one. Then the word designating this place changes its meaning, and its old meaning has become more or less obsolete. (§438)

While allowing for the logical possibility of innovation, features inherent to psychological concepts nevertheless seem to make this possibility an unlikely one. What are the grounds for so pessimistic a conclusion? Simply put, measurement requires a formalization which does not seem well suited to what Wittgenstein calls the 'messy' grammars of psychological concepts, grammars that evolved in an organic fashion through the 'grafting of language onto natural ("animal") behaviour' (Baker & Hacker, 1982). One aspect of this mismatch arises from the flexibility in the grounds of instantiation of many psychological concepts, the property that Baker and Hacker (1982) call an open-circumstance relativity (see also Gergen, Hepburn, & Comer Fisher, 1986, for a similar point). Take, for example, the concept *dominance*. Given the appropriate background conditions, practically any 'raw' behaviour could instantiate the concept. Hence, Joe's standing with his back to Sue could, in certain instances, be correctly conceptualized as a dominant action. On the other hand, Bob's ordering of someone to get off the phone is not a dominant action if closer scrutiny reveals the motivation for his behaviour to be a medical emergency which necessitated an immediate call for an ambulance. The possibility for the broadening of background conditions to defeat the application of a psychological concept is known as the defeasibility of criteria (Baker & Hacker, 1982). Together, open-circumstance relativity and the defeasibility of criteria suggest that psychological concepts are simply not organized around finite sets of behaviours which jointly provide necessary and sufficient conditions for their instantiation (Baker & Hacker, 1982). Yet, this is precisely the kind of formalization required if a concept is to play a role in measurement. Hence, Wittgenstein would certainly have endorsed Schonemann and Cliff's pessimism with regard to psychological measurement, but for very different reasons.

A psychologist might take exception to these views in a number of ways. Ter Hark (1990, p. 191), perhaps somewhat unkindly, states that psychologists are professionally disinclined to concern themselves with the meanings of common psychological concepts. But does a psychologist really require permission? Cannot a psychologist go ahead and employ a term such as *depression* in a new technical sense? Of course there exist no *laws* governing the use of concepts by psychologists: The psychologist may do as he or she wishes. In fact, the practising psychologist frequently retains common terms while substituting in technical uses, or introduces conceptual innovations to common concepts without reconciling these innovations with the normative practice that gives the concept meaning. Unfortunately, science does not thrive under conditions of conceptual confusion, and the covert use of common-or-garden terms to stand for technical concepts is destined to confuse: 'What does investigator A mean by ". . ."?' Even more regrettable is the practice of shifting from a technical sense to a common sense when revealing empirical discoveries to the public; for example, the technical concept *g* (Jensen, 1969; Spearman, 1927) is used in carrying out the research, while the results are reported to the public as discoveries about *intelligence*. This practice is an attempt to have one's cake, and eat it too (Baker & Hacker, 1982), for *g*, a factor analytically defined unidimensional manifold, is something entirely different than intelligence. Hence, while a psychologist does not require permission to employ various locutions under the guise of common psychological terms, confusion is the inevitable result.

Conclusion

Until the measurement problems of psychology are properly diagnosed as conceptual in nature, centring, as they do, on the grammars of psychological concepts, solutions will remain a scarcity. Wittgenstein harped on this theme throughout his writings on psychology. As it stands, the cleverest mathematical formulation, be it axiomatic measurement theory or factor analysis, at best constitutes a means for formulating laws, representing aspects of the empirical, or testing hypotheses about data. But these aspects of investigation, while essential to the carrying out of psychological research, do not bear on the supporting of measurement claims. Wittgenstein's (1953) words to psychology are as relevant today as they were when originally spoken:

> The existence of the experimental method makes us think we have the means of solving the problems that trouble us; though problem and method pass one another by. (p. 232e)

Note

1. I would like to thank Dr J.S. Jackson for allowing me the use of this data.

References

American Psychological Association. (1985). *Standards for educational and psychological tests and manuals*. Washington, DC: American Psychological Association.

Baker, G., & Hacker, P. (1980). *Meaning and understanding*. Chicago, IL: University of Chicago Press.

Baker, G., & Hacker, P. (1982). The grammar of psychology: Wittgenstein's 'Bemerkungen über die Philosophie der Psychologie'. *Language and Communication*, *2*(3), 227–244.

Bechtoldt, H. (1959). Construct validity: A critique. *American Psychologist*, *14*, 619–629.

Bohrnstedt, G. (1983). Measurement. In P.H. Possi, J.D. Wright, & A.R. Anderson (Eds.), *Handbook of survey research*. New York: Academic Press.

Campbell, N. (1921). *What is science?* New York: Dover.

Cattell, R. (1965). *The scientific analysis of personality*. Baltimore, MD: Penguin.

Cliff, N. (1992). Abstract measurement theory and the revolution that never happened. *Psychological Science*, *3*, 186–190.

Cronbach, L., & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302.

Falmagne, J. (1992). Measurement theory and the research scientist. *Psychological Science*, *3*(2), 88–93.

Gergen, K., Hepburn, A., & Comer Fisher, D. (1986). Hermeneutics of personality description. *Journal of Personality and Social Psychology*, *50*(6), 1261–1270.

Guttman, L. (1971). Measurement as structural theory. *Psychometrika*, *36*(4), 329–348.

Hacker, P. (1988). Language, rules, and pseudo-rules. *Language and Communication*, *8*(2), 159–172.

Jensen, A. (1979). *g*: Outmoded theory or unconquered frontier? *Creative Science and Technology*, *11*(3), 16–29.

Krantz, D. (1991). From indices to mappings: The representational approach to measurement. In D. Brown & J. Smith (Eds.), *Frontiers of mathematical psychology: Essays in honor of Clyde Coombs*. New York: Springer-Verlag.

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

Mach, E. (1960). *Science of mechanics. A critical and historical account of its development.* Lasalle, IL: Open Court.

Maraun, M.D. (1996). Metaphor taken as math: Indeterminacy in the factor analysis model. *Multivariate Behavioral Research, 31*(4), 517–538.

McDonald, R.P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology, 34*, 100–117.

Posner, M. (1975). Psychobiology of attention. In M. Gazzaniga & C. Blakemore (Eds.), *Handbook of psychobiology.* New York: Academic Press.

Quine, W. (1980). Two dogmas of empiricism. In W. Quine (Ed.), *From a logical point of view* (2nd ed.). Cambridge, MA: Harvard University Press.

Schonemann, P. (1994). Measurement: The reasonable ineffectiveness of mathematics in the social sciences. In I. Borg & P. Mohler (Eds.), *Trends and perspectives in empirical research.* Berlin: Walter de Gruyter.

Snyder, M. (1974). The self-monitoring of expressive behaviour. *Journal of Personality and Social Psychology, 30*, 526–537.

Spearman, C. (1927). *The abilities of man.* New York: Macmillan.

Ter Hark, M. (1990). *Beyond the inner and the outer.* Dordrecht: Kluwer Academic.

Thisson, D., Steinberg, L., Pyszczynski, T., & Greenberg, J. (1983). An item response theory for personality and attitude scales: Item analysis using restricted factor analysis. *Applied Psychological Measurement, 7*(2), 211–226.

Weiner, B. (1980). *Human motivation.* New York: Holt, Rinehart, & Winston.

Wittgenstein, L. (1953). *Philosophical investigations.* Oxford: Blackwell.

Wittgenstein, L. (1961). *Tractatus logico-philosophicus.* London: Routledge & Kegan Paul.

Wittgenstein, L. (1967a). *Remarks on the foundations of mathematics.* Oxford: Blackwell.

Wittgenstein, L. (1967b). *Zettel* (G. Anscombe & G. von Wright, Eds.). Oxford: Blackwell.

Wittgenstein, L. (1976). *Wittgenstein's lectures on the foundations of mathematics, Cambridge 1939* (C. Diamond, Ed.). Brighton: Harvester.

Wittgenstein, L. (1979). *Wittgenstein's lectures, Cambridge 1932–35* (A. Ambrose, Ed.). Oxford: Blackwell.

Wittgenstein, L. (1990a). Unpublished manuscript (169, p. 71), cited in M. Ter Hark (1990), *Beyond the inner and the outer.* Dordrecht: Kluwer Academic.

Wittgenstein, L. (1990b). Unpublished manuscript (115, p. 72), cited in M. Ter Hark (1990), *Beyond the inner and the outer.* Dordrecht: Kluwer Academic.

Wittgenstein, L. (1990c). Unpublished manuscript (115, p. 73), cited in M. Ter Hark (1990), *Beyond the inner and the outer.* Dordrecht: Kluwer Academic.

Wittgenstein, L. (1993). *Philosophical occasions, 1912–1951* (J. Klagge & A. Nordmann, Eds.). Cambridge, MA: Hackett.

Michael D. Maraun is an assistant professor of psychology at Simon Fraser University. His areas of interest are psychometrics and philosophy of science (including Wittgenstein's philosophy of psychology). His most recent publications include: 'Metaphor Taken as Math: Indeterminacy in

the Factor Analysis Model' (*Multivariate Behavioral Research*); 'The Claims of Factor Analysis' (*Multivariate Behavioral Research*); and 'Appearance and Reality: Is the Big-Five the Structure of Trait Descriptors?' (*Personality and Individual Differences*). ADDRESS: Department of Psychology, Simon Fraser University, Burnaby, BC, Canada V5A 1S6. [email: maraun@sfu.ca]