A CRITIQUE OF MEASUREMENT THEORY IN PSYCHOLOGY: THE CASE OF THE SENSATION SEEKING SCALE

by

Jeremy S. H. Jackson
M.A., Simon Fraser University, 1993

THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
in the Department
of
Psychology

SIMON FRASER UNIVERSITY
August 1996

ISBN 0-612-16929-4

PARTIAL COPYRIGHT LICENSE

Title of Thesis/Project/Extended Essay

A Critique of Measurement Theory in Psychology

The Case of the Sensation Seeking Scale

Author: ______
(signature)

______
(name)

Aug 15/96
(date)

Approval

Name: Jeremy S.H. Jackson

Degree: Doctor of Philosophy

Title: A Critique of Measurement Theory In Psychology: The Case of The Sensation Seeking Scale

Examining Committee:

Dr. Richard Wright, Chair

Dr. Roger Blackman, Senior Supervisor
Associate Professor of Psychology

Dr. David Cox
Associate Professor of Psychology

Dr. Michael Maraun
Assistant Professor of Psychology

Dr. Bruce Whittlesea
Associate Professor of Psychology

Dr. Peter Schonemann, External Examiner
Professor of Psychology

Date Approved: August 12/96

Abstract

Previous investigations of the psychometric properties of the Sensation Seeking Scale (SSS) have employed statistical tests that are inappropriate for dichotomous items. In addition, the dimensionality of the SSS has been assessed with models that entail overly restrictive linear latent variable, item response relationships. In the present study, theory from Holland and Rosenbaum (1986) was employed in a test of the dimensionality of the SSS items that is consistent with the theoretical structure of the SSS and its subscales. It was shown that with the possible exception of the Experience Seeking (ES) subscale, each of the SSS subscales has a psychometric structure that is in conformity with its theoretical structure. However, the psychometric properties of the SSS were shown to be inconsistent with its theoretical structure. From a psychometric point of view the results indicate that, in general, the SSS subscales each measure a single theoretical construct while the full scale does not. However, in an analysis of Wittgenstein's philosophy of psychology and the fundamental principles of construct validity it was shown that psychometric results are actually irrelevant to the issue of whether the SSS and its subscales are measures of the constructs for which they are named. Although it was shown that a total score based on any given set of items cannot be a measure of sensation

seeking proper, a conceptual analysis was performed and revealed that, in fact, none of the forty SSS items denote sensation seeking. In addition, it was shown that the justification for measurement claims given by operationism, the act frequency approach and axiomatic measurement theory is flawed. Since the vast majority of psychological tests have been constructed in accordance with the principles of construct validity, operationism, the act frequency approach and axiomatic measurement theory, it is concluded that in the vast majority of psychological research that employs psychological tests there is no justification for the claim that the results are relevant to the concepts the tests are believed to measure.

Dedication

To my darling wife Kim-Maru who has persevered through these difficult times.

Acknowledgments

I would like to take this opportunity to thank Dr. Roger Blackman for his support and guidance over the last five years of my graduate work. I appreciate very much his concern for my welfare and the work he has done on my behalf. Throughout my graduate career, he has loosened the reins when need be and pulled back on them when I have gone too far. I offer this thesis with respect, as an achievement that would not have been possible without his guidance.

To Dr. Michael Maraun, you have changed my life. I now strive to achieve a level of sophistication that will make you proud. To do justice to your supervision, I know that I must improve upon this work a great deal. You are a scholar in the true sense of the term and I consider every moment I have spent in your presence to be a blessing. Thank you for showing me what scholarship can be and teaching me academic principles and values.

Table of Contents

Table of Contents

List of Tables

1. Introduction

A survey of our indigenous measurement practices reveals few complexities or philosophical problems. Measurements are taken in accordance with rules that are commonly taught and explained. Such rules are readily surveyable and can be given in defence of particular measurement claims. If questioned, for example, to support a particular operation as a measurement of height, we expect that the claimant will point to certain rules. One might, for example, respond: "I used a metric tape measure, and stretched it in a straight line from the bottom of the object to the top." Such justifications are commonly given for a great variety of measurements. Measurements of speed, distance, weight, voltage, acceleration and time, for example, are all commonly taken and readily justifiable in terms of a public and shared system of rules. Although, in many cases (see for example Mach, 1960; Falmagne, 1992), the physical sciences have adopted more complex rule systems, their measurement practices remain logically linked with our own.

This, however, is manifestly not the case in psychology. While the discipline of psychology has developed numerous complex and technical measurement theories in an attempt to provide support for measurement claims about psychological phenomena, there are no indigenous measurement practices with which this theory can be linked. There is no indigenous public, surveyable or shared practice of psychological measurement - we simply do not speak of the units or instruments of measurement for psychological phenomena. We do not, for example, *teach* dominance measurement or *explain* how to take measurements of a person's dominance. It seems, therefore, that modern psychology faces a greater challenge than that of the physical sciences. It is a commonly held belief that while measurement operations in the physical sciences are known, psychology must develop them anew. Furthermore, since measurement is necessary for fruitful scientific investigation and since there are presently no established indigenous measurement practices for psychological phenomena, it is of fundamental importance to the advancement of the discipline that such practices be developed/discovered.

In this thesis, I will provide an example of the method by which a contemporary psychologist might address the measurement problem. The example is based on the well-known and widely used Sensation Seeking Scale (SSS) (Zuckerman, 1996). This analysis will focus on: (1) whether the statistical tests employed in the development of the SSS are appropriate for dichotomous items; (2) whether the statistical approach taken in previous sensation seeking (SS) research is justified by the nature of the constructs measured by the SSS full scale and subscales; and (3) whether the psychometric properties of the SSS are in conformity with the scoring rules employed in SS research. This analysis will provide a basis for determining whether there are psychometrically justified grounds for the claim that the SSS is an adequate measure of SS.

The purpose of the psychometric analysis is to illustrate methods and techniques that are used to support measurement claims in contemporary psychology. I will show, however, that the application of these methods (and some other major approaches to measurement in modern psychology) to the measurement problem rests on a misunderstanding. In contrast to the principles upon which the illustration is based, it is shown that measurement claims cannot be substantiated by empirical investigations of any kind. In particular, it is argued that, with the exception of operationism, the justification of measurement claims by empirical means is a fatal flaw shared by the major measurement traditions in modern psychology. The correct method upon which to justify measurement claims in any scientific context is, in fact, grounded in the autonomous grammar of concepts. Although operationism properly recognizes the importance of this issue to proper scientific practice, it provides a solution that divorces scientific investigations of technical, operationalized concepts from those in which we are most interested. For while there is clarity and precision in the rules of application for operational constructs, they are ultimately irrelevant to the common-or-garden psychological phenomena that we wish to understand. Finally, I show that the absence of a standard set of indigenous rules for the measurement of psychological concepts renders incoherent claims to the effect that psychological concepts can be measured. In so doing, the answer to the long pondered riddle of psychological measurement is revealed.

2. THE PSYCHOMETRIC PROPERTIES OF THE SSS

i. The Definition of SS

No decision regarding the correct empirical structure of an item set can be made without at least some prior explication of the concept of interest. For example, an item set that is expected to tap a unitary concept should have a structure that is distinct from that of an item set that measures a multifaceted concept (McDonald, 1981). Spearman, for example, expected a set of items designed to measure intelligence to be adequately characterized by a one-dimensional linear factor analysis model (Spearman, 1904), while Thurstone expected the same set to be described by a multidimensional linear factor analysis model (Thurstone, 1947). Their expectations differed because they viewed intelligence in different ways. Spearmen felt that intelligence was a unitary phenomenon while Thurstone believed that it was multifaceted. In the same way, our expectations concerning the structure of the SSS items must be predicated on a clear definition of the concept of SS.

In the opening sentence of the first article on SS Zuckerman, Kolin, Price and Zoob (1964) state; "This article reports the development of a Sensation-Seeking Scale (SSS) designed to quantify the construct: 'optimal stimulation level'." (p. 477). Unfortunately, Zuckerman et al do not provide any further explication of the concept. However, two comments in this early paper hint at what their view of the concept might have been. They state: "While it is possible that sensation seeking is specific to the various types of sensations, we hypothesized that a general factor would emerge from responses to diverse items" (p. 477). In concluding remarks Zuckerman et al present a similar position when they refer to SS as a unitary phenomenon that underlies a broad range of behavior. They state: "The SSS was developed to sample a broad SS tendency" (p. 481). Unfortunately, the first quote is rather ambiguous because it conflates an empirical hypothesis with a definitional/conceptual question. The notion that a general factor might appear in an analysis of a particular item set is not a description of the *meaning* of the concept of SS, but rather is an empirical hypothesis regarding the outcome of a linear factor analysis of the SS items. However, the concluding remark does indicate that the concept of SS should apply in a general manner and to a broad range of behavior.

The original 54 items chosen for Form I of the SSS consisted of 14 preference for extremes of sensation items, 12 items relating to dangerous activities, 2 items relating to a need for general excitement, 4 items relating to a need for adventure, 8 items expressing preferences for the new and unfamiliar, 8 preference for irregularity items, and 6 items expressing a desire for exciting (as opposed to reliable and predictable) friends. The sole insight into the grounds for item selection that Zuckerman et al (1964) provide is that "Items were written, using the construct as a guide" (p. 477). In later work with Form III and Form IV of the SSS, 63 new items were included (Zuckerman, 1971). The basis for inclusion of the items provides some insight into a modified view of the concept of SS. In this later paper Zuckerman adds to his previous description of SS. He states: "It was postulated that the need for change, variety, and intensity of stimulation would manifest itself in many aspects of behavior" (p. 45). Once again, this indicates that sensation seeking was viewed as an *overall* need for change, variety and intensity of stimulation. The concept is general or unitary in the sense that the need is not particular to certain behaviors/sensations but can be expressed in a wide variety of situations. However, subsequent empirical work led Zuckerman to overturn this original view of the concept of SS.

Farley (1967), reported a study in which a factor analysis of the original item set revealed four interpretable dimensions. He named them thrill seeking, social sensation seeking, visual sensation seeking and antisocial sensation seeking. Zuckerman also conducted his own study in which he attempted to replicate the factor structure uncovered by Farley (Zuckerman & Link, 1968). Two problems that emerged in these studies prompted Zuckerman to include 63 new items. Since only the first two factors were definable in females and the original set did not include enough items to clearly define more than one factor, Zuckerman (1971) set out "to write new items in an attempt to define the dimensions of sensation seeking" (p. 45). It is interesting to note that Zuckerman expects definition to *follow* item selection. One wonders how it is logically possible to conceive items relevant to a concept without a prior definition of the concept. If the items are to define the concept, how does one choose the items? With no prior definitional criteria, the domain of relevant items is infinitely large and completely unrestricted.

In a subsequent comment, however, Zuckerman admits to some prior criteria for item selection. He states: "Additional forced-choice items were written on the basis of the factors suggested by the preliminary results in the study by Zuckerman and Link (1968)" (p. 45). In addition to the dimensions uncovered in the factor analyses, "sexual sensation seeking" items were also included (Zuckerman, 1971).

Subsequent factor analyses of the revised item set (Zuckerman, 1971; Zuckerman, Eysenck and Eysenck, 1978) led to the current definition of four SS "factors." Zuckerman, et al (1978) defined these four factors as thrill and adventure seeking (TAS), experience seeking (ES), disinhibition (DIS) and boredom susceptibility (BS).

Although Form V of the SSS does not include a general factor, Zuckerman, et al (1978) suggest that a total SS score could be taken by summing the number of endorsed items. In his first book on the subject, Zuckerman (1979) also provided a definition of SS: "Sensation seeking is a trait defined by the need for varied, novel, and complex sensations and experiences and the willingness to take physical and social risks for the sake of such experience" (p. 10).

Form V of the SSS is the most widely used instrument for the measurement of SS, DIS, ES, TAS and BS. It consists of 40 forced-choice items, each of which belongs to one of four ten-item subscales. Total scores are obtained by taking an unweighted sum of the number of endorsed items on each of the SS subscales.

As previously mentioned, the definition of SS is important because it determines what we should expect to find in a psychometric analysis of the SSS items. It is crucial to note that without prior expectations, there is no criterion for the adequacy of the obtained psychometric results. For example, the fact that a single factor, g, should be responsible for performance on a diverse set of tests requiring cognitive ability led Spearman to expect certain patterns of relationships between the diverse set of tests (Spearman, 1927). In such a case, a set of tests that could not be represented by a single common factor model would not constitute an adequate measure of g.

As we have seen, a concern for definitional issues does not characterize Zuckerman's early work on SS. Although the initial set of items was written with a unitary concept of SS as a guide, later items were predicated on a four-dimensional view of the concept. However, the four-dimensional view was predicated on an empirical result obtained in the factor analyses of the original item set. There is no definitional guide to the logical connections between the four dimensions or between each of the four dimensions and the unitary SS concept. We, therefore, can not form any prior expectations about the dimensionality of the SSS items on definitional/conceptual grounds.

The change from a unitary view to a four-dimensional view also brings into question the original *unitary* concept of sensation seeking. Although Zuckerman continues to offer definitions of the unitary SS concept, the empirical analyses point to a multifaceted SS concept. One wonders why a commitment to a unitary concept is required if the empirical results are not supportive of such a conclusion. The practice of testing the original unitary "hypothesis" with empirical factor analyses, indicates that only a provisional commitment to a unitary SS concept is required. What then is the reason to maintain the original unitary concept of SS in the face of the apparently unsupportive results? Since there have been no definitional grounds provided for the unitary and four-dimensional views, there is apparently no justification for maintaining *both* a unitary and a multifaceted concept. This means that since the latest empirical results reported by Zuckerman suggest that SS is four-dimensional and since Zuckerman failed to provide conceptual links between the original unitary SS concept and the subscale concepts, the series of SS definitions given by Zuckerman have been undermined by his own research. Although the ambiguity surrounding the conceptual foundation upon which the SSS is built precludes a definitionally based determination of the expected psychometric properties of the SSS items, the scoring rules used in the employment of the SSS do place restrictions on them.

ii. Scoring Rules

A scoring rule provides the set of operations that are required in order to convert p item scores into a single total score. In general, the scoring rule for psychological scales is an unweighted sum of the item scores (Thissen, Steinberg, Pyszczynski and Greenberg, 1983). In the case of the SSS, items endorsed in the affirmative are assigned a value of 1 and items endorsed in the opposite direction of the scale construct are assigned a value of 0. Total scores for the TAS, ES, DIS and BS subscales are obtained by taking an unweighted sum of the item scores for the ten items that comprise each of the four subscales. The SS score is obtained by taking an unweighted sum of the four subscale scores.

These scoring rules place restrictions on the psychometric properties of the SS, TAS, ES, DIS and BS scales. This is because the justification for taking the sum of a set of items as a measure of a construct is merely that all the items measure the same construct. Therefore, in taking sums to obtain total scores, Zuckerman commits to the view that each of the 10 items on the four subscales measure their respective unitary subscale constructs and that all 40 of the SSS items measure the unitary construct of SS.

The fact that the scoring rule for the SSS and its subscales specifies *unweighted* sums places one further restriction on their expected psychometric properties. Since each item on the full scale and each of the subscales contributes an equal amount to its respective total scale score, we must assume that each item contributes an equivalent amount to the measurement of each of the scale constructs. Consider for example two individuals who endorse only a single but different item on the TAS scale. Under the unweighted sum scoring rule, each individual receives a total score of 1 on the TAS scale. Since, their TAS total score does not discriminate between them, they are deemed to possess equivalent amounts of TAS. It follows then that since they possess equivalent amounts of TAS and since each item is a measure of a unitary TAS phenomenon, each item must contribute an equal amount to the measurement of TAS.

As mentioned above, an unweighted sum scoring rule has consequences for the psychometric properties of a scale. If it is assumed that item scores are a function of a latent variable, θ, it can be shown that the unweighted sum of the X_j items is proportional to θ only if the regression parameters of each of the p items are equal. Thissen, Steinberg, Pyszczynski and Greenberg (1983) show that for a one-dimensional linear factor analysis model:

$$X_{ij} = \mu_j + \lambda_j\theta_i + \varepsilon_{ij}$$

where x_{ij} is the score of an individual i on item j, μ_j is the mean of item j, λ_j is the loading (regression parameter) of item j on the factor score of person i, θ_i, and ε_{ij} is a normal error with mean 0 and variance σ_j^2, the unweighted sum of items is not proportional to the maximum likelihood estimate for θ unless $\lambda_j = c$ for all p items. In the development of the SSS Zuckerman employed linear factor analysis to identify suitable items. A total score on the scale was based on an unweighted sum of the identified items. However, item regressions were not subject to equality constraints. An unweighted sum of the SSS items, is not, therefore, an optimal measure of the SS latent variable in the sense that it is not the maximum likelihood estimate for θ. An optimal total score for a one-dimensional factor analysis model with unequal λ_j is a weighted average of item responses with weights given by:

$$\beta_j = (\lambda_j \; / \; \sigma_j^2 \;)/(\Sigma(\lambda_j^2 \; / \; \sigma_j^2 \;) + 1)$$

where σ_j^2 is the unique variance for item j (Thissen, et al, 1983).

iii. Subscale Relationships

Since the SS subscale items are subsets of the 40 SSS items, it is important to specify the nature of the relationship between the 4 sets of subscale items. Without prior definitional criteria, it is difficult to form expectations concerning the subscale relationships. However, Zuckerman (1994) provides a reason for the construction of Form V of the SSS that indicates what he expected the subscale relationships to be. He states: "we felt that some of the correlations between the subscales were too high" (p. 32) and "we wanted to reduce the correlations between subscales, although we still expected enough correlation remaining to justify a total score" (p. 32). This suggests that Zuckerman expected the four subscales to be correlated. In fact, a criterion for the success of the construction of Form V was that there were medium correlations between the subscales.

Although medium correlations can be expected between subscales of a test, Zuckerman's belief that they are required in order to justify a total score is not correct. In fact, the above quote illustrates that the two distinct traditions of classical test theory and latent trait theory were indiscriminately applied in the development of the SSS. The view that large correlations indicate that the subscales measure the same thing is a fundamental premise of classical test theory (Thissen, et al, 1983). However, in latent trait theory the identification of the SS, TAS, DIS, BS and ES latent variables with the technique of linear factor analysis is based on the view that *unidimensionality* (not large correlations) of an item set indicates that the items measure the same thing (McDonald, 1981). It is well known, however, that medium or even large correlations between items or tests does not indicate that the set is unidimensional. What determines unidimensionality is the *pattern* of correlations between tests, not their absolute size.

The notion of a hierarchy of correlations was central in Spearman's development of common factor theory (Guttman, 1954). He noticed that if a common factor, $\boldsymbol{\theta}$, exists that when partialled out of each pair of items or tests reduces their partial correlations to 0, the observed correlations could be ordered so that for any pair of items or tests, x_i and x_j $i \neq j$, $r_{ij} > r_{lk}$ for $i < j$, $l \geq i$ and $k \geq j$. This condition holds in the correlation matrix given in Table 1. Since the variables in the matrix displayed in Table 1 are ordered so that their correlations decrease from left to

---

Insert Table 1 About Here

---

right and from top to bottom, they can be represented by a single common factor model. However, contrary to Zuckerman's belief, the tests have small intercorrelations. It follows then that Zuckerman's expectation of medium subscale intercorrelations does not imply a particular scoring rule for the subscales. However, although he is not justified in doing so, Zuckerman does recommend summing subscale scores in order to produce a full scale score. As we have seen, such a sum is justified if the four subscales are equal measures of a unitary underlying latent variable.

The implications of these scoring rules for the SSS and the four subscales are as follows:

1) SS is a unitary concept that is measured equally by each of the 40 SSS items.

2) SS is measured equally by each of the four SS subscales.

3) TAS, ES, DIS, BS are unitary concepts that are measured equally by each of the items on their respective subscales.

We are now in a position to form expectations concerning the psychometric properties of the SSS items.

iv. Measuring Unitary Concepts

Classical test theory offers inter-item correlations, item-total score correlations and various reliability coefficients to assess the measurement properties of item sets. Inter-item correlations and Cronbach's alpha, for example, are often used to assess the homogeneity of a set of items (McDonald, 1981). Large inter-item correlations or a large Cronbach's alpha are typically taken as an indication that the items measure a single construct. Similarly, large item-total score correlations are also taken as an indication that the items measure a single construct. In general, classical test theory interprets large correlations between items or between items and a total score as indicating that the items/test measure the same thing.

In contrast to classical test theory, the latent variable tradition assumes that items that measure the same thing should fit some model in the class of one-dimensional latent variable models. The latent variable model approach is recognized as an improvement over classical test theory because it allows for the specification of a statistical model in which item responses are a function of the latent variable the test is presumed to measure (Thissen, et al, 1983). By using a statistical model, the researcher can specify item parameters and apply statistical tests of the correspondence between the model and the observed data. In addition, this framework has the advantage of modelling item responses as if they arise from a *single* source of variance. This is an improvement over classical test theory because highly correlated item responses may arise from more than one source of variance. Such cases are not welcomed by test constructors because it is believed to be rather difficult to defend the position that a set of items measures a unitary concept when responses to the items arise form more than one source of variance. Latent variable theory is generally known as the set of techniques designed to address the problem of identifying underlying sources of variance responsible for test performance. Item response theory addresses the same issue for dichotomous items.

v. Item Response Theory

It is common practice in the application of item response theory to construct measures of a unitary construct in such a way that they fit at least one of a class of monotone one-dimensional latent variable models (Holland, 1981). For the purpose of the following analysis, I will adopt this convention. However, in the second section of this manuscript, arguments against the validity of this approach will be presented. A reasonable definition of unidimensionality exists for monotone latent variable models. This definition centres on the criterion of latent conditional independence (Holland & Rosenbaum, 1986). A set of items is said to satisfy the condition of latent conditional independence (be of dimension r) if the item residuals about their monotone regression on r latent variables are uncorrelated. If a *single* latent variable can be found that when partialled out of the item responses of the examinees renders their partialled item responses uncorrelated, the item set is defined to be unidimensional.

Unidimensional latent variable models differ with respect to the shape of the regression of the item responses on the latent variable and the shape of the distribution of the latent variable (Holland, 1981). In common factor analysis, for example, the regression of item responses on the latent variable is linear and the distribution of the latent variable is normal. A major class of latent variable models are those that apply when responses to items are in dichotomous form. These are widely known as item response models. Item response models are defined as follows:

$$P(\underline{X} = \underline{x}_*) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{j=1}^{p} P(X_j = 1 \mid \underline{\theta})^{x_{j*}} [1 - P(X_j = 1 \mid \underline{\theta})]^{1-x_{j*}} \, dF(\underline{\theta}) \qquad (1)$$

where $P(X_j=1|\underline{\boldsymbol{\theta}})$ is the regression of item j on the m < p dimensional random vector of latent variables (i.e., the item characteristic surface), $F(\underline{\boldsymbol{\theta}})$ is the p-dimensional distribution of $\underline{\boldsymbol{\theta}}$ and x_{j*} is a particular vector of dichotomous item responses. Item response models defined in

1 are unidimensional when $\underline{\boldsymbol{\theta}}$ is a scalar. In such cases the regression of item j on the latent variable $\boldsymbol{\theta}$ is a curve defined by $P(X_j=1|\boldsymbol{\theta})$. In the monotone variety of unidimensional item response models, $P(X_j=1|\boldsymbol{\theta})$ is nondecreasing with $\boldsymbol{\theta}$ and $F(\boldsymbol{\theta})$ is arbitrary. Models within this class differ on the shape of their item latent-variable regressions and on the distribution of the latent variable $\boldsymbol{\theta}$. These models define the candidate class for an acceptable characterization of the SSS, TAS, ES, DIS and BS scales for the following reasons:

1) They model item responses as if they arise from responses to items selected to measure a unitary concept.

2) They allow for any monotone item latent-variable relationship.

Although we have explored the justification for the first of these reasons, the justification for the second has not been addressed. In constructing a latent variable model for responses to a set of items that measure a unitary concept, it is necessary to specify the *form* of the relationship between the probability of endorsing an item and the value of the latent variable. This requirement amounts to a specification of the shape of the item characteristic curves $P(X_j=1|\boldsymbol{\theta})$. The problem is to determine what sorts of relationship between the probability of endorsing an item and the latent variable are logically admissible given the nature of the construct the items measure. In order to address this question, it is necessary to examine the role of latent variables in the measurement context.

vi. Latent Variables

Mathematically, latent variables are algebraic constructions that "account for" correlations between a set of items or tests. If $X = (X_1, \ldots, X_j)$ is a set of observed variables, θ (an "unobserved" latent variable) is said to account for or explain the correlations between the manifest variables in X if, given a particular value of θ, there is no correlation between the observed variables. The purpose of latent variable models is to represent performance of an examinee on a test in terms of an underlying unobserved latent trait or ability. The obtained mathematical construction is taken to represent an unobserved ability or trait that an individual possesses in some degree. If a latent variable is found that accounts for the correlations between a set of items, it is assumed that test takers possess a latent trait that is responsible for their performance on the items. Latent variable scores are taken as a measure of the amount of the latent trait possessed by any given test taker. In this framework, $P(X_j=1|\theta)$ represents the regression of the probability an examinee will endorse a given item on the latent variable.

In the case of the SS, TAS, BS, ES and DIS, the scales are named for the latent trait they are believed to measure. High scores on any of these scales are taken as an indication that the examinee has a high level of the respective latent trait, and low scores are taken to represent low levels of the respective latent trait. It follows then that the more items an individual endorses, the higher their value of the latent trait is believed to be. If it is assumed that the probability of endorsing any single dichotomous item increases with the total scale score, then it follows that $P(x_j=1)$ should increase with θ.

It has been argued that the term "latent variable" does not have a consistent or clear use in the psychometric or psychological literature (Maraun, 1996a). In order to address this issue it is necessary to distinguish between the technical concept of latent variable and the various senses of latent variable employed in the practice of latent variable modelling. A latent variable, θ, is a variable constructed in accordance with the mathematical equations of a particular latent variable model. For example, in the one-dimensional common factor analysis model:

$$\underline{X} = \underline{\Lambda}\theta + \underline{\sigma}^2 \qquad (2)$$

where $\underline{X}$ is a vector of p random variates, θ is the latent common factor, $\underline{\sigma}^2$ is a vector of p unique common factors and $\underline{\Lambda}$ is a px1 vector of factor loadings, the latent variable θ is constructed according to the mathematical rule:

$$\theta_i = \underline{\Lambda}\ \Sigma^{-1}\ \underline{X} + ps_i \qquad (3)$$

where Σ^{-1} is the inverse of the covariance matrix of $\underline{X}$, p = $(1 - \underline{\Lambda}'\Sigma^{-1}\underline{\Lambda})$ and $E(ps_i\underline{X}') = \underline{0}$, $E(s_i^2) = 1$ and $E(s_i) = \underline{0}$. The technical sense of latent common factor is given by the components of the equations (2) and (3) (Maraun, 1996a). A common factor is a latent variable if it satisfies the constraints in (2) and (3). Since, in the technical sense, what is meant by latent common factor is given by (2) and (3), the criteria of application for the term latent common factor is *internal* to the latent variable model specified in (2).

This technical sense of latent variable must be distinguished from the other senses of the term employed in the psychometric and psychological literature. These other senses are distinguished from the technical notion of latent variable in that they all appeal to criteria of application that are *external* to a latent variable model. For example "latent traits," "unobserved causal factors," "underlying abilities" and "dimensions of personality" are entities that are denoted by criteria that are external to the construction rules of a particular latent variable model (Maraun, 1996a). While it may be true that underlying or directly unobservable variables are "tapped," "picked up" or "detected" by latent variable models, *they* are not the variables defined by the construction rules of a latent variable model. Talk of latent variables as underlying, unobserved, etc. is, therefore, based on a conflation of the technical sense of latent variable and the senses of latent variables that are external to a latent variable model. In the following, I recognize this distinction by denoting the technical sense of latent variable by the symbol θ and external senses of latent variable by the symbol Ψ.

What then should the restrictions be on the form of the relationship between a particular latent variable and the items chosen to measure the underlying psychological construct the latent variable is assumed to represent? The solution to this problem is given by the contents of the bracket in the following statement:

Item X is a measure of concept Ψ if the regression of X on θ is --> {...........} (4)

To my knowledge, classical test theory and latent variable theory do not explicitly address issues pertaining to the contents of the bracket in (4). However, it is generally held that unless the regression of X on θ is at least monotone increasing, X cannot reasonably be advanced as a measure of Ψ (Holland, 1981; McDonald, 1981). In the case of the dichotomous items of the SS, TAS, ES, DIS and BS scales, this restriction amounts to the requirement that the $P(x_j=1)$ should increase monotonically with θ. A latent variable model representation of the relationship between Ψ

and X must, therefore, model the relationship between $P(x_j=1)$ and θ as monotone increasing.

vii. Monotone Latent Variable Models For Dichotomous Items

Holland (1981) has derived conditions that must be satisfied in order for a set of dichotomous items to be consistent with any model in the class of unidimensional monotone (increasing) latent variable models. Let X denote a vector of dichotomous responses on some test, T, of an examinee sampled at random from a population C. Let (Y,Z) denote a partition of X into complementary subsets of items and A denote a subset of items in Y. If $Y_A = 1$ denotes the event that $Y_j = 1$ for all $j \in A$, and $Y_{A+A'}$ denotes the event that $Y_j = 1$ for all $j \in$ A and A', then a unidimensional monotone latent variable (UMLVM) representation of X can be found if:

1) $P(Y_{A+A'} = 1|Z) \geq P(Y_A = 1|Z)P(Y_{A'}=1|Z)$

2) $P(Y_{B+B'} = 0|Z) \geq P(Y_B = 0|Z)P(Y_{B'}=0|Z)$

3) $P(Y_A = 1, Y_B=0|Z) \leq P(Y_A = 1|Z)P(Y_B = 0|Z)$

Condition 1 states that for any pattern of responses on the rest of the test, Z, the probability that an examinee will get all of one part of a test (A) correct increases if they are known to have perfect performance on another disjoint part of the test (A'). Condition 2 states that for any pattern of responses on the rest of the test, Z, the probability that an examinee will get all on one part of the test (B) incorrect, increases if they are known to have responded incorrectly to all the items in a disjoint part of the test, (B'). Finally, Condition 3 states that for any pattern of responses on, Z, knowing that an examinee responded incorrectly to one part of the test (B) decreases the probability that they will perform perfectly on a disjoint part of the test (A).

Unfortunately, it is not a simple matter to determine whether or not these conditions have been met. Essentially, the task is to form contingency tables containing the proportion of respondents with particular response vectors given a particular response vector on the other part of the test, Z. The problem faced in testing Holland's conditions is that as the number of items on the test increases, the number of contingency tables required increases exponentially. Consider a test of these conditions for just one of the 10-item SSS subscales. Let us first consider the number of response vectors that must be conditioned upon; that is, all possible response vectors that belong to Z. Conditioning on the first of the 10 items results in Z taking on two possible values, i.e., Z = {1} and Z = {0}. Conditioning on the first two of the ten items results in Z taking on the four possible values of {1,0}, {1,1}, {0,1} and {0,0}. In general the number of values for Z are 2^k, where k is the number of items in Z. However, the number of combinations of vectors of size, k, are n choose k or n!/[k!(n-k)!]. For k=2 this gives us (10!/[2!(8!)])*4 = 45*4 = 180 possible vectors of size 2 to condition upon. Therefore, for a 10-item subscale, there are:

(10!/[1!(9!)]*2^1 + (10!/[2!(8!)]*2^2 + ..+ (10!/[8!(2!)]2^{10}

= 20 + 45*4 + 120*8 + 210*16 + 252*32 + 210*64 + 120*128 + 45*256 = 46,504 subsets to condition upon. Now, for cases in which conditioning is on a single item, for each of the 20 possible vectors in Z the disjoint subsets of A and A' (or B and B') could contain 2, 3, 4, 5, 6, 7, 8, or 9 items. There are 10 choose 2 = 45 sets of possible A and A' vectors when A and A' consist of 2 items, 10 choose 3 = 120 sets of possible A and A' vectors when A and A' consist of 3 items and so on. Therefore, the number of contingency tables that can be formed when Z consists of a single item is: 45(20) + 120(20) + 210(20) + 252(20) + 210(20) + 120(20) + 45(20) + 10(20) = 20,240. Similarly, the number of contingency tables that can be formed when Z consists of two items are: 45(180) + 120(180) + .. + 45(180) = 180,360. The number of contingency tables that can be constructed in total, therefore, is: 45(1012) + 180(1002) + 960(957) + 3,360(837) + 8,064(627) + 13,440(375) + 15,360(165) + 11,520(45) = 17,081,528. The problem of constructing this many contingency tables is not a small one. However, more troubling is the problem of acquiring sufficient data to fill each of the cells in this many contingency tables.

Clearly, the conditions for monotone unidimensionality given by Holland (1981) are not testable for medium or large item sets. Certainly, sets consisting of 10 or more items are much too large. Fortunately, Holland and Rosenbaum (1986) have provided a test of monotone unidimensionality that is more easily carried out. They showed that if a set of dichotomous items is consistent with a unidimensional monotone latent variable model, the conditional covariances between all monotone increasing functions of a set of item responses, given any function of the remaining item responses, will be nonnegative. Two special cases of this condition are that every pair of items must have a nonnegative correlation (NNC) and that the correlation between any pair of items given any particular total score must also be nonnegative. The Mantel-Haenszel statistic is a conventional procedure that can be used to test the second of these cases (Zwick, 1986). In this case, the Mantel-Haenszel statistic examines the covariance between a pair of single item responses conditional upon the total score on the remaining items. Let n_{ijk} be the observed count in the ith row, jth column, and kth table, where i = 0,1, j = 0,1 and k = 1,2,...k. The Mantel-Haenszel test statistic is given by:

where

$$z = \frac{n_{11+} - E(n_{11+}) + 1/2}{\sqrt{V(n_{11+})}}$$

and

$$E(n_{11+}) = \sum_{k=1}^{k} \frac{n_{1+k}\, n_{+1k}}{n_{++k}}$$

$$v(n_{11+}) = \sum_{k=1}^{k} \frac{n_{1+k}\, n_{0+k}\, n_{+1k}\, n_{+0k}}{n_{++k}^{2}\,(n_{++k}-1)}$$

and the plus subscript indicates summation over the subscript from k=1 to k. Since the Mantel-Haenszel statistic has an approximately standard normal distribution, the approximate significance level is obtained by referring the calculated vale of z to the lower tail of the standard normal distribution. If $Z_{calc} < Z_{crit}$ for a given level of alpha, the hypothesis that the item pair has no partial correlation is rejected in favour of the alternate hypothesis that the items have a negative partial association. Since all conditional associations must be nonnegative, significant values of Z_{calc} indicate that the item set is not consistent with a UMLVM representation.

The Mantel-Haenszel statistic given above provides a direct assessment of the consistency of a set of dichotomous items with a UMLVM representation. This statistic, however, is not directly applicable to polychotomous or continuous items. Fortunately, Holland and Rosenbaum (1986) have derived a general theorem that places restrictions on the covariance matrix of any item set that can be described by a UMLVM. The theorem states that if a UMLVM representation can be found for a random vector X = $(x_i,..,x_j)$, the distribution of X is conditionally positively associated. A random vector, X, is conditionally positively associated if, for any partition (Y,Z) of X and any function h(Z), the conditional distribution of Y given h(Z) is positively associated. Finally, the distribution of a random vector, X, is positively associated if $Cov(f(x),g(x)) \geq 0$ for all nondecreasing, bounded functions, f(.) and g(.). This condition can be summarized as follows:

$$Cov(f(y),g(y)|h(z)) \geq 0$$

for any partition of X (Y,Z) and any function of Z, h(z).

In the present case, there are four subscales, $X_1,...,X_4$, upon which subjects receive integer valued scores between 0 and 10. Since only four subscales exist, the partition of X = $(X_1,...,X_4)$ in to two disjoint subsets, y and Z, means that Z can consist of a maximum of two and a minimum of 0 variables. For the case in which Z is the empty set, the above condition requires that the covariance matrix of the subscale totals must contain nonnegative covariances (NNCV). For the case in which Z consists of a single variable, the above condition requires that given a particular total score on the subscale belonging to Z, the covariance matrix of the three remaining subscales must be NNCV. For example, for all subjects with a total score of 5 on the ES subscale, the covariances between their BS, TAS and DIS subscale total scores must be nonnegative.

The case in which Z consists of two variables is a little more complicated. The above condition states that for any function of the two subscale totals belonging to Z, the covariance between the remaining two items must be nonnegative. For example, for all subjects with a total score of 10 on the ES and BS subscale, the covariance between their TAS and DIS subscale total scores must be nonnegative. These tests can be conducted by forming conditional (on Z) covariance matrices, and for each covariance in the matrix testing the hypothesis: Ho: Cov(X,Y) = 0. This hypothesis can be tested with the following statistic:

$$t_{N-2} = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

Since the t statistic has a t_{N-2} distribution, the approximate significance level is obtained by referring the calculated vale of t to the lower tail of the t_{N-2} distribution. If $t_{calc} < t_{crit}$ for a given level of alpha, the hypothesis that the item pair has no covariance is rejected in favour of the alternate hypothesis that the items have a negative partial association. Since all conditional covariances must be nonnegative, significant values of t_{calc} indicate that the item set is not consistent with a UMLVM representation.

viii Analytic Procedure

Three necessary conditions for the psychometric justification of the scoring rules employed in the SSS and its subscales have been identified. First, the 40 items of the SSS should exhibit a covariance structure that is consistent with a unidimensional monotone latent variable model. Second, each of the 10 item subscales of the SSS should exhibit a covariance structure that is consistent with a UMLVM. Finally, the four subscales of the SSS should be consistent with a UMLVM. The first two conditions can be tested by assessing NNC and, given acceptance of NNC, tested with the Mantel-Haenszel statistic previously described. The final condition can be assessed by testing for nonnegativity in the subscale total score conditional correlation matrix with the t statistic described earlier. In the present study, an initial test for nonnegativity of the subscale total score conditional correlation matrix will be conducted with conditioning on the empty set. Given acceptance of nonegative manifold with conditioning on the empty set, a test of NNC of the subscale total score conditional correlation matrix will be conducted with conditioning on the subscale total score.

iX Previous Analyses of The SS Items

Table 2 provides a summary of previous psychometric analyses of the SSS performed on Caucasian samples. Previous research conducted on non-English speaking samples will not be considered because there is no clear equivalence between the translated items they responded to and the original English item set. The early development (Studies 1

Insert Table 2 About Here

- 7) of the SSS is characterized by an attempt to choose a particular number of factors and the items that load on them. The general analytic strategy has been to extract factors from a correlation matrix with communalities on the diagonals, attain simple structure by conducting oblique rotations of the extracted factors, and select items to measure each factor on the basis of the size of their factor loadings. Curiously, there appears to be some uncertainty about what particular correlation matrix to analyze. In early studies, tetrachoric correlations were factor analyzed, while in later studies either no mention was made of the matrix analyzed or a Pearson product-moment correlation matrix was factored. Later studies have factored the 40 items of Form V with maximum likelihood factor analysis in order to determine whether or not the subscale items load as they should on their respective subscales. In the following section I will show that all previous analyses of the SSS have applied models that are too restrictive and that most have factored the wrong correlation matrix.

x Dimensional Analysis of Dichotomous Items

The common factor analysis model given in (1) assumes that there is a single common factor, $\boldsymbol{\theta}$, and that:

$$E(X_j \mid \boldsymbol{\theta}=\boldsymbol{\theta}_i) = \boldsymbol{\mu}_j + \boldsymbol{\lambda}_j\boldsymbol{\theta}_i$$

where $E(X_j \mid \boldsymbol{\theta}=\boldsymbol{\theta}_i)$ is the population mean score on item j given $\boldsymbol{\theta}=\boldsymbol{\theta}_i$, and $\boldsymbol{\lambda}_j$ is the loading of item j on the common factor. This assumption specifies that the relationship between the latent variable $\boldsymbol{\theta}$ and the item response X_j, is linear with item mean $\boldsymbol{\mu}_j$, and item slope $\boldsymbol{\lambda}_j$. In calculating the Pearson product-moment correlation (phi) between dichotomous items, it is assumed that the item responses are true dichotomies (there is no continuous variable underlying the dichotomous responses). The problem for the application of the common factor model to dichotomous item responses is that the relationship between a continuous latent variable and a dichotomous variable cannot be linear (Zwick, 1986). Since the linear factor analysis model assumes a linear latent variable, item regression, it is not an appropriate model for dichotomous items. If we assume that a phi matrix was actually analyzed in the unknown cases listed in Table 2, the analytical model was incorrectly specified in Studies 4, 5, 6, 7, 9 and 10. Since the first three studies were conducted on Form 1 and the eighth study did not perform any latent variable analyses, it is reasonable to conclude that no previous attempts to assess the dimensionality of the SSS items have been conducted correctly.

The consequence of this problem is that the conclusion of four dimensionality reached in the *empirical* analyses of the SSS items cannot be justified. Since no definitional grounds have been given for the dimensional structure of SS, there is, therefore, no sound justification for the claim that SS is four dimensional. If the conclusion of four dimensionality cannot be justified, there is no justification for taking total scores on the four Form V subscales. This means that there are no psychometric grounds for the claim that the Form V subscale scores are measures of anything meaningful. It follows, therefore, that there is no sound justification for the claim that previous empirical research in which the SS subscales are employed is meaningful.

Strong assumptions are also required in a linear factor analysis of tetrachoric correlations. In computing tetrachorics it is assumed that the probability of endorsing an item is a function of an underlying continuous variable.

The model asserts that individuals with a value on the underlying continuous variable that is greater than some threshold endorse that item, and that individuals with values lower than the threshold do not endorse the item. Under the assumption of bivariate normality, the correlation between any two unobserved continuous variables can be inferred from their 2x2 contingency table. If bivariate normality does not hold, tetrachoric correlations do not provide a valid measure of association. Since the input to the factor analyses conducted in Studies 1, 2 and 3 was a correlation matrix of tetrachorics, bivariate normality must hold between the continuous variables that are assumed to underlie the 1/2(p(p-1)) pairs of items in order for the factor analyses to provide valid results. Since the underlying continuous variables are not observable, it is not possible to test whether or not bivariate normality holds.

Table 2 shows that in all previous attempts to assess the dimensionality of the SSS items the analytic procedure involved some form of linear factor analysis. Early studies incorporated methods in which principal components were extracted from correlation matrices with communality estimates entered on their diagonals. Later studies performed maximum likelihood factor analyses. Both types of analysis assume a linear relationship between item responses and the unobserved latent variable. However, in this development of the definitional basis of SS and the scoring rules employed in the SSS, it has been shown that they place only *monotone* restrictions on the relationship between item responses and the unobserved latent variable. Therefore, all previous analyses of the SSS have employed models that place unnecessary linear restrictions on the item latent variable regressions. Such restrictions have implications for conclusions concerning the dimensionality of item sets and the items that load significantly on the chosen dimensions.

Further exploration of this point is necessary because claims to the dimensionality of SS have been justified on the basis of the number of dimensions uncovered in linear factor analyses of the SSS items. Since definitional criteria were not employed in the development of the SSS, the linear factor analyses provide the sole justification for Zuckerman's claim the SSS is four-dimensional. The problem for Zuckerman is that the conclusion reached under linear latent variable, item response regressions is not necessarily the same conclusion that would be reached under monotone latent variable, item response regressions. In particular, if a linear model provides a solution of s dimensions, a monotone model will require $p \leq s$ dimensions to represent the same inter-item covariance structure (McDonald, 1981; Lingoes & Guttman, 1967). In the following section of this manuscript, a psychometric investigation of the SSS items will be undertaken in order to determine whether a four-dimensional conclusion is warranted.

xi Subjects

Subjects were 927 British Columbia high school students who responded to a student lifestyle questionnaire. The SSS was included as one part of the survey. With only 5 exceptions, ages ranged from 14 to 19 years. The sample consists of 52.1 % males and has an average age of 16.08 years.

xii Results

Table 3 presents the frequency distributions of the responses to the 40 SSS items. Table 4 presents the means

----

Insert Table 3 About Here

----

and standard deviations of the SSS full scale and subscale

Insert Table 4 About Here

----

scores for the present sample and for the four samples studied by Zuckerman, Eysenck and Eysenck (1978). For each of the subscales and the full scale, both the means and standard deviations found in the present study are captured by the sample ranges found by Zuckerman, et al (1978). This result indicates that the response patterns of this sample are well within those found in previously studied samples.

Table 5 presents means, standard deviations and minimum and maximum values of the Pearson (phi) correlations

----

Insert Table 5 About Here

----

between the items of each of the four SSS subscales and the SSS full scale. In the case in which the population correlation (ρ) between each item pair is 0, the sampling distribution of the Pearson r is centred around 0 and is symmetrical. Therefore, if ρ is 0, approximately 50 percent of the sample inter-item correlations should be less than 0. A result in which less than 50 percent of the item correlations are less than 0 indicates that the average population inter-item correlation is greater than 0. In all cases, considerably less than half of the inter-item correlations were less than 0. This result supports the premise that the population inter-item correlation matrices for the four SSS subscales and the SSS full scale are NNC.

Stronger support for the conclusion of NNC is indicated in Table 5 by the proportion of sample inter-item correlations that fall below t_{crit} for α =.025. Specifically, if a sample inter-item correlation falls below t_{crit} it may be considered a candidate for rejection of the null hypothesis that ρ=0 in favour of the alternate hypothesis that $\rho < 0$. Negative inter-item population correlations are inconsistent with the conclusion of NNC and are, therefore, inconsistent with the conclusion that the item set can be described by a UMLVM. However, in the case in which $\rho=0$, we should expect α =.025 sample inter-item correlations to fall below t_{crit}. Since for each of the four SSS subscales no sample inter-item correlations fall below t_{crit}, we can conclude that no population inter-item correlations are negative. The results presented in Table 5, therefore, provide preliminary support for the conclusion that each of the four SSS subscales can be described by a UMLVM.

The results reported in Table 5 also provide preliminary support for the existence of a UMLVM characterization for the SSS full scale. Since the proportion of inter-item correlations less than t_{crit} ($P < t_{crit}$ = .028) is only slightly more than α =.025, it is reasonable to conclude that the observed significantly negative inter-item correlations are a result of sampling error. This result indicates that the population correlation matrix of SSS full scale items is NNC.

Insert Table 6 About Here

Table 6 presents Mantel-Haenszel (MH) statistics for the subscales and full scale. In general, the outcome of the MH tests is consistent with the inter-item correlation results reported in Table 5. Of the four subscales, only the ES scale had sample proportions of MH statistics (P_{mh}) exceeding expected sample proportions (P_E) under the hypothesis of a 0 population conditional covariance. In particular, $P_{MH} = 4.4 > P_E = .025$ and $P_{MH} = 17.8 > P_E = 14.0$. However, the observed values of P_{MH} are not considerably larger than the expected values of P_E.

The outcome of the MH results for the full scale items, however, was not consistent with the inter-item correlation results. In particular, 6.2 percent of the MH statistics exceeded $Z_{crit} = -2.33$, $\alpha = .001$ (1-tailed). Since under the assumption of a 0 population conditional covariance and for $\alpha = .001$, we would expect .1 percent of the MH statistics to exceed Z_{crit}, the observed proportion of MH statistics exceeding Z_{crit} is sixty times larger than expected. The MH results, therefore, suggest that the SS full scale cannot be described by a UMLVM.

As a basis for comparison, MH statistics were also calculated on a series of simulated item response patterns. Testfact statistical software (Wilson, Wood & Gibbons, 1991) was used to produce 927 simulated response patterns based on a normal ogive item response model. Normal ogive item response models relate values of a latent variable θ to the $P(X_j = 1 \mid \theta)$ as follows:

$$P(X_j = 1 \mid \theta) = \int_{-\infty}^{a(\theta - b)} f(z)\,dz \qquad (5)$$

where f(z) is the standard normal density function, a is the slope or regression parameter of the ICC defined in (5), and b is a difficulty parameter that gives the value of θ at which half the examinees endorse item j (Crocker & Algina, 1986). The slope of the ICC defined in (5) is related to the discriminability of an item. Items that discriminate sharply in the range of b, between subjects above and below a particular value of θ have large values of a (i.e., a > 1), and items that are ineffective at discriminating between subjects with different levels of θ have low values of a (i.e., a < .5). The regression of items for which a = 0 on θ is linear with no slope. The relationship between the regression parameter and the correlation between an item and the latent variable θ (ρ_j) is given by the following expression:

$$a_j = \frac{\rho_j}{\sqrt{1 - \rho_j^2}}$$

where ρ_j is the population biserial correlation between and item X_j and θ, and θ is distributed normally with mean 0 and variance 1. For the simulated response patterns generated in this study, a was fixed to some combination of the following values: -.9, -.7, -.5, -.3, -.1, 0, .1, .3, .5, .7 and .9. For these values of a, the item-latent variable correlations are 0 (a=0), .1 (a=.1), .29 (a=.3), .45 (a=.5), .58 (a=.7) and .66 (a=.9). Table 7 provides a description of the particular combinations of regression parameter values chosen for each of the models. Patterns of slopes were chosen in order to provide a basis of comparison

----

Insert Table 7 About Here

----

ranging from the theoretically correct model for the SSS scales to the theoretically most inappropriate. As previously discussed, the theoretical psychometric structure of the SSS scales implied by the SSS, TAS, DIS ES and BS definitions and scoring rules is, in all cases, unidimensional with equal, non zero, positive and preferably medium to large slopes. Models with constant slopes of .3 or greater are, therefore, "prototype" models for the SSS scales. Less appropriate, but perhaps marginal standards of comparison are the simulated response patterns based on models with low constant positive slopes and variable positive slopes. Two-dimensional models and models with variable positive and negative slopes are, however, in direct contradiction of the theoretical structure of the SSS scales.

Tables 8, 9, 10 and 11 include proportions of Mantel-Haenszel statistics calculated on the simulated response patterns exceeding Z = c for c = -4, -3, -2, -1, 0 and -2.33 (Z_{crit} for α = .001, 1-tailed). Also included are the observed MH statistics exceeding the chosen values of c and the expected number of MH statistics exceeding c given a zero population conditional covariance. In the following section, the observed and simulated proportions of MH statistics are compared.

Table 8 reports proportions calculated on response patterns of 40 items based on a one-dimensional normal ogive model with various slopes and item difficulties set to the observed difficulties reported in Table 3. Inspection of

----

Insert Table 8 About Here

Table 8 reveals that only three of the simulated models (item slopes: -.5-.5,-.7-.7 and -.9-.9) resulted in more rejections than the SSS full scale. Since these models all contain negative item latent variable regressions, they are not suitable candidates for the representation of the SS full scale items. This is simply because negative item latent variable regressions model the probability of endorsing an item as *decreasing* with increases in the latent variable. Such representations are clearly not in conformity with the summative scoring rule of the SSS full scale. Table 8, therefore, indicates that none of the models that are in conformity with the scoring rule of the SSS full scale resulted in as many rejections as were found for the SSS full scale. The simulation results, therefore, support the conclusion derived from the MH analyses that the SSS full scale cannot be described by a UMLVM.

Table 9 reports proportions calculated on response patterns of 40 items based on a two-dimensional normal ogive model with various slopes and item difficulties set to the observed difficulties reported in Table 3. The observed

----

Insert Table 9 About Here

----

pattern of rejections and proportions of MH statistics exceeding Z = c for c = -4, -3, -2, -1, 0 and -2.33 for the SSS full scale falls between the proportions generated from the simulated two-dimensional normal ogive models with slopes ranging from 0 to .3 and 0 to .5. For the simulated two-dimensional models with negative and positive slopes, the SSS full scale proportions fall between models with slopes ranging from -.1 to .1 and -.3 to .3. These results indicate that the SSS full scale observed response patterns are consistent with a two-dimensional monotone latent variable model representation.

Table 10 reports proportions calculated on response patterns of 10 items based on one-dimensional normal ogive models with various slopes and item difficulties set to the observed TAS scale difficulties. With the exception of

---

Insert Table 10 About Here

---

the ES subscale, the observed proportion of MH statistics is fully consistent with the unidimensional models presented in Table 10. These simulation results, therefore, support the conclusions derived from the inter-item correlation analyses and the MH analyses that the TAS, DIS and BS subscales can each be described by a UMLVM. The ES subscale proportions, however, are only consistent with the simulated results based on a unidimensional normal ogive model with slopes ranging between 0 and .1. This indicates that if the ES subscale is described by a UMLVM, the relationships between the ES items and the underlying latent variables they purport to measure are very weak.

On the basis of the simulations, however, it appears that the pattern of ES subscale proportions is consistent with a number of possible two-dimensional representations. Table 11 reports proportions calculated on response patterns of 10 items based on a two-dimensional normal ogive model with various slopes and item difficulties set to the observed TAS subscale difficulties. Four of the models

----

Insert Table 11 About Here

----

presented in Table 11 are consistent with the pattern of observed ES scale proportions. The simulations show that models with positive slopes ranging between 0 - .1 and 0 - .5 are possible candidates for a characterization of the ES subscale. In addition, a model with positive and negative slopes ranging between -.1 and .1 can be expected to produce the pattern of proportions observed for the ES subscale. The simulations, therefore, indicate that the ES subscale has a questionable status as a set of items described by a UMLVM.

Insert Table 12 About Here

Table 12 presents the Pearson correlations between the four subscales and the SSS full scale. Since the matrix of subscale correlations is clearly NNC, Table 12 provides support for a UMLVM representation of the subscale total scores. In a further test of the monotone unidimensionality of the subscale totals, a series of conditional correlations were calculated and tested for significance with the t statistic described earlier. Conditioning was on all possible total subscale scores for each of the four subscales. The outcomes are reported in Table 13. Of the 129 correlations reported in Table 13,

---

Insert Table 13 About Here

---

only 1 exceeded the value of t_{crit} for α = .01 (1-tailed). Since of the 129 calculated correlations, we would expect approximately 1 significant correlation under the hypothesis that the population conditional correlation is 0, it is reasonable to conclude that the population conditional correlation matrix of subscale total scores is NNC. This finding indicates that a UMLVM representation of the subscale total scores can be found.

3. Discussion

These analyses indicate that the psychometric structure of each of the SSS subscales, with the possible exception of the ES subscale, is in keeping with their theoretical structure. Specifically, it appears that the structure of the SSS subscales is in keeping with a UMLVM representation. However, the structure of the SSS full scale items does not appear to be clearly unidimensional in a monotone latent variable sense. Although it was found that the subscale totals may well have a UMLVM representation, it appears that the covariance structure of the full 40 item correlation matrix is not unidimensional. Is this good news for the SSS subscales and bad news for the SSS full scale?

From a psychometric point of view, these results indicate that the SSS subscales consist of items that all measure the same thing. Although the ES subscale could benefit from an item analysis and restructuring, in general, the statistical properties of the scale should satisfy proponents of the latent variable measurement tradition (McDonald, 1981). For the subscales, the psychometric news is good. The SSS full scale, however, is clearly in need of restructuring. Poorly performing items should be identified by consulting the inter-item correlation matrix and the Mantel-Haenszel item statistics. Given the relatively poor performance of the ES subscale, a sensible approach would be to target ES items.

For the moment, however, it is reasonable to ask the question: What are the implications of the poor psychometric performance of the SSS for research on SS? For the sake of argument, let us assume that the results presented above can be replicated and that a UMLVM representation of the SSS items is in keeping with their theoretical structure. Returning to our initial question, should it matter to the SS researcher that total scores on the SSS can be shown to arise from more than one source of variance? Should there be some question as to *what* is measured by the SSS if the psychometric results indicate that the scale measures at least two entities? Although there are numerous levels at which one might approach this question, the most fundamental centres on the nature of evidence one must provide in order to support a measurement claim. Once the grounds upon which the justification for measurement claims rest are fully articulated and in clear view, it should be a relatively simple matter to determine the implications of these results for research on SS.

In general, support for the claim that aspects of the covariance structure (including unidimensionality) of a set of items are relevant to measurement claims about those items is laid down in the principles of construct validity. In fact, if one were to ask a contemporary psychologist to support a particular measurement claim, it is most likely that the justifications given would be in the spirit of the principles of construct validity (see for example, Zuckerman, 1996). In the following section of this manuscript I argue that these principles contain an insurmountable logical flaw. In particular, the failure of the proponents and practitioners of construct validation to recognize the autonomous, grammatical nature of the justification for measurement claims results in the application of empirical methods to a problem that can only be addressed by conceptual/logical means. The result is a fatal conflation of discovery and meaning that renders incoherent the results of construct validation research. Consequently, it is concluded that the empirical, psychometric results produced above are actually irrelevant to research in which a measurement of SS is required. Since the psychometric procedures employed in this section and in previous attempts to establish the validity of the SSS are justified by a long history of validity research, the second section of this manuscript begins with an overview of the types of validity currently sought by test developers.

i The Validity of Psychological Tests: The Accepted View

As previously mentioned, the development of the SSS was based on the principles of both classical test theory and latent variable theory. Although each position takes a different stance on the sort of empirical evidence that is necessary to justify measurement claims, they both adhere strongly to the view that empirical evidence is necessary to the justification of the claim that an item/test measures a concept Ψ. In particular, both traditions maintain that the validity of a psychological test is given by the extent to which the score on a given test is in agreement with particular empirical expectations or hypotheses. In 1985 a joint committee of the American Psychological Association, the American Educational Research Association and the National Council on Measurement in Education published a manuscript in which the standards of acceptability for the validity of educational and psychological tests were given (American Psychological Association, 1985). The committee agreed upon three types of validity that were required in order to establish whether test scores exhibited sufficient agreement with the phenomena they were believed to measure. These types of validity were named content, criterion and construct validity.

ii Content Validity

Content validity is the only one of the three types of validity that is based on logical/definitional/conceptual criteria. A test has content validity if it consists of items that are a representative sample of the universe of items of interest to the test developer. The universe of items are to be chosen on the basis of their logical status as items that are relevant to the content domain of interest. This requires that the test developer define the content domain and universe of items, and select representative items from the facets of that universe.

The domain of item content can be determined in two distinct ways. A universe of items can be either an existing set of items or a set of items that are logically linked to a particular conceptual domain. Establishing content validity for an existing set of items, such as questions in a course study guide, amounts to weighting the topics covered in the course with respect to their desired importance and selecting items in accordance with the established weights. For example, if a course guide consists of ten chapters with questions at the end of each chapter and if the instructor wishes to place equal emphasis on the first 5 chapters and no emphasis of the last 5 chapters, then content validity of a test of course knowledge would be established by selecting equal numbers of items from the first five chapters of the study guide.

In the case of item sets that are conceptually linked to a particular content domain of interest, the universe of items must be established on the basis of the conceptual/logical links between the items and the content domain. The universe of items consists of all the items that have conceptual/logical links with the domain. Guttman

(1971) has dealt formally with the concept of a universe of content in a collection of work called facet theory. Facet theory was developed by Guttman and his colleagues as an approach to research that includes a method for the formal statement of the conceptual content of the target domain and a series of empirical procedures for analyzing the interrelationships between the elements of the domain. The universe of content or domain of a particular phenomenon is presented in the form of a mapping sentence. The mapping sentence provides an explicit statement of the logical facets of the universe of content and the elements contained by each of the facets. In the terms of facet analysis, the universe of content is a Cartesian set that consists of structuples defined by all the possible combinations of the elements of each of the facets of a mapping sentence. The universe of items consists of all possible items that have logical/conceptual links to each of the structuples defined by a mapping sentence. Facet analysis also serves as a tool for generating empirical predictions. Brown (1985) states that:

> "The content of and relationship between the facets is spelled out in a mapping sentence. Empirical data are collected in terms of the facet specification. It is expected that conceptually conceived similarity will be borne out in the data; that is, there should

> be a demonstrable correspondence between the conceptual structure and the structure of the empirical observations...." (p. 20)

In a later section of this thesis it will be shown that although the mapping sentence methodology is useful for specifying the conceptual characteristics of psychological concepts (or a universe of content for item sets) the expectation that empirical results should correspond to the conceptual structure is mistaken. In principle, this argument is precisely the same argument I will use against the view that empirical, psychometric results can be taken as evidence of what a test measures.

Operational definitions have also been introduced as a method for specifying the definitional/conceptual domain of psychological constructs. The scientific tradition of operationism addresses the universe-of-content issue by formalizing the constructs of interest in an operational definition. Operational definitions typically consist of a set of necessary and sufficient conditions that place logical bounds on the domain of interest. The purpose of providing such a priori definitional frameworks is described by Betchtoldt (1959).

> "The only contributions made by operational definitions to an empirical science are those of clarity, objectivity, and precision or accuracy of statement; such definitions

> enable one to determine and eventually eliminate, the 'ignorance' and 'error' represented in any 'imperfect' formulation.... Without specification of rules, or changes therein, for using such defined terms, neither accurate communication nor precise experimentation is possible in any science." (p. 136)

Although strictly speaking, operational definitions do not serve the purpose of identifying a universe of content for psychological constructs, they do address the definitional issues associated with identifying the domain of scientific inquiry. In particular, the use of terms defined by operational definitions is governed by the rules of application laid down in the definition. Since these rules provide the basis for the employment of scientific terms, they serve properly as standards of correctness for the employment of terms in scientific discourse. In this sense, the rules of application given in operational definitions constitute the meaning of the terms involved (Baker & Hacker, 1982).

Operationism also recognizes that the clarification of the meaning of a scientific term is not addressed by the introduction of new terms. Betchtoldt (1959) remarks:

> "After several experimental studies have resulted in one or more changes of

> definition, one might say the early concepts were imprecise, incomplete, vague, or of limited usefulness. But, strictly speaking, each change of definition introduces a new concept. These definitions are *not* alternative definitions of the same concept." (p. 136)

It is clear that what is required at each stage of a scientific investigation is the clarification of the concept in current use. This amounts to no more than a statement of the rules of application for a particular construct. To the extent that the rules of application change, the meaning of the concept also changes. And since meaning is constitutive for a scientific investigation of a particular phenomenon, a change in the rules implies a change in the phenomena under investigation (Ter Hark, 1990).

Wittgenstein's influence on the logical positivists of the Vienna circle is noticeable in the operationist approach to meaning. The connection between operationism and logical positivism can be clearly seen in the tone of the operationist's treatment of meaning. The value placed on prior rules of application, and the recognition that the replacement of old terms with new terms is not a suitable method for addressing the definitional problems associated with the old, are both important aspects of Wittgenstein's philosophy of psychology (Baker & Hacker, 1980).

In practice, psychological test constructors rarely attempt to formally address the universe-of-content problem. Guttman's facet analysis, for example, is relatively unused by the community of test developers. Similarly, although definitions are often given prior to test construction, they are generally viewed as rather crude guides that will almost certainly be displaced as empirical research enriches our understanding of the constructs under investigation (Zuckerman, 1996). The types of empirical evidence most often sought are criterion and construct validity.

iii Criterion Validity

A test is said to have criterion validity if scores on the test are correlated with a criterion measure of interest. Weight, for example, has criterion validity as a test of an individual's height because weight is highly correlated with height. In contrast to content validity, criterion validity is established purely on the basis of empirical results. In order to establish criterion validity there is no need to assess the conceptual content of the test or its items. However, it *is* necessary to have access to an agreed-upon criterion. In the height example given above, the criterion measure of height presents little difficulty because there is agreement that measurements of height should be taken in accordance with an accepted set of rules. However, in the case of psychological constructs such as dominance or intelligence, the psychological community has failed to reach such agreement. Cattell (1965) states:

> "In Thorndike's own field the proliferation of empirical definitions of intelligence finally reduced many psychologists to the desperate statement: 'Intelligence is what intelligence tests measure." (p. 357)

The formal introduction of construct validity by Cronbach and Meehl (1955) was embraced by psychologists as a method that could eradicate the problem of establishing agreed-upon criteria for psychological concepts.

iv Construct Validity

Cronbach and Meehl (1955) define a construct as a postulated, qualitative attribute or structure, or quantitative attribute of people that is assumed to be reflected in test performance. A test is said to have validity for a particular construct, Ψ, if the current theory of Ψ "can embrace the variates which yield positive correlations, and does not predict correlations where we found none" (p. 286). It is argued that construct validation of a test, τ, is necessary when there is no available, agreed-upon criterion measure for Ψ. In the following section, a detailed analysis of the logical principles of construct validity is given.

v. The Logic of Construct Validation

Cronbach and Meehl (1955) provide the following logical justification for construct validation:

> "The fundamental principles are these:
>
> 1. Scientifically speaking, to 'make clear what something is' means to set forth the laws in which it occurs. We shall refer to the interlocking system of laws as a nomological network.
>
> 2. The laws in a nomological network may relate (a) observable properties or quantities to each other; or (b) theoretical constructs to observables; or (c) different theoretical constructs to one another. These 'laws' may be statistical or deterministic.
>
> 3. A necessary condition for a construct to be scientifically admissible is that it occur in a nomological net, at least some of whose laws involve observables....
>
> 4. 'Learning more about' a theoretical construct is a matter of elaborating the nomological network in which it occurs, or increasing the definiteness of the components....
>
> 5. An enrichment of the net such as adding a construct or a relation to theory is

> justified if it generates nomologicals that are confirmed by observation or if it reduces the number of nomologicals required to predict the same observations....
>
> 6. We can say that 'operations' which are qualitatively very different 'overlap' or 'measure the same thing' if their positions in the nomological net tie them to the same construct variable." (pp. 296)

These six principles provide the logical justification for the claim that a test, τ, is a measure of a construct, Ψ. Since the vast majority of psychological tests have been and currently are being constructed in accordance with these principles, they provide the philosophical/logical basis for measurement claims in the majority of contemporary psychological research. The SSS is an excellent example of a test constructed in accord with the principles of construct validity. Zuckerman (1996) states that "The Sensation Seeking Scale (SSS) was developed explicitly within the Cronbach and Meehl (1955) idea of 'construct validity'" (p. 2). In the following section, arguments against the coherence of these principles are made. We begin with a description of a series of philosophical arguments that will provide the logical basis for the critique.

The philosophical arguments contained in the following section are taken from various sources. Where necessary, comments particular to an author are cited. However, in general, the following section borrows heavily from the following sources: Baker & Hacker, 1980; Baker & Hacker, 1982; Hacker, 1988; Maraun, 1989; Maraun, 1996a; Maraun, 1996b; Ter Hark, 1990 and Wittgenstein, 1953. In particular, the arguments I make against construct validity are based on the work of Maraun (1989), Maraun (1996a) and Maraun (1996b).

vi. Wittgenstein's Philosophy of Psychology

1. Explanation

Wittgenstein dedicated considerable effort to showing how language teaching can inform our understanding of meaning. In particular, Wittgenstein's interest in the link between explanation and meaning motivated an analysis of the status of explanations as logically fundamental to issues of meaning. Since explanations are the basis for teaching the meaning of terms, the fact that language is taught and learned is of considerable importance to Wittgenstein's arguments. In the following, I provide a synopsis of Wittgenstein's thoughts on the matter.

In the first place, methods of language teaching show that explanations both *presuppose* language and are made ***within*** a language. There can be no such thing as an explanation of the meaning of a term without a linguistic practice in which the term is employed. To forge an explanation of any kind presupposes a considerable linguistic sophistication in which particular explanations serve as explanations of particular terms. Furthermore, explanations cannot be given independently of this practice. Explanations of meaning belong to language because explanations not given in a language are, tautologically, not explanations.

Explanations given in teaching the meaning of terms are not merely attempts at correlating objects with terms. Explanations give a linguistic context in which words are actually used. An explanation of the term "spirit" is, for example, a description of the manifest uses of the term. It can be plainly seen that one is not taught to say "spirit" merely by being shown one. Although ostensive definitions are often an acceptable method of explanation, they can never provide the sole explanation of a concept. A language is always required to distinguish the object "pointed at" from its context. The action of pointing across the street in an effort to provide an ostensive definition of "car" is, for example, certainly an admissible form of explanation. However, to distinguish the car from the truck parked behind it presupposes a linguistic practice in which such objects are, in fact, distinguished. Such practices make evident the purpose of explanation; that is, to convey the linguistic context in which terms are actually used.

Wittgenstein is careful to dissociate his analysis of the relation between teaching and explanation from a theoretical analysis of *how* people learn. He does not wish to propose a theory of child psychology, of language learning, but to highlight *what* one learns when one is taught a language. His purpose is to show what explanations given in a language say about meaning. The answer is simple: meaning is what is given by an explanation of meaning. Although superficially simple, this response is crucial to Wittgenstein's analysis of meaning. Most fundamentally, it *grounds* meaning by linking it to our ordinary linguistic practices.

There are important consequences to the notion that a linguistic practice is the basis for meaning. In particular, since linguistic practices are public and shared they can be readily surveyed by language users. Language use does not take place in any hidden, etherial medium of the mind but rather in a public, shared and surveyable practice. Consequently, meaning itself is public, surveyable and shared. But must meaning be grounded in a practice? Wittgenstein responds that since explanations do not apply themselves, meaning necessarily is grounded in a practice, the practice of language use.

Explanations both give rules for the application of expressions they explain and are themselves rules. Explanations are rule-governed because they belong to a practice in which certain explanations count as explanations of certain expressions, and they are rules because they serve as standards of correctness for the application of expressions. Wittgenstein developed an analysis of the philosophy of psychology and of language and meaning that was based firmly in an analysis of the features of rules and of the relation between rules and actions.

2. Rules

Rules are human creations that are grounded in a practice. The practice is not merely a correlate of rule-guided behavior but a necessary precondition for the existence of a rule. This should not lead to the assumption that rules are mere *descriptions* of a practice. The rule "receive $200 when passing Go" is not a description of the behavior of Monopoly players but a standard of correctness for play. The practice presupposed by this rule is the involvement of humans in the playing of games, with procedures, equipment, goals, winners, rules, etc.

Ter Hark (1990) describes the relation between rules and actions that accord with a rule as internal. In the context of rule-guided behavior, internal relations have three characteristics: (i) The identity of a rule is eo ipso the identification of an action that accords with the rule; (ii) the relation between a rule and an action is not mediated by a third variable, and (iii) the internal relation exists in a practice. The first relation, that a criterion for the identification of a rule is a criterion for the identification of an action that accords with the rule, amounts to the statement that the comprehension of a rule is eo ipso to know what accords and conflicts with the rule. There is, for example, no distinction between understanding the rule "receive $200 when passing Go" and knowing what actions accord with the rule.

The second relation, that no intermediary can bridge the gap between a rule and an action, strengthens the notion of an internal relation. Internal relations are not mediated by a third "correlated" or "explanatory" variable. Here Wittgenstein wants to do away with the assumption that rule-guided behavior can be explained by positing the existence of a separate process in which understanding a rule leads to an interpretation that mediates between the rule and action. He argues that no interpretation of a rule can bridge the gap between rules and actions because rather than squaring the action with the rule one must now square the action with the interpretation. The question remains: how do I know that my action is in accord with the interpretation?

Finally, internal relations are said to exist in a practice. Without a practice in which there exist rules and in which rules are actually followed (and disobeyed), there can be no possibility of distinguishing between correct and incorrect behavior and hence no rule. Without the possibility of determining whether a rule has been followed there is no possibility for a rule. A practice must be a precondition for a rule because a rule cannot interpret itself.

Rule-guided practices such as language use are normative, for they provide the basis upon which to distinguish between correct or incorrect behavior. The primacy of a practice is crucial to Wittgenstein's analysis of rule-guided behavior. He states: "I want to say: it is characteristic of our language that it grows on the basis of stable forms of life, regular ways of ac ng" (Ter Hark, 1990, p. 57). He shows that while a rule is a standard of justification, nothing justifies the rule. In the end, justification for continuation of the rule runs out and "the spade is turned by the bedrock of action." Wittgenstein refers to the endpoint of justification inherent in every practice as "agreement in forms of life."

Grounded as they are in a practice, the rules of language are a standard of justification but cannot themselves be justified. Rules are autonomous. Wittgenstein remarks that "language must speak for itself," and since language use is a rule-guided practice, "the practice must speak for itself." Language is autonomous because the fact that corresponds to a sentence cannot be described without merely repeating the sentence (Ter Hark, 1990). As standards of correctness, grammatical rules do not stand in need of justification of any sort. They are neither founded nor unfounded since they provide the beginning point for verification, justification and doubt. The constitutive, autonomous nature of grammatical rules entails that they are prior to any empirical statements about aspects of reality. There can be no empirical verification of grammatical rules because empirical statements presuppose grammatical rules. It follows then that rules are also not discoverable. Although observations of Monopoly in play may lead one to *hypothesize* that "there is a rule that $200 is received when passing Go," the ultimate justification comes via an appeal to the rule. Similarly, although acquaintance with speakers of German may lead one to believe that "tella = Df'n plate" one must ultimately consult the rules of German to verify the belief. To have evidence for or against the rule "tella = Df'n plate" presupposes the rule. While evidence can serve to justify whether a rule is useful or useless, practical or impractical it can never serve to justify the truth of the rule. True or untrue applies to propositions that presuppose a practice in which rules are already in play.

Ter Hark (1990) summarizes the autonomous, nondiscoverable nature of rules by noting that rules are *constitutive* for statements about reality. Without grammatical rules for the application of the term Ψ, there can be no such thing as a discovery about Ψ or an investigation of the properties of Ψ.

There is, therefore, a significant distinction between rules (criteria) for the application of a term Ψ and the empirical discoveries or facts about Ψ. Symptoms or empirical correlates of Ψ are learned through experience as objects of knowledge. Criteria on the other hand are standards of knowledge, necessary preconditions for knowledge about Ψ. Criteria are not learned as objects of knowledge but taught via explanation.

Wittgenstein notes a parallel between constitutive grammatical rules and so-called methodological rules, such as rules for determining units of measure. Methodological rules are similar to grammatical rules in their autonomous, constitutive nature. The measurement units for length, for example, cannot be verified against reality, or be justified by empirical evidence because they are constitutive for such verifications and justifications. A measurement is a measurement of length if and only if it is taken in accord with the prior rules for such measurements. In fact, Wittgenstein used measurement as a paradigm case for his discussions on the autonomy of grammar.

Constitutive rules such as grammatical and methodological rules can be meaningfully distinguished from modifying rules that have external standards of correctness. Wittgenstein offers the example of culinary rules as rules that are not constitutive by virtue of the existence of an external standard of correctness. Consider the modifying rule for soft boiling an egg: "boil the egg for three minutes." The external criterion associated with this rule is that the final product has a soft centre. However, it would not be correct to argue that the egg must not have a soft centre if it were not boiled for three minutes. In contrast to modifying rules, failure to follow a methodological rule renders the result meaningless with respect to the rule. If, for example, one has not followed the rule: "measure the height of the object from the bottom to the top," one does not have a measure of the height of the object.

Wittgenstein's analysis of measurement was based on the notion of family resemblance concepts. Family resemblance concepts have a number of distinctive features that will be described with reference to the concept of measurement (Baker & Hacker, 1980).

(i) The activities of measurement have no necessary common properties. It is not the case, for instance, that a single property can be found that is shared by all measurement practices. Measurement practices are linked somewhat like the fibres in a strand of rope. They are overlapping and intertwined but are not all touched by any single fibre.

(ii) Measurement practices, therefore, cannot be defined by a single set of necessary and sufficient conditions (a "Merkmal" definition).

(iii) From this it follows that the concept of measurement has no sharp boundaries.

(iv) Rather, what makes measurement practices into measurement is a complex network of similarities.

Although the meaning of the concept of measurement cannot be given by "Merkmal," measurement practices do have certain distinctive features. These features are based on the characterization of measurement as a rule-guided, normative practice.

As a rule-guided practice, the justification for the claim that ξ is a measure of Ψ is given by the meaning of Ψ. While measuring the Ψ of λ is a simple experiment, the claim that ξ is a measure of Ψ is grammatical. As such, the measurement of the Ψ of λ presupposes the rule "ξ is a measure of Ψ." As a standard of justification, the rule 'ξ is a measure of Ψ' is autonomous and nondiscoverable. Units of measurement, ξ, are also autonomous and nondiscoverable. A unit of measure is a standard of comparison and so cannot be measured for accuracy. The standard metre, for example, can not be checked to determine whether it is correctly a metre long since it is the standard for such judgements.

The autonomous nature of methodological rules has two implications that are crucial to measurement practices in psychology:

1) Empirical evidence of any kind is irrelevant to the justification of a measurement claim. Since the justification for a measurement claim is based in grammatical rules and empirical investigation presupposes the existence of grammatical rules, empirical evidence cannot bear on the accuracy of a measurement claim. There are three implications of this result: (i) empirical evidence cannot verify that a set of numbers, S, are measurements of a construct, Ψ; (ii) empirical evidence cannot verify that a set of numbers S are not measurements of a construct, Ψ and (iii) no discovery can be made concerning whether certain actions, A, constitute the measurement of a construct Ψ.

2) The meaning of existing psychological constructs, Ψ, fully determines what constitutes a measure of Ψ. The determination of whether ξ is a measure of Ψ is, therefore, given in the grammatical rules for the application of ξ and Ψ. These rules are shared and open to public scrutiny. The belief that the measurement of Ψ is a mystery waiting to be solved in careful empirical research, rests on a

misunderstanding of the nature of measurement claims. The justification for taking measurements of Ψ rests in a practice. If no such practice exists, there can be no justification for the claim that Ψ is measurable.

It is often useful to distinguish between measurements of symptoms or correlates of Ψ and criteria for the measurement of Ψ. Criteria are logically internal to Ψ, whereas symptoms or correlates are logically external to Ψ. A criterion is a presumptive implication that is learned as a standard of knowledge via teaching and explanation. For example, stretching a tape measure from the top of an object to the bottom, in a measurement of height context, is non-inductive, logical support for the claim that X is measuring the height of Y. Conversely, correlates of measurements of Ψ are learned via experience and do not provide non-inductive logical support for measurement claims. Correlates can, therefore, be discovered and may be the solution to unsolved mysteries about Ψ.

3. Supporting Measurement Claims

The previous analysis of the relation between rules and actions shows that there is an internal relation between meaning and reality. That is, that there is an internal relationship between the questions, "what is Ψ" and "what is the meaning of the term Ψ." The internal relation holds because the criterion for the identity of Ψ is the criterion for the identity of the meaning of the term Ψ. Wittgenstein remarks:

> "When one asks someone 'how do you know that the description renders what you see' he could answer for instance 'I *mean* that by these words.' But what is this 'that,' if it is not itself articulated, that is, language. Consequently, 'I *mean* that' is no answer at all. The answer is an explanation of the meaning of the word." (Ter Hark, 1990, p. 64)

The mode of investigation that informs the question, what is Ψ, is an analysis of the grammatical rules for the application of Ψ. Baker and Hacker (1982) provide a summary of the features of an analysis of meaning:

> "The correct method in philosophy of mind is the description of the use of mental expressions, of the circumstances in which they are employed, the complex grammatical structures in which they occur (and those in which they cannot significantly occur), of the behavior in different circumstances which provides grounds for their use, and of the purposes and roles of the utterances in which they occur." (p. 229)

Of course, empirical results, hypotheses and theories are irrelevant to the clarification of concepts. Based as they are on analyses of symptoms or correlates, hypotheses, theories and observations presuppose meaning and so can have no bearing on the clarification of concepts. In a later section of this manuscript, a conceptual analysis of sensation seeking provides an example of the proper form for investigations of meaning.

In the previous section it was shown that the justification for measurement claims cannot be found in a careful analysis of the empirical. Further, it was shown that measurement is a rule-guided practice in which support for a measurement claim is based on the existence of a set of rules for measuring the Ψ of λ and having correctly followed the rules that govern the practice. These rules were shown to be the basis for a grammatical relation between ξ and Ψ, the grammar of Ψ determining whether ξ does in fact constitute a measurement of Ψ. As Wittgenstein remarks, there is no concept of height without an associated grammar for the measurement of the height of things. The same, however, cannot be said about psychological concepts. Psychological concepts do not have units of measurement and are not learned in combination with measurement procedures. There is, for example, no sense to the statement: "we don't know how much dominance Tom has until we have measured it." Or: "you should measure Tom's dominance again because you did not measure it properly last time." The absence of a measurement grammar for the concept of dominance means that there are no standards of verification or justification for these statements. There can be no doubt about whether a measurement of dominance is possible in a particular case or whether a measurement of dominance is correct or incorrect, without a grammar that incorporates a set of linguistic rules for the claim that ξ does, in fact, constitute a measurement of dominance. I now bring this analysis to bear on the measurement principles upon which construct validity is based.

Vii. Principles And Problems Of Construct Validity

1. Principle 1

The first principle upon which the logic of construct validity is based centres on the scientific justification for explicating what something is. It is argued that explicating what something is amounts to setting forth the laws in which it occurs. This principle is manifest in the practice of establishing that a psychological measure has construct validity. Among the most prominent methods are 1) establishing group differences; 2) Analysis of correlation matrices; 3) Factor analysis; 4) Internal consistency analyses; 5) Analyses of change over time; 6) Analyses of change over situations and 7) Studies of process.

The purpose of each of these techniques is to determine whether the empirical performance of a test, τ, as a measure of a construct Ψ is consistent with the performance that would be expected for a measure of Ψ. If τ performs empirically as it should, convergent validity is established and if τ performs empirically as it should not, discriminant validity is established. For example, it might be hypothesized that if τ is a measure of religiosity, scores on τ should be higher for churchgoers than nonchurchgoers. If a group difference is found (technique 1) the result is said to increase the convergent validity of τ as a measure of Ψ (religiosity). Empirical analyses of the type specified by the above techniques serves the purpose of uncovering and clarifying the laws in which a construct occurs. From this principle it follows that what a test measures is determined by the empirical relationships/laws in which it occurs. In this way, only an investigation of the entire empirical network (nomological network) of research findings can provide a comprehensive explication of what constructs a test measures.

From the claim that nomological networks provide the logical basis upon which to support a claim that τ is a measure of Ψ it follows that what τ measures is for ever open to modification. Since new empirical results can always be obtained with τ, the nomological net can always be modified. Since what τ measures depends on the nomological net and since the nomological net is always open to modification, what τ measures can never be definitively established.

The position that the meaning of concepts is dependent on the entire network of scientific laws in which the entities they denote reside is a popular position in modern philosophy of science. Proponents of various variants of realism, hold that the meaning of newly introduced theoretical terms is given by the content of their parent theory (Lewis, 1970). In this formulation, changes to the parent theory brought about by new evidence or discoveries results in changes to the meaning of theoretical terms.

The position described above is, in principle, based on a philosophy Wittgenstein called the Augustinian view (Lakoff, 1987; Baker and Hacker, 1980). A primitive form of the Augustinian view has two essential features: (i) meaning is something correlated with a word and (ii) meaning is the thing for which the word stands. From these two theses it follows that explanations of word meaning are fundamentally grounded in ostensive definition and that naming and describing (objects) are the two essential functions of language. In order to gain leverage on the extent to which the first principle of construct validation is grounded in the Augustinian view, a richer description of the features of ostensive definition is required.

1a. Ostensive Definition

Ostensive definitions pair the statement "That is ..." with a gesture that exhibits the sample to public view (i.e., pointing at an object, tasting a sweet). Although there also exist verbal definitions, they merely explain the meaning of expressions within a language and so provide no means of correlating words with things. Ostensive definitions, therefore, are necessary if language is to represent reality. Ostensive definitions allow us, it seems, to accomplish the goal of representing reality with language. Furthermore, they appear to go beyond language by concentrating on the *things* that language represents. Such definitions seem fundamental because they ground meaning in more than just mere words. The meaning of terms is the object with which the term is correlated. If there exists no such object, then the word that signifies it has no meaning. The answer to the riddle of language is simple; if we want to know what a word really means we must look at the object it signifies. Herein lies the fundamental principle of construct validation: an investigation of what something is, is an investigation of the empirical laws in which it resides.

Let us explore the link between the Augustinian view and the first principle of construct validity a little further. The notion that the meaning of a word is given by ostensive definition implies that words are not only correlated with objects but that the meaning of a word is given in total by its referent (the object it refers to). If supplementation of an ostensive definition were possible, then such definitions could not provide the foundation of language. The clarification of the meaning of an expression is, therefore, dependent on an unambiguous description of its referent. The rules of application for a concept are grounded in and hence given *completely* by the essential nature of objects.

The Augustinian view endows upon science a privileged status in investigations of meaning. Since meaning is the essential nature of objects, only a scientific investigation can settle questions pertaining to the meaning of unanalyzable concepts. The meaning of time, it is believed, provides a prime example of the marriage between science and language. Originally, the story goes, time was a poorly understood concept with ambiguities that were not fully appreciated until Einstein reformulated the laws of physics. Although in Newton's age there was a concept of time that was used with relative success, Einstein revealed that we were, in fact, mistaken in our use of the concept. His empirical investigation into the nature of the universe resulted in discoveries about the meaning of time that revealed an error in our use of the term. In this way, scientific investigations of the laws of nature are fundamental to the determination of the meaning of concepts.

On this formulation of Einstein's work, Wittgenstein comments:

> "You forget what Einstein, as I surmise, has taught the world: that the method of time measurement belongs to the grammar of time sentences." (Wittgenstein, 1990, p. 72)

Here Wittgenstein repudiates the view that the special theory of relativity represents an empirical discovery about the meaning of time. Consequently, he denies that the meaning of time was compromised by Einstein's scientific investigation. Rather, Einstein's contribution was conceptual: the clarification of conceptual problems entailed in the grammar of the concept of time. The sceptic may find a review of special relativity useful.

In his famous 1905 paper Einstein proposed that physicists do away with the concept of absolute space and with the accompanying assumption of a resting absolute frame of reference (the ether). In particular, the idea embodied two associated postulates. The first of these states: All the laws of physics have the same form in all inertial reference frames. This postulate extended the Newtonian relativity principle from the laws of mechanics to include the laws of electricity and magnetism. The second principle is a special case of the first: Light propagates through empty space with a definite speed, C, independent of the speed of the source of observer. Although Einstein's theory obviously has numerous interesting consequences, this discussion will be restricted to an analysis of their relevance to the conceptual features of simultaneity and time.

The special theory of relativity has significant implications for determinations of the simultaneity of events. Prior to special relativity, simultaneous events were events observed to take place at the same time. Significantly, this picture makes no distinction between events taking place in close proximity and those taking place at great distances. However, the constancy of the speed of light with respect to any particular inertial frame of reference raises the potential for confusion in the determination of the simultaneity of events occurring in vastly different inertial frames of reference. Consider a case in which two observers O_1 and O_2 are located in the same inertial frame of reference - they are close together in space, are moving at the same speed and in the same direction. Now suppose that two sources of light located at equal distances in front and behind O_1 and O_2 each emit a single burst of light, B_f and B_b, at time T_1. Since light travels at a constant speed, regardless of the observer's inertial frame of reference, O_1 and O_2 observe B_f and B_b at the same moment in time T_2. We might say that B_f and B_b are simultaneous. However, could we say that B_f and B_b are simultaneous if O_1 and O_2 were not located in the same inertial frame of reference? Suppose that O_1 and O_2 are moving along a single, straight path through space at some speed, C, but are travelling in opposite directions. Suppose also that as O_1 and O_2 pass each other, two sources of light located at equal distances in front and behind O_1 and O_2 and in O_2'S inertial frame each emit a single burst of light, B_f and B_b at time T_1 (note: B_f is the light source emitted in the direction of the movement of O_1). We have the identical situation described above except that O_1 and O_2 are moving in opposite directions. In this case, however, O_2 observes B_f and B_b at the same time, T_2, but O_1 observes B_f before T_2 and B_b after T_2. The problem is clear: since O_1 does not observe the two light sources at the same time, can we say that B_f and B_b are simultaneous?

The conclusion Wittgenstein wants to draw here is that Einstein made no discoveries about the meaning of the concepts of time or simultaneity. Rather, he argues that the concept of simultaneity is rendered ambiguous in the novel context shown to us by Einstein. Although we had a clear concept of simultaneity for the situations in which it was used in practice, Einstein's work forced us to examine the application of the concept in an unfamiliar context. We do not require an explication of the fundamental, essential nature of simultaneity; we require an explication of what it *means* for two events taking place in different inertial frames of reference to be simultaneous. Waismann (1965) summarizes the point:

> "Surveying the argument today it is perfectly clear that a way out of this dilemma could only be found by turning away from the world of facts to a consideration of concepts....'What exactly does it mean to say that they are simultaneous?'...If it is used to refer to events in quite different places, we require a statement of what it is to mean in this new context. This step was taken by Einstein. He neither discovered hitherto unknown facts, nor did he suggest a hypothesis which explains better the known facts; rather he cleared away from the concept of simultaneity the confusion which had surrounded it." (p. 12)

This brief survey of special relativity also clarifies what Einstein taught us about the nature of time. The argument is similar to the above. In particular, two observers O_1 and O_2 in different inertial reference frames (reference frames moving at different speeds relative to each other) observe time passing at the same rate T_1. However, O_1 observes that time passes more slowly in his own reference frame than in the reference frame of O_2. The well known twin paradox exemplifies this principle. Suppose that one of a pair of twins is chosen to fly a mission to the moon. The other is stay on earth and await his return. If the first twin leaves earth at T_1 and travels to the moon and back at a high rate of speed, upon his arrival at T_2 he will be *younger* than his earth bound twin. This is because the first twin's clock has moved more slowly than the earth bound second twin's clock. Stated simply, the faster a clock moves relative to a particular inertial reference frame, the more slowly it runs. This is a consequence of the second principal of special relativity.

The question remains: What does this result teach us about the nature of time? Most importantly, does it teach us that time now has a different *meaning* than we thought? Do the laws of physics as reworked by Einstein have implications for the meaning of our old concept of time? Wittgenstein wants to say yes, but not in the way we might think. In particular, the introduction of a new context of time measurement (i.e., time measurement in vastly different inertial reference frames) requires a *clarification* in the grammar of time sentences, but not a change in the meaning of time. Einstein's work introduced a new context in which it is necessary to clarify exactly what it means to say that time passes at a given rate when measured in two vastly different inertial reference frames. Significantly, special relativity does not require that the criteria for the measurement of time be changed. Only that time, as we measure it, passes at different rates depending on the speed a clock travels. We learned something about time that we did not know before. And this would not be possible if special relativity had required a change in the meaning of the concept of time (see Jackson and Maraun, 1996).

We now return to the link between the Augustinian view and the first principle of construct validation. In the previous example it was shown that the momentous restructuring of the laws of physics brought about by the introduction of relativity theory did not inform our understanding of the meaning of time (i.e., what time is). While our understanding of how time passes in various inertial reference frames did change, the criteria of application for the concept of time (what time is) did not. This illustrates that finding out more about time must be distinguished from understanding the meaning of time. Investigations of the laws of nature that attempt to reveal the essential nature of the world are *not* fundamental to the determination of the meaning of concepts.

This conclusion is supported by previously introduced theory. The constitutive, autonomous nature of grammatical rules for the application of the concept of time entails that they are prior to any empirical statements about time. There can be no empirical verification of the meaning of time because grammatical rules are constitutive for empirical statements. Rules are also not discoverable. Although observations of moving clocks may lead one to the hypothesis that "time passes differently in different internal reference frames," the ultimate verification of the hypothesis must presuppose a concept of time that has prior meaning. While evidence can serve to justify whether present time measurement practices are useful or useless, practical or impractical, it can never serve to justify the truth of our practice. True or untrue applies to propositions that presuppose a practice in which rules are already in play. As Ter Hark (1990) notes, grammatical rules are constitutive for statements about reality. Without grammatical rules for the application of the term time, there can be no such thing as a discovery about time or an investigation of the properties of time.

There is, therefore, a significant distinction between rules for the application of the term, time (grammatical rules, criteria), and the empirical discoveries or facts about time. The agreement of clocks based in distinct inertial reference frames are symptoms or empirical correlates of time that have been learned through experience as objects of knowledge. Criteria for the application of the concept of time, on the other hand, are standards of knowledge, necessary preconditions for knowledge about time.

Although there are obvious differences between physics and psychology, the import of this argument to the respective disciplines is equivalent. No matter how elaborate the specification of the empirical laws governing Ψ becomes, they can have no bearing on what Ψ is. For a set of empirical results to be about Ψ necessarily requires that the grammatical argument that they denote Ψ can be supported. It follows, therefore, that the first principle of construct validity is false. An investigation of the empirical laws into which Ψ enters presupposes the ability to support the claim that the obtained empirical results denote Ψ, and can, therefore, have no bearing on the meaning of Ψ.

2. Principle 2

The second principle of construct validation describes the admissible constituents of empirical laws that relate phenomena. It is difficult to verify Principle 2 because the terminology employed by Cronbach and Meehl (1955) is not sufficiently described. Specifically, an analysis of the validity of this principle rests on the meanings of "law," "observables" and "theoretical constructs." In the following section an analysis of the meanings of these concepts will be used to show that, for any reasonable set of meanings, Principle 2 incorporates a conflation of empirical discovery, theory and meaning.

2a. Laws

A set of phenomena behave lawfully if they are related in a highly predictable or consistent fashion. Guttman (1982), for example, describes the first law of achievement as the consistent and predictable finding that correlation matrices of achievement items are positive manifold. The movement of the planets, for example, is also lawful because they are observed over time to consistently follow predictable paths. Of course, to establish that phenomena behave lawfully requires that observations of the behavior of the relevant phenomena have been made over time, circumstances, etc. In short, the claim that a law has been discovered can only be supported by recourse to a core of empirical evidence.

As we have seen, empirical evidence about a phenomenon, Ψ, presupposes a criterion for Ψ. To review, a finding can not be shown to be about Ψ if it cannot be justified by recourse to the criteria of application for Ψ that the observations were indeed made of Ψ. The discovery of the first law of attitude, for example, presupposes the existence of criteria for attitude items. Empirical laws can only be discovered, discussed and incorporated into theory if there exist criteria of application for the terms of the law.

2b. Theoretical Terms

Although Cronbach and Meehl do not provide a clarification of the meaning of theoretical terms, they do cite a body of philosophical work in which theoretical terms are of considerable importance. Lewis (1970) introduces theoretical terms (T-terms) as terms of a new theory T that have never been used before. T-terms are unfamiliar in the sense that our only clue to their meaning is the theoretical postulates of T that introduce them. Further, T-terms are said to name entities that they denote. A T-term is denotationless if it names an entity (T-entity) that has no realization - does not exist. It follows, then that T-terms of unrealized theories do not name anything. Lewis (1970), for example, provides the following example:

> "Phlogiston presumably is a theoretical term of an unrealized theory; we say without hesitation that there is no such thing as phlogiston. What else could we possibly say? Should we say that phlogiston is something or other, but (unless phlogiston theory turns out to be true after all) we have no hope of finding out what. Let us say, then, that the theoretical terms of unrealized theories do not name anything." (p. 432)

The close ties between the meaning of T-terms and the truth of the theories in which they are introduced has implications for their meaning at any particular time after their initial introduction. Lewis questions:

> "If T is thus partially reduced and partially falsified, or revised for any other reason, do the T-terms retain their meanings?" (p. 445).

Questions concerning the meaning of T-terms also arise when a theory has never received support or verification. If there are no empirical results to verify a newly introduced theory, what are we to say the T-terms mean? What, for example, did id, ego and superego mean when Freud introduced the terms? Such problems are an inevitable consequence of the Augustinian practice of grounding meaning in the empirical and were at the heart of Wittgenstein's criticism of the Augustinian view.

2c. Observables

Presumably, the need to distinguish between theoretical terms and observables is that the entities denoted by theoretical terms may not always be observable. For example, in construct validation observed test scores are assumed to reflect the existence of underlying "theoretical" constructs. Such constructs are often termed hypothetical because they are not yet observed or are perhaps directly unobservable. However, the test scores upon which the hypothesis is based are observed.

2d. Law, Theory and Meaning

Since a law is established on the basis of empirical evidence, it is not correct to characterize a law as relating observables to theoretical constructs or theoretical constructs to themselves. This is simply because theoretical constructs refer to entities that have not yet been directly observed. And, if no observations have been made of the theoretical construct, its empirical behavior cannot be established. An example is required.

Consider the theoretical construct "g." Let us agree that "g" is well described by the term generalized intelligence and that its existence is hypothesized on the strength of empirical evidence relating performance on achievement tests. Now, the empirical observations that have led to the hypothesis that individuals have generalized intelligence are not observations of generalized intelligence. They are observations on particular tests that indicate the existence of g. Of course, if they were observations of g, then g would be an observable and not a "theoretical construct." Since only the scores on the test that indicate the existence of g are observed, it is only these scores that can become a component of an empirical law. The theoretical construct, g, however, is not observed and as such cannot be a component of an empirical law.

One might counter that although g itself is not observed we can measure g by taking individuals' scores on the first component extracted in an analysis of achievement tests. Although the component itself is not g, the scores on the component are measures of g. This is mistaken, however, because it elevates evidence for the existence of g to the status of measurements of g. A set of numbers can be either evidence for the existence of Ψ or measurements of Ψ but not both.

It was shown earlier that the autonomous nature of methodological rules implies that the meaning of existing psychological constructs, Ψ, fully determines what constitutes a measure of Ψ. Further, it was shown that measurement is a rule-guided practice in which support for a measurement claim is based on the existence of a set of rules for measuring the g of λ and having correctly followed the rules that govern the practice. These rules were shown to be the basis for a grammatical relation between ξ and g, the grammar of g determining whether ξ does in fact constitute a measurement of g. As Wittgenstein remarks, there is no concept of height without an associated grammar for the measurement of the heights of things. The determination of whether ξ is a measure of g is, therefore, given in the grammatical rules for the application of ξ and g. These rules are shared and open to public scrutiny. Symptoms or correlates of g, however, are logically distinct from criteria for the measurement of g. Symptoms such as evidence for the existence of g rest on empirical evidence. Evidence for the existence of g, for example, is generated in accordance with empirical hypotheses, verified on the strength of empirical evidence, may be weak or strong depending on its agreement with prior expectations. In short, issues pertaining to the existence of g are addressed via an appeal to empirical evidence, while criteria for the measurement of g are conceptual. From the logical distinction between empirical issues and conceptual issues it follows that evidence for the existence of g cannot also be measurements of g. This argument is of no small importance in the psychological testing literature. Psychologists have struggled with this problem from the inception of the concept of g. Consider Jensen's (1969) assertion that: "Intelligence, like electricity is easier to measure than to define (p. 5)." and Schonemann's (1987) bemused response: "It is complete nonsense to claim to be able to 'measure intelligence' without being able to define it" (p. 317). The argument provided above shows that, although he did not mount a strong defence of his claim, Schonemann was essentially correct.

The previous example and associated argument show that an empirical law cannot relate phenomena that are not empirically observable. Empirical laws are supported on the strength of empirical evidence, and without the possibility for empirical observations of Ψ, there can be no empirical law governing observations of Ψ. We can, therefore, maintain only the first of the three tenets of Principle 2.

That is, that the laws of a nomological net may relate observables to observables.

3. Principle 3

Principle 3 states that a necessary condition for a construct to be scientifically admissible is that it occur in a nomological net, at least some of whose laws involve observables. In the previous section, it was shown that all concepts involved in an empirical law must be observable. Since the practice of establishing scientific laws is fundamentally an empirical practice it must be the case that the phenomena of investigation be, in principle at least, observable. At this point, it is certainly reasonable to enter into a more detailed discussion of observability and its relevance to scientific practice. In fact, a considerable amount of work in this area has resulted from Heisenberg's discovery of the uncertainty principle (Hawking, 1990). Although I consider issues pertaining to the meaning of phrases like "observable in principle," "unobservable," etc., to be of considerable importance, they are sufficiently peripheral to the principles of construct validity that an in-depth treatment in this manuscript is not warranted. Let us now consider the fourth principle of construct validity.

4. Principle 4

In this principle we are told that learning more about a theoretical construct is achieved through empirical analyses. Cronbach and Meehl (1955) provide an example of this principle:

> "We will be able to say 'what anxiety is' when we know all 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)

Thus, learning more about anxiety constitutes an empirical investigation of the nomological net in which anxiety resides. As the nomological net is elaborated it is expected that we will gradually come to understand more accurately what anxiety really is.

As the previous analysis of Wittgenstein's philosophy of psychology has shown, this principle and the associated quote conflate the meaning of a concept (i.e., what anxiety is) with knowing more about a concept. To review, one can know more about Ψ, discover interesting new facts about Ψ, observe changes in the relationships Ψ has with other concepts over time, only as long as one has a prior criterion for Ψ. Without a prior criterion for Ψ no observed empirical relationship can be said to be about Ψ. This argument is crucial to investigations of the construct validity of a psychological concept because it shows that the tools of construct validity can *never* lead to an accurate understanding of what anxiety really is. Consequently, no investigation of the construct validity of any psychological concept in current use has ever had any bearing of any kind on what the particular phenomena of interest is.

What can lead to an understanding of what anxiety is, is an investigation of the meaning of the term. The proper form for investigations of the meaning of terms is conceptual. Only an analysis of the correct use of Ψ, what stands as an explanation of the meaning of Ψ and the complex grammatical structures in which Ψ belongs (and does not belong) can inform the question of what Ψ is. Only when such issues are clarified can coherent empirical research on the empirical behavior of Ψ begin. Since, in construct validity, empirical evidence provides the basis for clarifying psychological constructs, it is no surprise that such conceptual investigations never precede studies of the construct validity of a test. Therefore, given that construct validity provides the justification for the vast majority of measurement claims in modern psychology, and that measurement claims can not be justified by construct validation procedures, the vast majority of measurement claims in modern psychology lack justification.

5. Principle 6

Principle six is the remaining substantively important principle of construct validity. It states that operations that are qualitatively very different measure the same thing if their positions in the nomological net tie them to the same construct variable. Qualitatively different operations are measurements of or scores on conceptually distinct phenomena. Positions in the nomological net refer to the web of empirical associations a variable has with the other variables in the net. Principle six, therefore, tells us that conceptually distinct variables measure a single construct, Ψ, if they have similar patterns of empirical associations with other variables.

The arguments against this position have already been elucidated. However, they are worth reviewing. Earlier it was shown that the autonomous nature of methodological rules has two implications that are crucial to measurement practices in psychology:

1) Empirical evidence of any kind is irrelevant to the justification of a measurement claim. Since the justification for a measurement claim is based in grammatical rules and empirical investigation presupposes the existence of grammatical rules, empirical evidence cannot bear on the accuracy of a measurement claim. The three implications of this result show that Principle 6 is false. First, empirical evidence cannot verify that a set of numbers, S, are measurements of a construct, Ψ. No matter how similar a set of numbers are with measurements of the height of objects, for example, there can be no

justification that the set are measures of the height of objects unless it can be shown that they were taken in accordance with the prior grammatical rules for the measurement of height. For example, the strong positive correlation between height of mountain and depth of snow are not evidence that measurements of the height of mountains are measurements of the depth of snow. Only the grammatical relations between the terms height, mountain, depth and snow can settle this claim.

Second, empirical evidence cannot verify that a set of numbers S are not measurements of a construct, Ψ. One might object that empirical observations can, in fact, rule out a set of measurements as measurements of a particular construct Ψ. If, for example, a set of measurements of IQ included the value 500, this might be evidence that they are not measures of IQ. However, to know that humans do not have IQ'S of 500 presupposes a practice of measuring the IQ of humans that is grounded in rules for the measurement of IQ. There could not be the possibility for such a judgement without an accumulation of empirical evidence in which the number 500 is absent. The claim that IQ does not reach such extreme values is empirical and so presupposes a set of public and shared rules for the measurement of IQ.

Third, no discovery can be made concerning whether certain actions A constitute the measurement of a construct Ψ. Measurement is a rule-guided practice in which measurements are taken in accordance with rules. It has been shown that the relation between a rule and an action that accords with the rule is internal. Internal relations are grammatical, not mediated by explanatory variables, and exist in a practice. Internal relations are, therefore, autonomous and nondiscoverable. A statement of a rule is, eo ipso a statement of the actions that accord with the rule. There can be no discovery that certain actions constitute the measurement of Ψ simply because the actions that constitute measurements of Ψ rest in an internal, grammatical relation between the actions and the rule. Since grammatical rules are autonomous and nondiscoverable, there can be no possibility of discovering the actions that constitute a measurement of Ψ.

2) It was also shown that the meaning of psychological constructs, Ψ, fully determines what constitutes a measure of Ψ. The determination of whether ξ is a measure of Ψ is, therefore, given in the grammatical rules for the application of ξ and Ψ. The belief that the issue of how to measure Ψ can be bolstered or supported by careful empirical research, rests on a misunderstanding of the nature of measurement claims. The justification for taking measurements of Ψ rests in a practice. If no such practice exists, there can be no justification for the claim that Ψ is measurable.

Clarification of the measurement issues that haunt psychological investigators can only be achieved by an analysis of the conceptual or, as Wittgenstein preferred, grammatical contours of a concept. Causal explanations, theories or empirical hypothesis are irrelevant to the clarification of concepts. While it is true that empirical evidence can have no bearing on the clarification of concepts, the converse is most certainly not the case. In so far as psychology makes use of everyday concepts in common use, it can benefit directly from a logical description of these concepts. Wittgenstein is convinced that the need to clarify conceptual problems associated with everyday concepts is what distinguishes psychology from, for example, physics or chemistry:

> "the confusion and barrenness of psychology is not to be explained by calling it a "young science"; its state is not compatible with that of physics, for instance, in its beginnings....For in psychology there are experimental methods and *conceptual confusion*. The existence of the experimental method makes us think we have the means of solving the problems which trouble us; though problem and method pass one another by."
> (Baker & Hacker, 1982, p. 228)

The problem Wittgenstein refers to here is the clarification of everyday common-or-garden concepts that are widely ramifying, lacking in unifying employment and not readily surveyable. These are not the technical terms of an advanced science but terms with complex grammatical structures. By their very nature such terms are not readily reducible to the necessary and sufficient conditions of an operational definition or a well-defined universe of content. In the following section, I attempt to remedy the injustice that Zuckerman has done to the concept of sensation seeking in a clarification of the complex logical/grammatical/conceptual structures of the term.

viii. A Conceptual Analysis of Sensation Seeking

Our language is rich, complex and replete with metaphor, cliche and rhetoric. Metaphor and cliche, for example, are fundamental to language. We "hit the road," are simultaneously "cool" and "hot," can be so "good" that we are "bad." Since a linguistic practice is a standard of justification, such usage is not fundamentally substandard or incorrect. However, in certain contexts, the failure to recognize a metaphorical or rhetorical usage can have damaging consequences. For example, it is currently in vogue to treat criminal behavior as a *disease* (Monahan & Steadman, 1994) in need of *treatment*. Although certain scientific and economic advantages may obtain from this kind of move, the failure to recognize its metaphorical status can have severe consequences. In this particular case, social scientists working in the area of violence prediction have all but forgotten that criminal behavior is immoral, while contracting a disease is not. Consequently, they forget that, to the extent that their investigations ignore the moral issues central to criminal behavior, their treatment is incomplete. For example, the failure to consider personal responsibility, punishment or remorse in the scientific analysis of criminal behavior, renders such investigations only a partial treatment of the problem. Furthermore, it is reasonable to conclude that, in the sense described above, a partial treatment is no treatment at all since researchers treating a disease are not logically studying criminal behavior. And as I have shown, no empirical investigation can ever overturn this claim.

Hacker, (1988) comments in a similar vein on the place of analogy in scientific investigation:

> "Fruitful analogies are the go-cart of creativity. The hydrodynamic analogy proved immensely fruitful in the development of the theory of electricity, even though electrical current does not flow in the same sense as water flows and an electrical wire is not a kind of pipe .... But if a schoolchild - after having his first lesson about

> electricity - proceeded to cut electrical wires, turn them up vertically and shake them in order to get some of the electricity to pour out, we would, with justice, think that he had not understood the lesson, not grasped the character of the 'model' of electrical current, and wholly misconstrued the concepts of flow of electricity, electrical potential, and so forth." (p. 485)

The import is clear, a scientific investigation that does not distinguish metaphor, analogy and rhetoric from technical usage is bound to border on the incoherent.

This discussion is significant here because the term sensation seeker is at best a rather contrived metaphor. Although it is common to employ dispositional concepts such as risk taker or thrill seeker, sensation seeker is a concept with no clear standards of application. In the following, an analysis of the technical sense of sensation seeker will serve as a backdrop against which to view its poorly grounded, metaphorical and cliched use in sensation seeking research.

A sensation seeker is tautologically someone who seeks sensations. What are sensations? This question has occupied the minds of many of the major figures in the history of psychology. Hume, for example, classified the basic elements of mental life as "simple" sensations. A direct result of Hume's empiricist tradition was his treatment of sensory experiences as the basic units by which the mind acquires knowledge. This reduction of psychological concepts to their fundamental sensory elements, however, constitutes a theory of mind and so is not relevant to an analysis of the logical contours of the term sensation. The problem of sensation was also central in the work of James, Brentano and Wittgenstein. In fact, Wittgenstein used James as a source of, what were to his mind, a series of confusions on the matter (Ter Hark, 1990). In the following section, I borrow from these works in an attempt to formulate my own conceptual analysis of sensation.

1. Sensations

In the first place, sensations are felt. They are felt, that is, as an aspect or content of experience. As an element of experience, sensations are not capacities, tendencies, abilities or dispositions. Rather, like emotions or mental images, sensations are states. Unlike dreams or sleep, however, sensations are waking states. And unlike hypnotic trances or drug-induced unconsciousness, sensations are conscious states. The feeling of pain, for example, is internally related to consciousness in the sense that if one is in pain this implies that one is also conscious.

The characteristic of sensations that they are felt distinguishes them from "impressions". By "impressions" I mean sensory experiences resulting from the stimulation of the visual, auditory and olfactory organs. The sight of a painting, the sound of a drum, the smell of a rose and the taste of an artichoke are all sensory experiences but are not sensations. Let me illustrate this distinction with an example. Consider the smell of a rose. Would one not want to argue that the experience of smelling a rose is actually felt and so is, in fact, a sensation? Surely what is felt are the movements of air through the nostrils and the expansion of the lungs. However, the *smell* of a rose is not felt but only experienced.

There are, in general, two kinds of sensations that can be distinguished on the basis of their localization. Bodily sensations such as feeling hot, euphoria or exhaustion are diffusely located. These sensations are not localized in the same way as kinaesthetic sensations. Kinaesthetic sensations such as a sprained ankle or bent knee do have specific locations. In fact, there has been a fierce debate over the issue of the location of sensations. In general, the debate centres on the issue of whether a sensation is an independent sense-datum that provides information about the state of our bodies. Proponents of the local sign theory hold that sensations carry with them a peculiar shade or feeling ("local colour") that is derived from their topographic position. This local colour is a pure quale that offers clues about the location of the sensation. Wittgenstein rejects the local sign theory on the basis that it presupposes the possibility for private ostensive definition. Wittgenstein's rejection of the local sign theory does not amount to a repudiation of the thesis that sensations have locations. Rather, it is based on the premise that the relation between, for example, the feeling of a bent knee and having a bent knee is internal. Wittgenstein remarks:

> "I feel that I am moving all right, and I can also judge roughly *how* by the feeling - but I simply *know* what movement I have made, although you couldn't speak of any *sense-datum* of the movement, of any immediate inner picture of the movement. And when I say 'I simply *know*...''knowing' here means something like 'being able to say' and is not in turn, say, some kind of inner picture." (Ter Hark, 1990, p. 211)

In effect, Wittgenstein is saying that the local-sign theory has the sensation serving the role of a *representation* of the object of sensation and the experience of the sensation serving the role of an *interpretation* of the sensation. This picture is misguided because the link between the content of an impression and the object of the impression is *internal*. Since, internal relations are not explained via a third explanatory or correlated variable, there can be no intermediary between the object of sensation and the experience of a sensation.

Although there is disagreement over how to say that sensations have location, there is agreement that kinaesthetic sensations have non-diffuse locations. The locatability of sensations is significant because it provides a further basis upon which to distinguish them from impressions. While we feel pain in our elbow or have an aching tooth, we do not see red in our eyes or hear screams in our ears. Rather, we see with our eyes and hear with our ears. This rests on the premise that what is perceived with our sensory organs is not a sensation in the part of the body in which the organ resides.

2. Sensation seeking

As a state of consciousness or content of experience, a sensation is a temporary phenomenon that has real duration. The beginning point and end point of pain, for example, are clockable events. Dispositions, on the other hand, are enduring and not interrupted by changes in attention or states of consciousness. Although dispositions can mature or recede, their existence is not ascertained by periodic spot-check. Although sensations are states of consciousness, sensation seeking is a dispositional concept.

A sensation seeker is a person who exhibits enduring, persistent and regular sensation-seeking behavior.

Someone in search of sensations or who seeks out sensations, actively pursues them. They do not merely come across sensations in their everyday travels but actively search for them. The search is enduring, occupies them regularly and is an integral aspect of their lives. They do not merely like, prefer or enjoy sensations; they *seek* them. The combination of the terms "sensation" and "seeker" in this phrase, however, is not without its conceptual difficulties. Language use is grounded in a practice and the meaning of terms is given in this practice. Problems arise in determining the meaning of sensation seeker because this phrase is not a common part of our practice. An analysis of the meaning of such terms is often usefully explicated by comparing them to similar phrases that are part of our practice. Thrill seeker, for example, is an often-used phrase that is roughly in the conceptual domain of sensation seeker. Although "thrills" are diffusely located sensations with real duration, there is a significant difference between thrilling sensations and sensations in general. Thrills are *extraordinary* sensations. They are sought because they are uplifting, uncommonly exciting, arousing and extraordinarily pleasing. Unlike mere sensations, the arousing aspect of thrills is unusual and uncommon and so is a sensation to be sought after. Logically we seek-out the uncommon, unusual, exciting or pleasing, not the regular or common. There is no sense in seeking the common; one has it on a regular basis. And since sensations are common, regular experiences that are an integral part of every person's conscious experience, they need not be sought after. Hence, the term sensation seeker is internally contradictory and taken literally, meaningless.

The above analysis of sensation seeker is based on the literal meaning of the terms "sensation" and "seeker." Although one would expect scientific investigations to avoid cliche and metaphor, it is clear that a literal meaning of sensation seeker is not commensurate with Zuckerman's work.

3. Zuckerman's Treatment of the Concept of Sensation Seeking

The scientific development of the concept of sensation seeking and its associated measurement instruments (Forms 1 through 5) are a rich source of conceptual confusions. Zuckerman's unsophisticated forays into the conceptual realm are characterized by unclarity, ambiguity, contradiction and incoherence. In the first place, we are offered a series of definitions that begins years *after* the SS items were written and a hodge podge of empirical evidence, hypothesis, theory and supposition to support their continual reformulation. Even the definitions themselves are riddled with conceptual errors. Consider the latest in the series:

> "Sensation seeking is a trait defined by the seeking of varied, novel, complex, and intense sensations and experiences, and the willingness to take physical, social, legal and financial risks for the sake of such experience." (Zuckerman, 1994, p. 27)

Zuckerman is apparently unaware that this "definition" is largely a statement of the empirical correlates of sensation seeking rather than a definition of the concept itself. In particular, the "willingness to take physical, social, legal and financial risks" *must* be an empirical correlate of sensation seeking. How can it be established that sensation seekers are *willing* to engage in such behavior if it has not been established/discovered in empirical research? Since sensation seeking is a *dispositional* concept, the issue of what sensation seekers are *willing* to do for the sake of sensational experiences can be addressed only in empirical work. In the following section, it will be shown that Zuckerman is guilty of numerous conceptual errors throughout the development of the concept of SS.

3a. Optimal Level of Stimulation

In the opening sentence of the abstract of the first published sensation seeking paper Zuckerman et al (1964) state:

> "This article reports the development of the Sensation-Seeking Scale (SSS) designed to quantify the construct: 'optimal level of stimulation'." (p. 477)

In the final sentence of the introduction to the same paper Zuckerman et al repeat this stance:

> "Interest in the personality implications of the 'optimal stimulation' concept and its possible application to ongoing perceptual isolation experiments led us to attempt to develop a questionnaire scale which might measure this postulated trait. While it is possible that sensation seeking is specific to the various types of sensations, we hypothesized that a general factor would emerge from responses to diverse items." (p. 477)

At first glance, one must assume that Zuckerman et al provided some rationale for the link they forge between optimal level of stimulation and sensation seeking. In the second quote, for example, sensation seeking is introduced as a place-holder for optimal level of stimulation with no apparent logical support. Logical support, of course, is necessary because sensation seeking and optimal level of stimulation are two quite distinct concepts. It is, for example, nonsense to speak of an optimal level of a disposition such as sensation seeking. One does not have a preference for, or seek out "optimal levels" of dispositions. Furthermore, the logical connection between sensations and stimulation is far from trivial. While they are certainly not identical concepts, they do share contexts of application that are conceptually significant. A conceptual analysis of the type reported above is necessary to reveal their logical interconnections and clarify their distinct natures.

So then, we ask, what did Zuckerman have to say on the matter? In Zuckerman's first two major papers on the sensation seeking scale there is not a single comment of any kind on the logical link between optimal level of stimulation and sensation seeking. Thus, a study in the spirit of construct validity begins. At the outset of Zuckerman's world famous and respected psychology is a profoundly obvious category error in which a sensation seeking scale is developed to measure an optimal level of stimulation. In the past thirty years, the community of psychologists has not only accepted this absurdity with blind faith but celebrated the introduction of new ones. For now the very same items that have been used to measure the dispositional concept of sensation seeking belong to a *susceptibility* scale (BS), *seeking* scales (TAS, ES and SS) and an *inhibition* scale (DIS). And when we ask how these things are measured, we are handed a set of items that were chosen *before* there was any attempt to define the concepts.

Let us continue to examine the content of the second quotation given above. We are told that "it is possible that sensation seeking is specific to the various types of sensations." Once again, one must assume that, in the spirit of science, Zuckerman gave careful consideration to the famous works of Brentano, Spinoza, James and Wittgenstein before providing a list of the candidate sensations for his scale. However, none of their work was ever mentioned and no treatment of the various types of sensation was ever given. Once again construct validity is substituted for scientific rigour. This omission is not trivial because it compromises the premise of the entire SS investigation. Consider what is meant in the above quotation by "possible." Is it that sensation seekers = Df'n people that seek specific sensations and that sensation seekers exist but it is not known which specific sensations they seek. Or is it that sensation seekers = Df'n people that seek specific sensations x, y and z and it is not known whether sensation seekers exist. Or is it that sensation seekers = Df'n people that seek all sensations, and that it is not known whether sensation seekers exist. Zuckerman's failure to clarify his definition of sensation seeker and distinguish between empirical and conceptual forms of investigation leave us is in the lurch. We simply have no way to distinguish the possibilities and hence lack a firm foundation upon which to proceed with sensation seeking research. In fact, the confusion, ambiguity and uncertainty are invited by Zuckerman when, in the above quotation, he refers to sensation seeking as a "postulated trait" which "might measure" SS. Let us see whether the SS items do, in fact, denote sensation seeking.

ix. A Conceptual Item Analysis of The SSS

Regardless of whether sensation seeking is meant in a technical or metaphorical sense, it is a dispositional concept. Since the instantiation of dispositional concepts rest on behavioral criteria, items that instantiate needs, attitudes, preferences, likes/dislikes or desires do not denote sensation seeking. In order to be a sensation seeker an individual must have engaged in sensational activities. Furthermore, individuals who engage in sensational activities by mere happenstance are not sensation seekers. An individual is a sensation seeker if he/she actively pursues sensations for the purpose of and with the intention of having a sensational experience. This means that it is not enough to merely pursue activities that, for some, are sensational. Hence, an individual who engages in prima facie sensational activities may not be a sensation seeker, while another individual who engages in these same activities for the purpose of and with the intent of having a sensational experience is a sensation seeker.

Table 15 reports the agreement of the forty SSS items with three fundamental sensation seeking criteria outlined above. First, an item instantiates a disposition if it

---

Insert Table 15 About Here

---

denotes behavior. In order for an item to qualify, it must be a condition of endorsement of the item that the individual actually engages in an activity on a regular basis. Furthermore, dispositions are enduring characteristics that mature or recede over long periods of time. To have engaged in a sensational activity only once or many years previously does not support the instantiation of a dispositional predicate to such an individual. Second, the individual must actively seek out circumstances that he/she expects will lead to sensational experiences. It is not enough simply to have the experience, it must be actively pursued. Third, the sought after experience or circumstance must either be or be intended to lead to, a *sensational* experience. The distinction between impressions and sensations is maintained in the judgement criteria. Thus, a person who actively seeks out smells of a particular kind is not a sensation seeker, while a person that actively seeks out the sensations associated with a smell experience is.

It can be seen in Table 15 that, in total, only 11 of the 120 possible criterial judgments are supported. In particular, 4 of the items instantiated a disposition, none of the items instantiated seeking behavior and 7 of the items instantiated sensations. None of the 40 items met all three criteria. It is therefore concluded that none of the sensation seeking scale items justify the predicate sensation seeking to an individual who endorses them.

4. Conclusion

It has been shown that the basis for any claim that ξ is a measure of Ψ is grammatical, the grammar of Ψ providing the standard of correctness for measurement claims about Ψ. It has also been shown that construct and criterion validity address empirical issues that are logically irrelevant to grammatical issues. In addition, it was shown that although content validity is a conceptual issue, the accompanying notion of a universe of content presupposes a grammatical basis for measurement claims about Ψ. It follows, therefore, that there can be no justification for the claim that a psychological test constructed in accord with the American Psychological Association standards (APAS) measures the construct it purports to measure. Since the vast majority of tests are constructed in accord with these standards, there is no justification for the claim that any of the vast majority of psychological tests measure the construct they purport to measure. Although this conclusion appears severe, the consequences of this argument for the measurement of psychological phenomena extend beyond the more conventional APA standards.

Concepts that have associated measurement grammars do not consist of two separable components - a measurement grammar and a nonmeasurement grammar. A measurement grammar is a fundamental component of a concept that is the basis for explanations of the concept's meaning. For example, without instruments for the measurement of time and rules for the taking of time measurements, there is no concept of time. And, it is not merely that the meaning of time would be different if it were divested of its measurement grammar. Rather, the new concept would have either no meaning at all or would be a radically different concept. The same applies to attempts to graft a measurement grammar on to an existing concept. A new concept of "depression," for example, that incorporates rules for the taking of depression measurements and units of depression measurement, is not our concept of depression. It is a wholly distinct concept that has no meaning for our current concept.

One might counter that since we do correctly speak of gradations of psychological concepts, they do, in fact, incorporate a measurement grammar. Since we can logically distinguish between severe and mild depression, for example, depression is a concept that incorporates an underlying continuum upon which people can be measured. The measurement of depression, therefore, is a grammatical matter that rests on the criteria of application for the various conceptual gradations of the concept of depression. This response rests on a misunderstanding. When we apply the predicate "severe" to an individual's level of depression, we are not *measuring* depression. Ter Hark (1990) remarks:

> "This confusion occurs for instance in psychology when on the one hand 'thinking' is used in the normal sense of the word and on the other hand is regarded as measurable in terms of (physiological) reactions. Wittgenstein does not say that this kind of measurement is impossible, but only that it involves an *entirely different* phenomenon from what 'thinking' is normally understood to mean. As he puts it 'There are gradations of expecting, but it is nonsense to speak of a measurement of hoping, if one allows 'hope' its use." (p. 32)

While there certainly are grounds for claims to the effect that some psychological concepts have gradations of intensity, the claim that these gradations are *measurements* is logically unsupportable.

i. Axiomatic Measurement Theory

Axiomatic measurement theory provides a formal treatment of the notion that psychological concepts can be measured on gradations of intensity (Krantz, 1991). Let us suppose that a, b, c and d are four sounds and that subjects are exposed to each of the sounds in pairs. The subject's task is to rate the relative loudness of each of the sounds in each of the nxn/2 (in this case, 8) possible pairs. If stimuli a is judged to be louder than stimuli b by a greater amount than stimuli c is judged to be louder than stimuli d we have:

$$(a,b) \geq (c,d) \tag{1}$$

The attempt in axiomatic measurement theory is to find a numerical scale such that:

$$\text{if } (a,b) \geq (c,d) \text{ then } u(a)-u(b) \geq u(c)-u(d) \tag{2}$$

where u is a real valued scale measuring a particular attribute (in this case, loudness). The full set of conditions (axioms) under which a scale u can be found is the measurement theory for the attribute in question. Of course, it is not always possible to find a scale that will satisfy the conditions implied by 2. However, it is expected that if a measurement scale for an attribute exists, careful empirical research of the type described above will lead to its discovery.

A principal feature of axiomatic measurement theory is the attempt to derive numerical measurement scales from natural orderings. Krantz (1991) notes that:

> "It is not very clear just what ontological commitment has been made but in most cases it seems that one expects at least a useful sort of ordering of objects or situations or organisms or social entities. In many cases, this expectation is made explicit by the use

> of words such as "intensity" or "strength" or "level" or "degree" or "extent." (p. 3)

and 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...." (p. 3)

The axiomatic approach, therefore, commits to the notion that there is a natural gradation to many psychological concepts, and that this gradation can provide the basis for measurements of Ψ.

In the first place, the "expectation" of an ordering for psychological concepts is not an "ontological" problem. The basis for an ordering is grammatical. The grammar of psychological concepts provides the grounds for the application of gradation or order predicates. Furthermore, it is not correct to "expect" psychological concepts to have an ordering because it implies that there are no standards of correctness for the application of order predicates to psychological concepts. That is, we cannot determine a priori on the basis of the grammar of psychological concepts that they allow gradations of intensity; we can only "expect" that our *empirical hypothesis* of gradation will hold. As we have seen, there is nothing about the empirical existence of an ordering that can bear on the meaning of a concept. The grammar of concepts is constitutive for empirical investigations of order. And, it is the grammar of concepts that determines whether or not a gradation or ordering is meaningful. Whether there actually exists an ordering of phenomena (whether or not a scale u exists) is, however, an empirical matter. If, for example, it were suddenly the case that the red shift of stars became constant, this would not signify the lack of a red shift measurement scale - the meaning of red shift would not be altered. It would, however, indicate that the universe is not currently expanding or contracting.

Secondly, although it might "seem natural" to move from the presupposition of an ordering to a numerical measurement, it is certainly not always correct to do so. Although a numerical scale, u, exists that satisfies the relation given in Equation 2 for the finishing positions of runners in a race, it is not correct to equate time to completion with finishing position. In fact, the grammar of the concepts, finishing position, and time to completion, provides grounds for the claim that order and finishing time have distinct meanings.

ii. The Act Frequency Approach

Another systematic treatment that offers a rationale for the justification of measurement claims is an approach devised by Buss and Craik (1980) called the act frequency approach (AFA). The AFA speaks to dispositional concepts such as dominance or aggression. The measurement of the level of intensity or gradation of dispositional psychological concepts is based on the frequency with which an individual engages in behaviors that denote the concept of interest. For example, a measurement of dominance is given by the frequency with which an individual engages in dominant behaviors. Maraun (1989) gives a comprehensive description of the AFA and a critique that is based on, among others, the philosophical principles I have used to undermine the logic of construct validity. Essentially, Maraun (1989) concludes that the grammar of dispositional concepts does contain rules for the application of gradation predicates such as severe or mild but does not allow for the application of frequency predicates. This can be seen in the trivial case in which each of two individuals engages in a single dominant act. According to the AFA, each individual has an equal amount of dominance, regardless of the context or type of act. Maraun (1989) illustrates the fallacy of this claim:

> "Similarly, if one person admitted to 'taking charge of things at the committee meeting'... and 'demanding a back rub', while another admitted 'seeking military control of a country' just once, the first person would be

> considered as manifesting more 'dominance' according to the scoring rule of the approach. Clearly the aggregation rule of the approach is not adequate in modelling the way 'dispositional dominance' is instantiated. The <u>true</u> relationship of the frequency tally to the instantiation of dispositional concepts is not one-one.... In fact, the frequency tally is entirely inadequate in capturing the instantiation of dispositional concepts." (p. 67)

Although there are numerous other problems with the AFA, the attempt to represent gradations of dispositional concepts with a simple frequency tally is sufficiently conceptually flawed to render meaningless the results generated by the approach.

The arguments contained in this dissertation are interwoven, complex and often lack an obvious ordering. The consequence is that the basic philosophical principles and measurement related claims can be easily overlooked. The broad relevance of these claims to modern psychology, however, is such that deserve to be given clearly, unambiguously and made readily accessible to direct public scrutiny. I therefore conclude with an attempt to lay bare the essence of the most important of these claims:

1) Measurement is a rule-governed practice.

2) The justification for a measurement claim about Ψ is based in the rules that govern the practice of Ψ measurement (if, in fact, rules exist).
3) Rules are autonomous and non-discoverable.
4) Autonomous, grammatical rules give meaning to empirical results.
5) Empirical results about Ψ presuppose grammatical rules for the application of the concept Ψ.
6) Empirical results have no bearing on the meaning of Ψ.
7) Empirical evidence is a fundamental component of the justification for measurement claims about Ψ in construct validity, axiomatic measurement theory and the act frequency approach.
8) Any measurement claim about Ψ that is grounded in empirical evidence lacks justification.
9) Psychological concepts do not have associated measurement grammars - rules for the measurement of Ψ.
10) Psychological concepts have gradations of intensity but cannot be measured.

Some less fundamental but related claims are:

11) In both the physical and social sciences, the basis for measurement claims is grammatical.
12) The empirical correlation of two variables is irrelevant to what the two variables measure.
13) Latent variable model results are irrelevant to what a

set of items measures.

14) In particular, there is no justification for the claim that a set of unidimensional items measures the same thing.

15) The results of classical test theory are irrelevant to what a set of items measures.

16) In particular, there is no justification for the claim that a set of highly intercorrelated items measures the same thing.

17) The question of what a construct is, can not be addressed by empirical research.

18) What a construct **Ψ** is, is a grammatical matter.

19) Existence and meaning are distinct.

20) Existence is irrelevant to meaning.

21) Meaning is public and shared, not private.

22) Psychological constructs are not 'unobserved', 'hypothetical', 'postulated', 'in need of empirical justification', 'subjectively defined', etc.

23) Latent variables are not 'unobserved', 'hypothetical', 'postulated', 'underlying', etc.

24) The meaning of a term is not the object it denotes.

25) It is incoherent to theorize about the meaning of **Ψ**.

26) It is incoherent to construct empirical hypotheses about the meaning of **Ψ**.

27) Meaning is not discovered.

28) Construct validity conflates meaning and discovery.

29) Psychological tests constructed in accordance with the principles of construct validity lack justification as measurements of the constructs they purport to measure.
30) Psychological tests constructed in accordance with the principles of axiomatic measurement theory lack justification as measurements of the constructs they purport to measure.
31) Psychological tests constructed in accordance with the principles of the AFA lack justification as measurements of the constructs they purport to measure.
32) The meaning of our everyday psychological constructs can not be given by 'Merkmal' definitions.
33) Operationism grounds the meaning of constructs in 'Merkmal' (operational) definitions.
34) The results generated with psychological tests constructed from the facets of an operational definition lack meaning for everyday psychological constructs.

These conclusions face the modern psychologist with a considerable challenge. It is ironic that in an age in which only empirical work is deemed to have any real value, only a logical/philosophical argument can overturn the conclusion that all psychological tests constructed in accordance with the major measurement traditions in modern psychology lack justification as measurements of the constructs they purport to measure.

iii Where Did We Go Wrong: The Problem and Its Solution.

In any broad reading of the history of psychology it would be difficult not to notice that attempts to address the measurement problem have not met with unanimous approval. In fact, some would suggest that the years of quarrelling over measurement problems have not yielded a single fundamental measurement scale (see for example Schonemann, 1994). In no context is this more clear than in attempts to measure the most fundamental of all psychological constructs - intelligence. Against a history of argument, equivocation and stipulation in which it has been claimed that "the proliferation of empirical definitions of intelligence finally reduced many psychologists to the desperate statement: 'Intelligence is what intelligence tests measure'." (Cattell, 1965, p. 357), Jensen, for example, seems unperturbed. Buoyed by the support of the philosophical principles of construct validity, Jensen (1983) maintains that, in the spirit of science, careful empirical investigation will enable us to move closer to an understanding of intelligence. Others do not agree. Schonemann (1994), for example, documents a litany of dissention with measurement practices in modern psychology. He describes methodological problems, political problems, intellectual problems and even problems with the psychological makeup of psychologists. And, inevitably, it seems, whenever such fundamental problems arise in psychology, there is the comparison to physics. Why, it is asked in frustration, does physics not suffer the problems of psychology?

A correct diagnosis of measurement problems in psychology is the ultimate goal of this dissertation. I have argued that the justification for measurement claims in any scientific discipline is based in the normative, autonomous grammar of concepts. To the extent that previous attempts to address measurement problems in psychology have focused on statistical methods, mathematics, hypothesis, theory or the practices of modern psychologists, they have misdiagnosed the problem. Grounded as they are in empirical considerations, these offerings cannot speak to the autonomous grammar upon which the justification for measurement claims rests. The solution is to turn away from empirical investigation and look towards the grammar of our concepts. Only then will we appreciate why intelligence, for example, seems so difficult to measure. Only when we eschew more subtle experiments, sophisticated theorizing and advanced methodologies, will there be the possibility for a clear picture. For no investigation of the scientific practices of psychologists can deliver a surview of our exceedingly complex and widely ramifying psychological concepts.

And what of the comparison of psychology to physics? This much is clear; our psychological concepts are not the technical concepts of an advanced science such as physics, and are not based in a sophisticated calculus such as the concepts of higher mathematics. They were developed in contexts and for purposes largely independent of our everyday, complex forms of life. Psychological concepts, however, have grown and developed from primitive articulations into complex and widely ramifying forms. Consequently, they are tightly interwoven with our own way of life. Unlike the concepts of physics, a surview of psychological concepts can serve to *inform* as well as describe and explain this way of life. The contemporary psychologist's struggle to emulate disciplines with a technical language or complex calculus is admirable but dangerous. If the struggle results in an abandonment of the everyday common-or-garden concepts that inform the way of life that we strive to understand, it is misguided. When it is recognized that an understanding of our own complex concepts cannot be achieved by wrenching them from their proper context and replacing them with entirely distinct technical notions, psychology will have taken its first stride beyond infancy. First, however, there must be a change in "the habits of mind" of modern investigators. Construct validity, the act frequency approach and axiomatic measurement theory must be seen for what they are. It is time to put aside the methodological gimmickry and focus on the formidable, complex conceptual and logical problems that we have inherited. The answer is clear:

> "the confusion and barrenness of psychology is not to be explained by calling it a "young science"; its state is not compatible with that of physics, for instance, in its beginnings....For in psychology there are experimental methods and *conceptual confusion*. The existence of the experimental method makes us think we have the means of solving the problems which trouble us; though problem and method pass one another by." (Wittgenstein, 1953)

References

American Psychological Association. (1985). Standards for educational and psychological tests. Washington, D.C.: American Psychological Association.

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

Baker, G., & Hacker, P. (1982). The grammar of Psychology: Wittgenstein's bermerkungen uber die philosophie der Psychologie. Language and Communiation, 2(3), 227-244.

Bechtoldt, H. (1959). Construct Validity: A critique. American Psychologist, 14, 619-629.

Brown, J. (1985). An introduction to the uses of facet theory. In Canter, D (Ed.), Theory and approaches to social research. New York: Springer-Verlag

Buss, D., & Craik, K. (1980). The frequency concept of disposition: Dominance and prototypically dominant acts. Journal of Personality, 48, 379-392.

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

Crocker, L. M., & Algina, J. (1986). Introduction to classical & modern test theory. New York: CBS College Publishing.

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

Ellett, F. (1985). Psychological terms, logical positivism, and realism: Issues related to construct validation. Educational Theory, 35(3), 273-284.

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

Farley, F. H. (1967). Social desirability and dimensionality in the sensation seeking scale. Acta Psychologica, 26, 89-96.

Guttman, L. (1954). The principal components of scalable attitudes. In P. F. Lazarsfeld (Ed.), Mathematical thinking in the social sciences. Glencoe, Illinois: The Free Press.

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

Guttman, L. (1982). What is not what in theory construction. In R. M. Hauser, D. Mechanic, & A. Haller (Eds.), Social Structure and Behavior. New York: Academic.

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

Hawking, S. (1990) A brief history of time.

Holland, P. (1981). When are item response models consistent with observed data? Psychometrika, 46(1), 79-92.

Holland, P., & Rosenbaum, P. (1986). Conditional association and unidimensionality in monotone latent variable models. The Annals of Statistics, 14(4), 1523-1543.

Jackson, J. S. H., & Maraun, M. (1996). The conceptual validity of empirical scale construction: The case of the sensation seeking scale. In press, Journal of Personality and Individual Differences.

Jensen, A. (1969). How much can we boost IQ and scholastic achievement? Harvard Educational Review, 39, 1-123.

Jensen, A. (1983). The definition of intelligence and factor score indeterminacy. The Behavior and Brain Sciences, 6(2), 313-315.

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

Lakoff, G. (1987). Women, fire, and dangerous things. Chicago: University of Chicago Press.

Lewis, D. (1970). How to define theoretical terms. The Journal of Philosophy, 13, 427-447.

Lingoes, J. T. and Guttman, L. (1967). Nonmetric factor analysis: A rank reducing alternative to linear factor analysis. Multivariate Behavioral Research, 485-505.

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

Maraun, M. (1989). Conceptual and methodological problems in the act frequency approach. Unpublished MA Thesis.

Maraun, M. (1996a). Metaphor taken as math: Indeterminacy in the factor analysis model. In press, *Multivariate Behavioral Research*.

Maraun, M. (1996b). Measurement as a normative practice: Implications of Wittgenstein's philosophy of psychology. Unpublished manuscript.

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

Monahan, J., & Steadman, H. (Eds.) (1994). *Violence and mental disorder: Developments in risk assessment.* Chicago, IL: University of Chicago Press.

Schonemann, P. (1987). Jensen's g: Outmoded theories and unconquered frontiers, In s. Modgil (Ed.), *Arthur Jensen: Consensus and controversy*. London: The Falmer Press.

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

Spearman, C. (1904). General ability, objectively determined and measured. *American Journal of Psychology*, *15*, 201-293.

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

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

Thissen, 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.

Thurstone, L. (1947). Multiple-factor analysis: A development and expansion of the vectors of mind. Chicago: University of Chicago Press.

Waismann, F. (1965). Principles of linguistic philosophy.

Wilson, D., Wood, R. & Gibbons, R. (1991). TESTFACT. Test scoring, item statistics, and item factor analysis [computer Program]. Chicago: Scientific Software International.

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

Wittgenstein, L (1990). Unpublished manuscript (169, p. 71), cited in M. Ter Hark (1990), Beyond the inner and the outer. Dordrecht, The Netherlands: Kluwer academic publishers.

Zuckerman, M. (1971). Dimensions of sensation seeking. Journal of Consulting and Clinical Psychology, 36, 45-52.

Zuckerman, M. (1979). Sensation seeking: Beyond the optimal level of arousal. Lawrence Erlbaum, Hillsdale, N. J.

Zuckerman, M. (1994). Behavioral expressions and biosocial bases of sensation seeking. New York: Cambridge University Press.

Zuckerman, M. (1996). "Conceptual clarification" or confusion in "the study of sensation seeking" by J. S. H. Jackson and M. Maraun. In press, Journal of personality and individual differences.

Zuckerman, M., Eysenck, S. B. G. & Eysenck, H. J. (1978). Sensation seeking in England and America: cross-cultural, age, and sex comparisons. Journal of Consulting and Clinical Psychology, 46, 139-149.

Zuckerman, M., Kolin, E., Price, L. & Zoob, I. (1964). Development of a sensation-seeking scale. Journal of Consulting Psychology, 28, 477-482.

Zuckerman, K. & Link, K. (1968). Construct validity for the sensation-seeking scale. Journal of Consulting and Clinical Psychology, 32, 420-426.

Zwick, R. (1986). Assessing the dimensionality of dichotomous item responses: Theoretical and empirical perspectives. Presented at the annual meeting of the American Educational Research Association.

Table 1

Correlation Matrix of Four Variables That Can Be Described By a One-Dimensional Linear Factor Analysis Model

Variable	1	2	3	4
1	(.16)	.12	.08	.04
2		(.09)	.06	.03
3			(.04)	.02
4				(.01)

Table 2

Summary of Previous Psychometric Analyses of The Sensation Seeking Scale

Study	1	2	3	4	5	6	7	8	9	10
Year	64	67	68	71	75	78	78	80	83	86
First	Zuc	Far	Zuc	Zuc	Sew	Zuc	Zuc	Rid	Bal	Row
Form	1	1	1	3	4	4	5	5	5	5
Subjects:										
Males	268	100	268	201	78	414	97	181	363	299
Fem	277	0	277	223	78	865	97	155	335	439
Analyses:										
# item	50	26	50	113	74	71	40	40	40	40
P.C.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No
F.A.	No	No	No	No	No	No	No	No	Yes	Yes
Rotate	No	No	Yes	Yes	Yes	Yes	Yes	Na	Yes	Yes
C.F.A.	No	No	No	No	No	No	No	No	No	Yes
Matrix	Tet	Tet	Tet	PM	?	PM	PM	Na	?	?
Reliability:										
S/H	Yes	No	No	Yes	No	Yes	No	No	No	No
Alpha	No	No	No	No	No	Yes	No	Yes	No	No
Kr-20	No	No	No	No	No	No	Yes	No	No	Yes

Table 2, Con't

Summary of Previous Psychometric Analyses of The Sensation Seeking Scale

Note: Zuc = Zuckerman, M.; Far = Farley, F. H.; Sew = Stewart, D. W.; Rid = Ridgeway, D.; Bal = Ball, I. L.; Row = Rowland, G. L.; P.C. = Holzinger's principal components analysis; F.A. = maximum likelihood factor analysis; C.F.A. = confirmatory maximum likelihood factor analysis performed with Lisrel IV; Tet = Tetrachoric, PM = Product moment, S/H = Split Half and Alpha = Cronbach's alpha.

Table 3

Frequency Distributions of Sensation Seeking Scale Items

Subscale	Item	Option Frequency	
		A	B
DIS	1	518	409
BS	2	818	109
TAS	3	465	462
ES	4	614	313
BS	5	123	804
ES	6	461	466
BS	7	668	259
BS	8	360	567
ES	9	255	672
ES	10	757	170
TAS	11	317	610
DIS	12	355	572
DIS	13	674	253
ES	14	566	361
BS	15	535	392
TAS	16	669	258
TAS	17	634	293
ES	18	468	459
ES	19	712	215
TAS	20	269	658

Table 3, Con't

Frequency Distributions of Sensation Seeking Scale Items

Subscale	Item	Option Frequency A	Option Frequency B
TAS	21	327	600
ES	22	363	564
TAS	23	601	326
BS	24	357	570
DIS	25	357	570
ES	26	538	389
BS	27	463	464
TAS	28	443	484
DIS	29	408	519
DIS	30	733	194
BS	31	681	246
DIS	32	361	566
DIS	33	482	445
BS	34	300	627
DIS	35	386	541
DIS	36	188	739
ES	37	399	528
TAS	38	394	533
BS	39	341	586

Table 3, Con't

Frequency Distributions of Sensation Seeking Scale Items

Subscale	Item	Option Frequency	
		A	B
TAS	40	367	560

Table 4

Means and Standard Deviations for The Sensation Seeking Scale Full Scale and Four Subscales For The Present Sample and For The Four Samples Assessed By Zuckerman et al (1978)

Scale	Present Sample		Zuckerman (1978)	
	Mean	Standard Deviation	Mean	Standard Deviation
Full	17.80	6.75	16.6-21.6	5.7-7.2
TAS	6.23	2.71	5.6- 7.8	2.3-3.0
BS	3.18	2.01	2.8- 3.8	1.9-2.5
ES	4.02	2.02	4.1- 5.0	2.0-2.2
DIS	4.37	2.64	4.1- 6.2	2.6-3.1

Table 5

Means, Standard Deviations and Minimum and Maximum Values of The Pearson Correlations Between the Items of The Sensation Seeking Scale Full Scale and Four Subscales

Scale	Mean	SD	Min	Max	% < 0	% < -.07
Full	.11	.11	-.12	.59	13.2	2.80
TAS	.25	.01	.07	.59	0	.00
BS	.11	.06	-.01	.25	2.2	.00
ES	.10	.11	-.03	.58	13.3	.00
DIS	.24	.01	.03	.50	0	.00

Note: With N=927, r < -.07 is significant at P = .025

Table 6

Percentage of Mantel-Haenzsel Statistics Less Than Particular Values of The Standard Normal Distribution For The Sensation Seeking Full Scale and Four Subscales

Scale	Value of Z					
	-4	-3	-2	-1	0	-2.33
Full	0.3	2.9	9.2	21.5	39.6	6.2
TAS	0.0	0.0	0.0	2.2	6.7	0.0
BS	0.0	0.0	0.0	0.0	13.3	0.0
DIS	0.0	0.0	0.0	2.2	6.7	0.0
ES	0.0	0.0	4.4	17.8	31.1	0.0
Z	0.0	0.1	2.3	15.9	50.0	1.0

Table 7

Regression Parameters for Simulated Response Patterns Based on 1- and 2-Dimensional Normal Ogive Item Response Models

Model Code	Regression Parameters
	40 Item Full Scale Simulations
	One-Dimensional Constant Slopes
.0	40(0)
.1	40(.1)
.3	40(.3)
.5	40(.5)
.7	40(.7)
.9	40(.9)
	One-Dimensional Variable Slopes
.0-.1	20(0), 20(.1)
.0-.3	10(0), 10(.1), 10(.2), 10(.3)
.0-.5	7(0), 7(.1), 5(.2), 7(.3), 7(.4), 7(.5)
.0-.7	5(0), 5(.1), 5(.2),., 5(.5), 5(.6), 5(.7)
.0-.9	4(0), 4(.1), 4(.2),., 4(.7), 4(.8), 4(.9)

Table 7, Con't

Regression Parameters for Simulated Response Patterns Based on 1- and 2-Dimensional Normal Ogive Item Response Models

Model Code	Regression Parameters	
	One-Dimensional Variable Positive and Negative Slopes	
-.1-.1	14(-.1), 12(0), 14(.1)	
-.3-.3	6(-.3), 6(-.2),., 4(0),., 6(.2), 6(.3)	
-.5-.5	4(-.5), 4(-.4),., 0(0),., 4(.4), 4(.5)	
-.7-.7	3(-.7),., 2(-.1), 0(0), 2(.1),., 3(.7)	
-.9-.9	2(-.9), 2(-.8),., 4(0),., 2(,.8),2(.9)	
	Two-Dimensional Positive Slopes	
	Dimension 1	Dimension 2
.0-.1	20(0), 20(.1)	20(.1), 20(0)
.0-.3	20(0), 20(.3)	20(.3), 20(0)
.0-.5	20(0), 20(.5)	20(.5), 20(0)
.0-.7	20(0), 20(.7)	20(.7), 20(0)
.0-.9	20(0), 20(.9)	20(.9), 20(0)

Table 7, Con't

Regression Parameters for Simulated Response Patterns Based on 1- and 2-Dimensional Normal Ogive Item Response Models

Model Code	Regression Parameters	
Two-Dimensional Positive and Negative Slopes		
	Dimension 1	Dimension 2
-.1-.1	20(-.1), 20(.1)	20(.1), 20(-.1)
-.3-.3	20(-.3), 20(.3)	20(.3), 20(-.3)
-.5-.5	20(-.5), 20(.5)	20(.5), 20(-.5)
-.7-.7	20(-.7), 20(.7)	20(.7), 20(-.7)
-.9-.9	20(-.9), 20(.9)	20(.9), 20(-.9)
10 Item Subscale Simulations		
One-Dimensional Constant Slopes		
.0	10(0)	
.1	10(.1)	
.3	10(.3)	
.5	10(.5)	
.7	10(.7)	

Table 7, Con't

Regression Parameters for Simulated Response Patterns Based on 1- and 2-Dimensional Normal Ogive Item Response Models

Model Code	Regression Parameters
.9	10(.9)
	One-Dimensional Variable Slopes
.0-.1	5(0), 5(.1)
.0-.3	3(0), 2(.1), 2(.2), 3(.3)
.0-.5	2(0), 2(.1), 1(.2), 1(.3), 2(.4), 2(.5)
.0-.7	2(0), 1(.1), 1(.2),., 1(.5), 1(.6), 2(.7)
.0-.9	1(0), 1(.1), 1(.2),., 1(.7), 1(.8), 1(.9)
	One-Dimensional Variable Positive and Negative Slopes
-.1-.1	4(-.1), 2(0), 4(.1)
-.3-.3	2(-.3), 2(-.2), 1(-.1), 0(0), 1(.1), 2(.2), 3(.3)
-.5-.5	1(-.5), 1(-.4), 1(-.3),., 1(.3), 1(.4), 1(.5)
-.7-.7	1(-.7), 1(-.5),., 2(0),., 1(-.5), 1(-.7)
-.9-.9	1(-.9), 1(-.7),., 1(-.1), 1(.1),., 1(.7), 1(.9)

Table 7, Con't

Regression Parameters for Simulated Response Patterns Based on 1- and 2-Dimensional Normal Ogive Item Response Models

Model Code	Regression Parameters	
	Two-Dimensional Positive Slopes	
	Dimension 1	Dimension 2
.0-.1	5(0), 5(.1)	5(.1), 5(0)
.0-.3	5(0), 5(.3)	5(.3), 5(0)
.0-.5	5(0), 5(.5)	5(.5), 5(0)
.0-.7	5(0), 5(.7)	5(.7), 5(0)
.0-.9	5(0), 5(.9)	5(.9), 5(0)
	Two-Dimensional Positive and Negative Slopes	
	Dimension 1	Dimension 2
-.1-.1	5(-.1), 5(.1)	5(.1), 5(-.1)
-.3-.3	5(-.3), 5(.3)	5(.3), 5(-.3)
-.5-.5	5(-.5), 5(.5)	5(.5), 5(-.5)
-.7-.7	5(-.7), 5(.7)	5(.7), 5(-.7)

Table 7, Con't

Regression Parameters for Simulated Response Patterns Based on 1- and 2-Dimensional Normal Ogive Item Response Models

Model Code	Regression Parameters	
-.9-.9	5(-.9), 5(.9)	5(.9), 5(-.9)

Table 8

Percentage of Mantel-Haenzsel Statistics Less Than Particular Values of The Standard Normal Distribution For Simulated Data Based on A Forty-Item One-Dimensional Normal Ogive Model With Constant, Variable Positive And Variable Positive and Negative Item Slopes

Slopes	Value of Z					
	-4	-3	-2	-1	0	-2.33
			Full Scale and Z			
Full	0.3	2.9	9.2	21.5	39.6	6.2
Z	0.0	0.1	2.3	15.9	50.0	1.0
			Constant Slopes			
.0	0.0	0.3	1.9	13.5	44.2	1.3
.1	0.0	0.1	1.4	11.4	37.4	0.6
.3	0.0	0.0	0.9	4.7	24.2	0.3
.5	0.0	0.0	0.1	2.8	17.6	0.1
.7	0.0	0.0	0.3	3.7	21.2	0.1
.9	0.0	0.0	0.6	2.7	21.2	0.1
			Variable Positive Slopes			
.0-.1	0.0	0.1	1.7	13.2	42.7	0.8

Table 8, Con't

Percentage of Mantel-Haenzsel Statistics Less Than Particular Values of The Standard Normal Distribution For Simulated Data Based on A Forty-Item One-Dimensional Normal Ogive Model With Constant, Variable Positive And Variable Positive and Negative Item Slopes

Scale	Value of Z					
	-4	-3	-2	-1	0	-2.33
.0-.3	0.0	0.3	1.3	8.6	32.1	0.4
.0-.5	0.0	0.0	0.9	7.8	30.9	0.5
.0-.7	0.0	0.1	0.5	6.2	23.5	0.4
.0-.9	0.0	0.0	0.8	5.9	24.7	0.5
	Variable Positive and Negative Slopes					
-.1-.1	0.6	0.1	1.9	14.1	43.3	0.6
-.3-.3	0.0	0.9	5.1	17.4	44.5	3.3
-.5-.5	3.3	8.1	17.1	30.5	47.3	13.7
-.7-.7	16.3	21.3	29.2	39.6	48.8	26.9
-.9-.9	18.5	22.6	27.6	35.6	48.8	25.9

Table 9

Percentage of Mantel-Haenzsel Statistics Less Than Particular Values of The Standard Normal Distribution For Simulated Data Based on A Forty-Item Two-Dimensional Normal Ogive Model With Constant Positive and Constant Positive and Negative Item Slopes

Slopes	Value of Z					
	-4	-3	-2	-1	0	-2.33
			Full Scale and Z			
Full	0.3	2.9	9.2	21.5	39.6	6.2
Z	0.0	0.1	2.3	15.9	50.0	1.0
			Constant Positive Slopes			
.0-.1	0.0	0.1	1.2	10.0	35.5	0.8
.0-.3	0.0	0.3	2.1	12.9	36.2	1.0
.0-.5	0.4	4.4	17.6	36.8	47.8	11.2
.0-.7	28.1	42.4	49.9	50.9	51.3	48.3
.0-.9	51.3	51.3	51.3	51.3	51.3	51.3
			Constant Positive and Negative Slopes			
-.1-.1	0.0	0.3	1.9	13.6	4.2	1.2
-.3-.3	8.3	24.2	40.6	48.3	51.4	36.5

Table 9, Con't

Percentage of Mantel-Haenzsel Statistics Less Than Particular Values of The Standard Normal Distribution For Simulated Data Based on A Forty-Item Two-Dimensional Normal Ogive Model With Constant Positive and Constant Positive and Negative Item Slopes

Slopes	Value of Z					
	-4	-3	-2	-1	0	-2.33
-.5-.5	48.7	48.7	49.0	49.2	51.7	49.0
-.7-.7	48.7	48.7	48.7	48.8	50.5	48.7
-.9-.9	48.7	48.7	47.7	48.7	50.0	48.7

Table 10

Percentage of Mantel-Haenzsel Statistics Less Than Particular Values of The Standard Normal Distribution For Simulated Data Based on A Ten-Item One-Dimensional Normal Ogive Model With Constant, Variable Positive And Variable Positive and Negative Item Slopes

Slopes	Value of Z					
	-4	-3	-2	-1	0	-2.33
			Sub Scales and Z			
TAS	0.0	0.0	0.0	2.2	6.7	0.0
BS	0.0	0.0	0.0	0.0	13.3	0.0
DIS	0.0	0.0	0.0	2.2	6.7	0.0
ES	0.0	0.0	4.4	17.8	31.1	0.0
Z	0.0	0.1	2.3	15.9	50.0	1.0
			Constant Slopes			
.0	0.0	0.0	2.2	20.0	48.9	0.0
.1	0.0	0.0	2.2	17.8	32.2	0.0
.3	0.0	0.0	0.0	0.0	13.3	0.0
.5	0.0	0.0	0.0	0.0	0.0	0.0
.7	0.0	0.0	0.0	0.0	0.0	0.0
.9	0.0	0.0	0.0	0.0	0.0	0.0

Table 10, Con't

Percentage of Mantel-Haenzsel Statistics Less Than Particular Values of The Standard Normal Distribution For Simulated Data Based on A Ten-Item One-Dimensional Normal Ogive Model With Constant, Variable Positive And Variable Positive and Negative Item Slopes

Scale	Value of Z					
	-4	-3	-2	-1	0	-2.33
			Variable Positive Slopes			
.0-.1	0.0	0.0	4.4	17.8	42.2	0.0
.0-.3	0.0	0.0	0.0	13.3	33.3	0.0
.0-.5	0.0	0.0	0.0	8.9	22.2	0.0
.0-.7	0.0	0.0	0.0	11.1	17.8	0.0
.0-.9	0.0	0.0	0.0	8.9	11.1	0.0
			Variable Positive and Negative Slopes			
-.1-.1	0.0	0.0	2.2	24.4	53.3	0.0
-.3-.3	0.0	0.0	4.4	26.7	53.3	4.4
-.5-.5	4.4	11.1	17.8	33.3	55.6	15.6
-.7-.7	11.1	11.1	13.3	31.1	60.0	13.3
-.9-.9	24.4	31.1	37.8	42.2	57.8	35.6

Table 11

Percentage of Mantel-Haenzsel Statistics Less Than Particular Values of The Standard Normal Distribution For Simulated Data Based on A Ten-Item Two-Dimensional Normal Ogive Model With Constant Positive and Constant Positive and Negative Item Slopes

Slopes	Value of Z					
	-4	-3	-2	-1	0	-2.33
Sub Scales and Z						
TAS	0.0	0.0	0.0	2.2	6.7	0.0
BS	0.0	0.0	0.0	0.0	13.3	0.0
DIS	0.0	0.0	0.0	2.2	6.7	0.0
ES	0.0	0.0	4.4	17.8	31.1	0.0
Z	0.0	0.1	2.3	15.9	50.0	1.0
Constant Positive Slopes						
.0-.1	0.0	0.0	2.2	20.0	48.9	2.2
.0-.3	0.0	0.0	4.4	8.9	26.7	0.0
.0-.5	0.0	0.0	4.4	26.7	46.7	0.0
.0-.7	6.7	33.3	51.1	55.6	55.6	44.4
.0-.9	55.6	55.6	55.6	55.6	55.6	55.6

Table 11, Con't

Percentage of Mantel-Haenzsel Statistics Less Than Particular Values of The Standard Normal Distribution For Simulated Data Based on A Forty-Item Two-Dimensional Normal Ogive Model With Constant Positive and Constant Positive and Negative Item Slopes

Slopes	Value of Z					
	-4	-3	-2	-1	0	-2.33
			Constant Positive and Negative Slopes			
-.1-.1	0.0	0.3	2.2	20.0	60.0	2.2
-.3-.3	17.8	40.0	46.7	55.6	55.6	46.7
-.5-.5	55.6	55.6	55.6	55.6	55.6	55.6
-.7-.7	55.6	55.6	55.6	55.6	55.6	55.6
-.9-.9	55.6	55.6	55.6	55.6	55.6	55.6

Table 12

Pearson Correlations Between The Sensation Seeking Scale Full Scale and Subscales

Scale	TAS	BS	DIS	ES	Full
TAS	1				
BS	.20	1			
DIS	.36	.50	1		
ES	.39	.24	.40	1	
Full	.72	.64	.81	.69	1

Note: All correlations are significant at $p = .01$

Table 13

Conditional Correlations Between Subscale Totals

Subscale Pair	Total Score 0	1	2	3	4	5	6	7	8	9	10
Conditioned on TAS Total Scores											
ES/BS	.40	-.02	.10	.11	.28	.28	.19	.08	.30	.15	.20
ES/DIS	.40	.33	.16	.40	.27	.31	.34	.26	.37	.26	.32
BS/DIS	.59	.44	.51	.51	.43	.53	.39	.41	.57	.41	.48
Conditioned on DIS Total Scores											
ES/BS	.10	.23	.09	.17	.14	-.01	.01	-.02	-.14	-.15	.27
ES/TAS	.43	.40	.20	.31	.27	.31	.27	.16	.27	.39	.38
BS/TAS	.12	.29	-.05	.01	.00	-.01	-.01	.00	-.09	-.10	.26
Conditioned on BS Total Score											
ES/TAS	.40	.41	.34	.45	.32	.30	.46	.29	.14	-.98	*
ES/DIS	.32	.52	.38	.40	.22	.33	.22	.15	.13	.42	*
TAS/DIS	.49	.30	.35	.39	.13	.40	.41	.09	.29	-.27	*

Table 13, Con't

Conditional Correlations Between Subscale Totals

Subscale Pair	Total Score 0	1	2	3	4	5	6	7	8	9	10
					Conditioned on ES Total Score						
BS/TAS	.20	.27	.09	.12	.23	.08	-.13	-.03	.08	-.18	.08
BS/DIS	.24	.29	.32	.28	.26	.22	.13	.03	.36	-.03	.82
TAS/DIS	.39	.46	.50	.53	.46	.46	.34	.43	.26	-.12	-.5

Note: * indicates that a correlation could not be calculated due to there being no cases with that subscale total score.

Table 14

Sensation Seeking Scale Items

Subscale	Item #	Item Content
TAS	3	I often wish I could be a mountain climber
TAS	11	I sometimes like to do things that are a little frightening
TAS	16	I would like to take up the sport of water skiing
TAS	17	I would like to try surfboard riding
TAS	20	I would like to learn to fly an airplane
TAS	21	I would like to go scuba diving
TAS	23	I would like to try parachute jumping
TAS	28	I like to dive off the high board
TAS	38	I would like to sail a long distance in a small but seaworthy sailing craft
TAS	40	I think I would enjoy the sensations of skiing very fast down a high mountain slope
ES	4	I like some of the earthy body smells

Table 14, Con't

Sensation Seeking Scale Items

Subscale	Item #	Item Content
ES	6	I like to explore a strange city or section of town myself, even if it means getting lost
ES	9	I have tried marijuana or would like to
ES	10	I would like to try some of the new drugs that produce hallucinations
ES	14	I like to try new foods that I have never tasted before
ES	18	I would like to take off on a trip with no preplanned or definite routes or timetables
ES	19	I would like to make friends in some of the "far-out" groups like artists or "hippies"
ES	22	I would like to meet some persons who are homosexual (men or women)
ES	26	I often find beauty in the "clashing" colours and irregular form of modern painting

Table 14, Con't

Sensation Seeking Scale Items

Subscale	Item #	Item Content
ES	37	People should dress in individual ways even if the effects are sometimes strange
DIS	1	I like wild "uninhibited" parties
DIS	12	I enjoy the company of real swingers
DIS	13	I often like to get high (drinking liquor or smoking marijuana)
DIS	25	I like to have new and exciting experiences and sensations even if they are a little unconventional or illegal
DIS	29	I like to date members of the opposite sex who are physically exciting
DIS	30	Keeping the drinks full is the key to a good party
DIS	32	A person should have considerable sexual experience before marriage

Table 14, Con't

Sensation Seeking Scale Items

Subscale	Item #	Item Content
DIS	33	I could conceive of myself seeking pleasures around the world with the "jet set"
DIS	35	I enjoy watching many of the "sexy" scenes in movies
DIS	36	I feel best after taking a couple of drinks
BS	2	I can't stand watching a movie that I have seen before
BS	5	I get bored seeing the same old faces
BS	7	When you can predict almost everything a person will do and say, he or she must be a bore
BS	8	I usually don't enjoy a movie or a play where I can predict what will happen in advance
BS	15	Looking at someone's home movies or travel slides bores me tremendously
BS	24	I prefer friends who are excitingly unpredictable

Table 14, Con't

Sensation Seeking Scale Items

Subscale	Item #	Item Content
BS	27	I get very restless if I have to stay around home for any length of time
BS	31	The worst social sin is to be a bore
BS	34	I like people who are sharp and witty even if they do sometimes insult others
BS	39	I have no patience with dull or boring persons

Table 15

Correspondence of Sensation Seeking Scale Items With Sensation Seeking Criteria

Scale	Item #	Criteria		
		Dispositional	Seeking	Sensation
TAS	3	No	No	No
TAS	11	Yes	No	No
TAS	16	No	No	No
TAS	17	No	No	No
TAS	20	No	No	No
TAS	21	No	No	No
TAS	23	No	No	No
TAS	28	No	No	No
TAS	38	No	No	No
TAS	40	No	No	No
ES	4	No	No	No
ES	6	No	No	No
ES	9	No	No	No
ES	10	No	No	Yes
ES	14	No	No	No
ES	18	Yes	No	No
ES	19	No	No	No
ES	22	No	No	No
ES	26	Yes	No	No

Table 15, Con't

<u>Correspondance of Sensation Seeking Scale Items With Sensation Seeking Criteria</u>

Scale	Item #	Criteria		
		Dispositional	Seeking	Sensation
ES	37	No	No	No
DIS	1	No	No	No
DIS	12	No	No	No
DIS	13	Yes	No	Yes
DIS	25	No	No	Yes
DIS	29	No	No	No
DIS	30	No	No	No
DIS	32	No	No	No
DIS	33	No	No	No
DIS	35	No	No	No
DIS	36	No	No	Yes
BS	2	No	No	No
BS	5	No	No	Yes
BS	7	No	No	No
BS	8	No	No	No
BS	15	No	No	Yes
BS	24	No	No	Yes
BS	27	No	No	No
BS	31	No	No	No

Table 15, Con't

Correspondance of Sensation Seeking Scale Items With Sensation Seeking Criteria

Scale	Item #	Criteria		
		Dispositional	Seeking	Sensation
BS	34	No	No	No
BS	39	No	No	No