THE LOGIC OF TEST ANALYSIS: AN EVALUATION OF TEST THEORY AND A PROPOSED LOGIC FOR TEST ANALYSIS

by

Kathleen L. Slaney

B.A., Simon Fraser University, 1994

M.A., Simon Fraser University, 2001

Dissertation Submitted in Partial Fulfilment of the Requirements for the Degree of

Doctor of Philosophy

in the Department

of

Psychology

SIMON FRASER UNIVERSITY

Spring 2006

APPROVAL

Name: Kathleen L. Slaney

Degree: Doctor of Philosophy (Psychology)

Title of Thesis: The Logic of Test Analysis: An Evaluation of Test Theory and a Proposed Logic for Test Analysis

Examining Committee:

Chair: **Dr. Cathy McFarland**
Professor

Dr. Michael Maraun Senior Supervisor Associate Professor, Department of Psychology	**Dr. William Krane** Associate Professor/Associate VP Academic
Dr. Rachel Fouladi Assistant Professor, Department of Psychology	**Dr. Bruno Zumbo** Internal Examiner Professor University of British Columbia
Dr. Jack Martin Professor, Faculty of Education	**Dr. James Steiger** External Examiner Professor Vanderbilt University

Date Defended: March 9, 2006

ABSTRACT

Despite the rich and abundant body of test theoretic results that have accumulated over the past 100 years, little work has been done on the development of a coherent *framework* for the carrying out of test *analyses*, resulting in a general state of test analytic practice which is unsystematic, unreasoned, and piecemeal. The current work was guided by two primary aims: 1) to document the apparent gulf that exists between the advances that have been made in test theory, and the seemingly calamitous and unmethodical state of current test analytic practice, and 2) to rescue applied test analytic practice from its ill-defined state by deriving a logical, sequential framework for the carrying out of test analyses within which the tools of test theory can be used to full advantage. To serve these aims, the historical landmarks of 100 years of test theory were documented. The mathematical foundations of two relatively distinct theoretical test theoretic perspectives, viz., *classical test theory* and *modern test theory*, were summarized in axiomatic fashion. Articles from five peer-reviewed journals were examined with the aim of gaining further insight into the current state of test analytic practices. Finally, the components of the proposed framework for analyzing tests were fully explicated, and certain current test analytic practices critiqued in light of the proposed framework.

Keywords: Test theory, classical test theory, modern test theory, test analysis, test analytic framework, validity, reliability, test theory model, construct validation, test performance, rule-guided practice

DEDICATION

This thesis is dedicated to the memories of all great scholars, to vast in number to name, but whose work continues to inspire, challenge, and provoke young academics to dedicate themselves to the scholarly endeavour.

ACKNOWLEDGEMENTS

I thank my committee members, Dr. William Krane and Dr. Rachel Fouladi, for their patience and their valuable input, and Dr. Jack Martin, a true mentor, for his kind and thorough feedback, and his invaluable advice on scholarly matters. In addition, I also thank Joan Wolfe and Penny Simpson for taking the time to assist in the formatting of this document, and Lynn Kumpula for her diligence in overseeing all the details.

I thank my parents, Vern and Rosemary Slaney, and Sharon Slaney, each for their unwavering support in what has turned out to be a very long and drawn out education. Thanks also go to my siblings, Colleen, Erin, and Mike, for their frequent words of encouragement, and to each of those extended family members and friends who has been kind enough to lend support throughout the duration of this work.

There is no adequate way to express in words my gratitude to my senior supervisor, Dr. Michael Maraun, without whose vital assistance, patient guidance, and steadfast support this dissertation would not have seen the light of day. He has been a mentor of superior quality, but, also, a loyal and supportive friend, and in him I have found a kindred spirit.

Finally, my most heartfelt thanks go to my son, Toby, whose enormous spirit inspires me beyond words, and to my husband, Andrew, whose love, loyalty, and support apparently know no bounds.

TABLE OF CONTENTS

1. INTRODUCTION

The history of testing and measurement in psychological science is a rich one, dating back to the era of psychophysics, and the birth of psychology as a science in its own right. It encompasses a number of different areas within the discipline, and is encountered, at least to some extent, by all engaged in empirical research. It would indeed be unusual to find a psychological researcher unfamiliar with such concepts as *reliability, validity, measure, scale,* and so on. Yet, it appears, at least on the surface of things, that few researchers in psychology could adequately articulate what exactly is involved in test analysis or anything about the theory (or theories) which provides its mathematical grounding. In fact, despite the abundance of concepts and quantities associated with test *theory,* it appears that there exists little consensus as to how they should be employed in test *analyses.*

One source of this confusion stems from the fact that, broadly speaking, there exist two relatively distinct test theories: C*lassical test theory* (CTT) and *modern test theory* (MTT). CTT consists primarily in a collection of indices and techniques pertaining to the assessment of the "reliability" and "validity" of a test. These indices and techniques are unified by the *classical true score* model

according to which an observed test score for a given individual is conceptualized as decomposable into two components: the individual's "true score" and an error component, the latter of which represents the degree of imprecision of measurement.

Within MTT, observed responses to the individual items of a test are conceptualized as "manifestations" or "indicators" of an unobservable attribute of interest, which, in MTT jargon is known as the "latent trait". At the epicentre of MTT is the employment of latent variable models, each of which specifies the mathematical form of the item/latent trait regressions (i.e., the "item characteristic curves", or "item response functions"). From each of such models, particular implications may be drawn and tested on the basis of a sample of responses to the set of items of which a test of some attribute consists. If the data are shown to conform to the model, then optimal compositing rules and estimates of precision may be derived directly from the model.

Although the differences between CTT and MTT are well understood (cf. Blinkhorn, 1997; Lumsden, 1976; McDonald, 1999; Weiss and Davison, 1981), it is apparent that there remains a great deal of confusion with regard to how the mathematical tools generated by CTT and MTT should be employed in applied test analysis (i.e., in the passing of judgment on the quality of a test). This confusion would seem to chiefly be the result of two factors: 1) Applied test analyzers very often misunderstand the mathematical products of CTT and MTT,

a point that has been made frequently enough in the literature (see, e.g., Hattie, 1981, 1984, 1985); and 2) despite advances in the sophistication of the mathematical tools available to the applied test analyst, little work has been done on the development of a coherent logical *framework* for the carrying out of test analyses. It is, then, not surprising that the applied test analyses found in the literature of the social sciences are almost uniformly unsystematic, unreasoned, and piecemeal.

The over-riding aims of this work are as follows: 1) to document the perplexing gulf that exists between the notable advances that have been made in test theory, represented, in particular, by the movement from CTT to MTT, and the rather primitive state of current test analytic practice[1]; and 2) to rescue applied test analytic practice from its primitive state by deriving a logical, sequential framework for the carrying out of test analyses within which the impressive tools of CTT and MTT can be used to full advantage. To realize this aim, the following topics will be addressed:

A. Chapter Two explicates in detail the historical path of test theory from its formal inception in the work of Charles Spearman at the turn of the 20th century, through the birth of modern test theory in the 1950's, to current

[1] Note that the emphasis here is placed on test analysis as it applies to the evaluation of pre-existing measures and not on the development of instruments, nor on various applications of modern test theory principles such as item calibration, test equating, and tailored testing, each of which will be given brief mention in Chapter 3.

developments regarding the application of latent variable modelling to test data. Chapter Three provides a summary of the mathematical foundations of both CTT and MTT.

B. Chapter Four provides a snap-shot of the state of current test analytic practice. It contains the findings from a systematic examination of research studies published over a specified time period in a sample of peer-reviewed journals in which test analyses frequently appear. These articles were examined for the following: 1) whether the aim of the analysis was identified; 2) whether the researcher was explicit with regard to how many attributes the test is expected to measure, and whether this feature of the test was assessed (i.e., the "dimensionality" or "structure" of the test was examined), in particular with some statistical model; 3) whether the items of the analyzed test were composited and, if so, what the was the nature of the compositing rule used to create this composite; 4) which, if any, indices were used to estimate the "reliability" (or more generally, the precision) of the test; 5) whether and how the issue of validity was handled; 6) whether the analyses (if any) appeared to be guided by an explicit logic.

C. Although there exists a well acknowledged distinction between the two test theories, CTT and MTT, the distinction between a test theory and a *test analytic framework* seems to have somehow been sublimated. The mere

existence of sophisticated test theory does not imply the existence of sophisticated test analytic practice, as the findings of Chapter Four make clear. The beginning of Chapter Five includes an elucidation of this essential distinction.

D. The remainder of Chapter Five will propose, and explicate, a logical, sequential framework for test analysis. This framework is comprised of the following components: 1) specification of the theoretical structure of the test to be analyzed; 2) choice of a (unidimensional) test theory model that squares, or is in keeping, with the theoretical structure; 3) a test of conformity of the joint distribution of the items of the test to the chosen test theory model; 4) conditional on the conformity of the distribution of the items to the test theory model, the derivation of a model-implied compositing rule for the test items; 5) the estimation of the reliability of the resulting composite of test items; 6) conditional on the composite possessing adequate reliability, the entering of the composite into "external" construct validation studies (e.g., multi-trait, multi-method analyses, general explorations of the test's place in the nomological network of the attribute it was designed to measure, etc.). This proposed framework will be illustrated through a number of examples.

E. Finally, in Chapter Six certain of the more prominent confusions inherent to current applied test analytic practice will be catalogued and discussed in light of the proposed test analytic framework.

Throughout this work, certain concepts and symbols will be employed repeatedly. In general terms, the test analytic context may be described as follows: A test, T, is a collection of *stimulus materials* and *response options*, plus a set of *scoring rules* that convert the responses of a respondent to the stimulus materials, as encoded by the response options, into a set of real numbers (scores). The stimulus materials of the objective tests standardly employed in the social and behavioural sciences are a set of k test items, with each item comprised of a content *stem* and set of response options (cf. McDonald, 1999). This is the case that is treated in the current work. The k items of a given test, T, were designed to be indicators of an attribute, γ, whose measurement[2] by T is of interest. The aim is to employ test T to yield measurements of the γ of the individuals who are

[2] The issue of measurement cannot be separated from an examination of test analytic practices in psychology. Indeed, most general treatments of test theory and analysis begin by defining measurement (as "the assignment of numbers to specific empirical manifestations of an attribute", or something to that effect), and providing a classification scheme of particular levels of measurement (usually with reference to Steven's *nominal, ordinal, interval,* and *ratio* levels of measurement, cf. Stevens, 1946). Although such definitions have not been universally accepted among psychological researchers (see Krantz, 1991; Michell, 1990, 1999), they are commonly adopted in most formal treatments of psychological measurement. Although the concept of "psychological measurement" is, in my opinion, in dire need of an overhaul, such an undertaking is certainly beyond the scope of the present work, and, thus, I will remain relatively agnostic to measurement matters insofar as I will not explicitly examine the concept of measurement, nor will I debate the issue of whether psychological phenomena can, in fact, even be measured. Instead, in the present work it will be assumed that the test analyst has before her quantitative representations of attributes, properties, etc., and that these data have, at least, ordinal properties.

the elements of some focal population **P**. The responding of the individuals in population **P** to the k items that comprise test T is called *test behaviour*.

When the individuals of population **P** are given test T, each individual is exposed to each item stem, his response to each stem is encoded in a set of response options, and an associated scoring rule converts the result into a number. The result is then a single score for each individual on each item. The symbol X_j will represent the collection of scores (over individuals in population **P**) to item j, and $\underline{X}$, a vector containing the k random variates X_j, $j = 1, 2, ..., k$. Any scalar function, $\phi = f(\underline{X})$, of the random vector $\underline{X}$ will be called a test (item) *composite* or *metric*. The marginal and joint distributions, in population **P**, of the X_j, $j = 1, 2, ..., k$, $\underline{X}$, any $\phi = f(\underline{X})$, and any other random quantities, will, as is usual, be specified by density functions. An analysis of the performance of test T is empirical in nature, and focuses on properties of these densities, notably those of the density of $\underline{X}$. Clearly, it is not the items that have a joint distribution, but rather the X_j, these which represent the full set of scored responses to each item stem in a focal population, **P**. Nevertheless, it will sometimes be convenient to employ looser terminology, and describe the items as having a distribution and empirical properties. Finally, there will be the need to discuss two distinct senses of *structure*. When talk is of "the structure of the items", what will be meant is the association structure of the X_j, an empirical

fact about the joint distribution of the X_j in focal population **P**. On the other hand, talk of the "theoretical structure of a test" refers to non-empirical, theoretical characteristics of the relationship between the items of a test and the attribute for which they were designed to be indicators.

2. A HISTORY OF TEST THEORY: CLASSICAL AND MODERN TEST THEORY

Spearman and the Birth of the Classical Approach

The origins of classical test theory are generally traced back to the early work of Charles Spearman, in particular to his 1904 article, "The Proof and Measurement of Association Between Two Things". In this work Spearman emphasized the notion that where an individual sits with regard to a particular (mental) attribute is something that cannot be infallibly measured, and, hence, what is observed with respect to an individual's standing on the attribute can be decomposed into two non-overlapping parts: that pertaining to how much of the attribute the individual truly possesses and that reflecting the imprecision associated with the particular measurement instrument employed. Specifically, Spearman distinguished between "systematic" and "accidental" deviations, the former of which he believed are due to the interrelationships of measured variables to some general ability or tendency, which he called *g*, and the latter of

which he thought represented "accidental deviation from the real general tendency" (p. 88), or variation due to error of measurement.[3]

He claimed that in practice only *approximations* to the "true objective values" of the measured variables can be obtained, and, as a consequence of such error of measurement, the "real" correlation between the composite score of a set of measures of p and another composite score of a set of measures of q, r_{pq}, will be attenuated. Spearman then provided the now familiar *correlation attenuation formula* with which one could eliminate the effect of error "disturbances", and thereby ascertain the true correlation between p and q, via two or more independent series of observations of both p and q. Specifically, the correlation between the "true objective values" of p and q is given by

$$r_{pq} = \frac{r_{p'q'}}{\sqrt{r_{p'p'} \cdot r_{q'q'}}}, \tag{2.1}$$

in which p' and q' are observed composite measures for p and q respectively, $r_{p'q'}$ is the average correlation between the individual measures of p with the individual measures of q, $r_{p'p'}$ is the average correlation between one and another of several independently obtained measures of p, and $r_{q'q'}$ is the same for q (Spearman, 1904). Spearman did acknowledge, however, that a practical

[3] Importantly, as McDonald (1999) notes, Spearman's interest in error of measurement was not isolated from his work on the existence of a "common factor", and his treatment of both presumed that cognitive performances depend on a single underlying factor *g*, or general intelligence.

difficulty associated with using (2.1) is that of obtaining two or more observed measures of p and q that are "sufficiently" independent of one another. He noted, however, that in the face of such dependency among trials, the formula is still valid, but simply does "not go quite far enough" (1904, p. 91) and will underestimate the true correlation.

In 1907, in response to criticism (most pointedly from Karl Pearson) that no proof was provided for the formula presented in the 1904 paper, Spearman published a proof of the attenuation formula. There he specified a number of conditions which the proof required, namely, that 1) the average of all measures of p (and of q) is equal to (or varies proportionally to) the true score of p (and of q); and 2) over different measures of p (and of q), the errors of measurement are independent.

Early Conceptions of Reliability and the Spearman-Brown Prophecy

In 1910, Spearman published another important paper in the *British Journal of Psychology* in which he responded to further criticism that his attenuation formula, although possibly providing a valid correction in cases in which the only "disturbances" to precise measurement are "accidental", was likely invalid in circumstances in which

> The discrepancies between successive measurements...cannot properly be termed "accidental", but may arise from the fact that the second later measurement does not deal with the same function as the earlier one, owing to the modifications introduced by practice, fatigue, etc. (Spearman, 1910, p. 272)

To this, Spearman reemphasized the need to distinguish between "systematic" effects and "accidental" errors, arguing that the former need to be controlled for and estimated, whereas the latter need be corrected for. First, Spearman described the logic underlying the creation of two such halves as follows:

> Let each individual be measured several times with regard to any characteristic to be compared with another. And let his measurements be divided into several—usually two—groups. Then take the average of each group...*The division into groups is to be made in such a way, that any differences between the different group averages (for the same individual) may be regarded as quite "accidental".* (p. 274; emphasis in original)

In other words, a set of measures of an attribute x may be partitioned into p mutually disjoint subsets, and an unweighted sum produced for each subset. Analogously, a set of measures of an attribute y may be partitioned into q mutually disjoint subsets, and unweighted sums produced for each of these. He then suggested a new correction formula,

$$r_{xy} = r_{x(p),y(q)} \sqrt{\frac{1+(p-1)r_{x(1),x(1)}}{pr_{x(1),x(1)}} \cdot \frac{1+(q-1)r_{y(1),y(1)}}{qr_{y(1),y(1)}}} \tag{2.2}$$

in which p denotes the number of composite measures of x, q denotes the number of composite measures of another variate y, $r_{x(1),x(1)}$ denotes the average

correlation between the composite measures of x, $r_{y(1),y(1)}$ denotes the same for y, and $r_{x(p),y(q)}$ denotes the average of the correlations of the p composite measures of x with the q composite measures of y (Spearman, 1910). Spearman claimed that, assuming that the errors of measurement for a composite of any given subset are uncorrelated both with the errors of measurement and true scores (which he called the underlying "regular" measurement) of the composite of any other subset,[4] then the corrected correlation that estimates the correlation between the average of the true scores on the p x-composites and q y-composites is given by (2.2).

In the same paper Spearman introduced the term "reliability coefficient" to describe "the coefficient between one half and the other half of several measurements of the same thing" (p. 281). He then gave as an example of how such halves might be obtained, viz., that one could simply divide a test consisting of k measures of some attribute of interest into two halves by letting the even-numbered items form one group measures, while the odd-numbered items form the other group of measures. Then each group could be formed into a composite score by taking the average score of the measures contained in the group. Spearman then provided a formula for estimating the coefficient of reliability (based on the correlation between the composite measures derived

[4] These two assumptions, it was suggested to Spearman in a private letter by Yule, must be explicit (Walker, 1929).

from the above-described method, or some such similar means of creating two roughly equivalent measures of the same thing) as a function of any given additional number of measures:

$$r_{x(p),x(p)} = \frac{pr_{x(q),x(q)}}{q+(p-q)r_{x(q),x(q)}}. \tag{2.3}$$

In this formula, "$r_{x(q),x(q)}$ is the known reliability coefficient of x when the latter has been measured ($2q \times i$) times, i being any number, and $r_{x(p),x(p)}$ is the required most probable reliability coefficient if x be measured ($2p \times i$) times" (Spearman, 1910, p. 281). Spearman illustrated the case in which $q = 1$, (2.3) reducing to

$$r_{x(p),x(p)} = \frac{pr_{x(1),x(1)}}{1+(p-1)r_{x(1),x(1)}}, \tag{2.4}$$

which expresses the degree to which the reliability of a test will increase (or decrease) if it were lengthened (or shortened) by adding (or subtracting) p similar groups of measures to the original set.

In an article adjacent to Spearman's 1910 paper, William Brown reported a number of empirical results pertaining to correlations between tests of "very simple mental abilities" for relatively homogeneous groups of individuals. In this work, he defined a coefficient, r_2, as a measure of the extent to which "the amalgamated results of...two tests would correlate with a similar amalgamated series of two other applications of the same test" (p. 299) and, in a simple and elegant proof, showed that this correlation was equal to

(2.5) $$r_2 = \frac{2r_1}{1+r_1},$$

in which r_1 he defined simply as the "Reliability coefficient (r_1) for each test" (p. 299), and which is equivalent to Spearman's formula (i.e., as shown in (2.4)) for the case in which $p = 2$. Like Spearman, Brown also provided the general result for p tests,

(2.6) $$r_p = \frac{pr_1}{1+(p-1)r_1},$$

which he claimed "furnishes a ready means of determining from the reliability coefficient of a single test, the number of applications of the test which would be necessary to give an amalgamated result of any desired reliability" (p. 299). Spearman and Brown's independently derived results regarding the effect of test length on reliability would come to be known as the *Spearman-Brown prophecy*, which remains in popular use today.

The Spearman-Brown formula requires that there be some means of obtaining an estimate of the reliability coefficient (i.e., Spearman's $r_{x(1),x(1)}$ and Brown's r_1), which is presumed to be the same for any pair from a set of measures of some attribute in question. Interestingly, Spearman and Brown appeared, at least initially, to hold differing conceptions of how such an estimate might be obtained. Whereas Spearman's conception may be readily deduced from his definition of the "reliability coefficient", which, recall, is the correlation

between the total unweighted sums of each of two sets of measures of the same thing. The sets could be produced through any of a number of different methods, e.g., by splitting an existing test into two parts or producing two similar forms of a given test. Brown, on the other hand, did not give an explicit definition of r_1. However, he did refer to the number of *applications* of the test that would be required to yield a desired value of reliability, implying that estimates of r_1 could be obtained by correlating two *administrations*, separated by some time interval, of the same test. Indeed, in his empirical examples, he noted that most of the tests were applied twice,[5] with the second application occurring "about a fortnight after the first, and at the same hour of the day" (p. 298). The implication is that reliability estimates for each of the tests were obtained by the "test-retest" method.[6] Hence, despite their independent contributions toward a correction formula for test length, Spearman and Brown held differing conceptions of how, in practice, estimates for the reliability of a single test might be obtained.

At any rate, these 1910 articles of Spearman and Brown underscore three important results for classical test theory: 1) An estimate of the reliability of a test could be obtained by correlating scores from two suitably similar tests of the

5 The one exception, however, is the Müller-Lyer Illusion test, for which an estimate of reliability was obtained by dividing the results into two halves and then correlating them.

6 Although, as Walker (1929) notes, Brown eventually adopted Spearman's definition of r_1 as the correlation between two comparable forms of a test.

same attribute; 2) the reliability of a test will increase with increased length (i.e., "amalgamated pairs of tests") and decrease with diminution in length; and 3) from (2), it is possible to determine from the reliability coefficient of a single test how much the test must be lengthened (or shortened) to obtain a desired degree of reliability.[7]

"True Score" Defined and the Index of Reliability

In 1911, Abelson showed that Spearman's general formula,[8]

$$r_{x(p),y(q)} = r_{xy}\sqrt{\frac{pr_{x(p),x(p)}}{1+(p-1)r_{x(p),x(p)}}}\sqrt{\frac{qr_{y(q),y(q)}}{1+(q-1)r_{y(q),y(q)}}}, \tag{2.7}$$

with $p = 1$, $r_{xy} = 1$, and $q \to \infty$, equals the square root of the reliability coefficient, $r_{x(1),x(1)}$,

$$\begin{aligned} r_{x(1),y(\infty)} &= 1\sqrt{\frac{1\times r_{x(1),x(1)}}{1+(1-1)r_{x(1),x(1)}}}\sqrt{1} \\ &= \sqrt{r_{x(1),x(1)}} \end{aligned} \tag{2.8}$$

[7] However, it should not go without noting that Brown was critical of Spearman's (1904, 1907) work on the attenuation formula, claiming that Spearman's assumption of independence of errors with underlying objective values and with other errors "are very large assumptions to make". Brown believed that the "accidental deviations" to which Spearman referred were, in fact, not "accidental" at all, and, instead, represent variability of performance of function *within* the individual, and, hence, "to assume them uncorrelated with one another or with the mean values of the functions is to indulge in somewhat *a priori* reasoning" (p. 319).

[8] As presented in (2.2), but with r_{xy} and $r_{x(p),y(q)}$ reversed.

(Abelson, 1911). Abelson used this formula in a numerical example to express the probable correlation between a single test score, x, and the average of the infinity of similar such measures of the same function, y, the latter of which he called the "true value" (cited in Walker, 1929). Five years later, Kelley (1916) made an independent derivation of the same formula, which came to be known (apparently quite by accident) as the *index of reliability* (cf. Walker, 1929). Later, Kelley claimed that the highest possible correlation that can be obtained between a test and a second measure "is with that which truly represents what the test actually measures—that is, the correlation between the test and the true scores of individuals in just such tests", in which "true scores" he defined as "the average scores of individuals upon a very large number (and infinite number) of just such tests" (1921, p. 372). Kelley speculated that (2.8) might constitute a "more significant index of reliability" than the usual reliability estimate that gives the correlation between two similar measures of the same thing, which themselves may or may not be good estimates of their respective true scores.

A Prelude to Parallelism: Early Conceptions of "Parallel Measures" and Methods for Obtaining Such Measures

Although Spearman did not explicitly define the reliability coefficient until 1910, he did, by explicating the terms in the correlation attenuation formula, make reference to the reliability of p' (and also q'), the approximation to p (and

to q), which there he defined as "the average correlation between one and another of...several independently obtained series of values for p [or q]" (p. 90). Hence, Spearman's initial conception as regards obtaining estimates of reliability from a set of measures of some "objective true value" required only that the measures be independent,[9] with reliability of any one group of measures being defined simply as the average correlation between all pairs of the individual group averages. In 1910, when Spearman presented results pertaining to the effect of test length on reliability, he observed that all the measures of an attribute, x (or of y), should, if possible, be of "general equal accuracy". He further noted in the proof of the formula presented in (2.3) that

> Although this formula applies immediately to groups of approximately equal liability to accidental disturbances, it can easily be extended to cases of unequal liability. For an actual measurement of any degree of accuracy is...equivalent to the average of a number of measurements of an inferior degree of accuracy. So that two actual measurements (or groups of such) of unequal accuracy may be conceived as the averages of two unequal numbers of measurements all of equal (inferior) accuracy. (1910, p. 291)

Hence, for Spearman, it did not matter if the groups of measures employed to derive the full reliability of the set had equal reliabilities, as long as they could be considered reasonably independent.

Brown, on the other hand, made no claims regarding the independence of measures. In fact, he was quite critical of Spearman's views on this matter,

[9] Noting further, however, that the condition was difficult to obtain in practice.

claiming that there are reasonable grounds for assuming that errors of measurement *will* be correlated. In addition, Brown defined "reliability coefficient" variously as "the correlation coefficient of the marks [of a test] obtained on two different occasions", and "the correlation of two halves of a split test" (Brown and Thompson, 1940, p. 132). He did, however, specify in his 1940 text *The Essentials of Mental Measurement*, co-written with G.H. Thompson, that his formula for correcting the reliability for test length is a special case of a formula expressing the correlation of a sum of groups of measures when reliability coefficients for the groups equal r_1, and all groups' standard deviations are equal.

Truman Kelley championed the idea that, in order for one to have faith that the correlation between two measures (or groups of measures) of the same attribute will give an estimate of the reliability of one (or the other), it must be the case that the two measures (or groups) are "comparable tests". In 1923, Kelley laid out a specific set of conditions that must hold in order for two tests to be considered comparable, and, consequently, for the correlation between them to be considered a "reasonable" reliability coefficient:

> The following rule for the construction of two comparable tests may be laid down: (1) sufficient fore-exercise should be provided to establish an attitude or set, thus lessening the likelihood of the second test being different from the first, due to a new level of familiarity with the mechanical features, etc.; (2) the elements of the first test should be as similar in difficulty and type to those in the second, pair by pair, as possible; but (3) should not be so identical

> in word or form as to commonly lead to a memory transfer or correlation between errors. (p. 203)

Hence, Kelley differentiated between reliability as defined as the correlation between two "comparable" measures, which may be considered to be two similar forms of a test that measure the same thing, and the reliability of a test given by the correlation between the results from two applications of the same test, the latter of which he did not consider to give a proper estimate of reliability, because the condition of independence of errors could not, in his view, be reasonably assumed.[10] For Kelley, the proper method for obtaining an estimate of the reliability of a test was to correlate two comparable *forms* of a test, with the reliability of the full test then given by application of the Spearman-Brown formula.

Classical Reliability and "Equivalence" Defined

In a 1924 article entitled "Note on the reliability of a test: A reply to Dr. Crum's criticism", Kelley discussed the implications involved in using the Spearman-Brown formula in cases in which the individual measures of a set have unequal variability. In this article he presented a number of scenarios with regard to the properties of "similar" tests, one of which considered two tests scores,

[10] However, he did note that this method may be considered a sound procedure for obtaining a *lower bound* to the true reliability coefficient (cf. Kelley, 1923).

(2.9) $$x_{1_i} = a + e_{1_i} \text{ and } x_{2_i} = a + e_{2_i},$$

in which x_{1_i} and x_{2_i} are deviations of the observed test scores from their respective means for the i^{th} individual, a is equal to the "ability factor" as a deviation from the mean (the lack of subscripts indicating that this value is presumed to be equal for tests 1 and 2, a presumption based on the notion that the tests are "similar", i.e., may be considered to be two forms of the same test), and e_{1_i} and e_{2_i} are deviations of the chance factors from the means for tests 1 and 2 respectively. Kelley offered up the following further specifications: 1) a and e_{1_i} (and, so too, a and e_{2_i}) are presumed to be entirely uncorrelated, with the consequence being that

(2.10) $$\sigma_1^2 = \frac{\sum_{i=1}^{N} x_{1_i}^2}{N} = \frac{\sum_{i=1}^{N} \left(a + e_{1_i}\right)^2}{N} = \frac{\sum_{i=1}^{N} a^2}{N} + \frac{\sum_{i=1}^{N} e_{1_i}^2}{N} = \sigma_a^2 + \sigma_{e_1}^2,$$

(p. 196) and similarly for σ_2^2; and 2) for a given individual the chance factors will vary over replications of measurement, so that $e_{1_i} \neq e_{2_i}$, and, hence, $x_{1_i} \neq x_{2_i}$. However, since the tests are presumed to be "similar", the standard deviations of the chance factors will in the long run be the same, such that $\sigma_{e_1}^2 = \sigma_{e_2}^2$,[11] and

[11] This is the first formal specification of the conditions that must be satisfied for two (or more) measures to be considered "similar" or "equivalent"; both expressions would eventually be replaced by the term "parallel", whose origin not clear. The term was used by Thurstone in 1931, however, to describe two forms of a test in which the two forms are paired, item by item, "in order 1) to make sure that the two paired items should be sufficiently similar in the abilities tested to warrant their classification as parallel, and 2) to make sure that they are not so nearly similar that they are for practical purposes identical" (p. 9).

(2.11) $$\sigma_2^2 = \sigma_a^2 + \sigma_{e_2}^2 = \sigma_a^2 + \sigma_{e_1}^2 = \sigma_1^2$$

(p. 196). Kelley then defined the correlation between the scores on the two similar tests as

(2.12) $$r_{12} = \frac{\sum_{i=1}^{N} x_{1_i} x_{2_i}}{N\sigma_1\sigma_2} = \frac{\sum_{i=1}^{N} \left(a + e_{1_i}\right)\left(a + e_{2_i}\right)}{N\sigma^2} = \frac{\sum_{i=1}^{N} a^2}{N\sigma^2} = \frac{\sigma_a^2}{\sigma^2},$$

(in which $\sigma^2 = \sigma_1\sigma_2$ since $\sigma_1 = \sigma_2$) and claimed that "We thus see that in this simple case the reliability coefficient is that proportion of the total variability [of a single test]...which is due to the common ability factor *a*" (p. 196). Kelley concluded that,

> *(a)* If the two halves are in truth measures of the same function and equally reliable and equally variable, the reliability of the sum or average of the two is exactly given by the Spearman-Brown Formula. *(b)* If they are in truth measures of the same function but unequally reliable and unequally variable but not radically different in these respects, the reliability of the sum will be given by the Spearman-Brown Formula to a remarkably close approximation. (p. 201)

Hence, according to Kelley's viewpoint, the "split-half" method of obtaining reliability estimates, i.e., correlating the scores from two halves of the same test, is applicable for tests in which the halves are equivalent, or very closely approximately so. The method gives an estimate of the reliability of one or the other half, the latter of which could then be "stepped up" by the Spearman-Brown formula to give the reliability of the entire test.

Estimating Reliability From a Single Trial

The Kuder-Richardson Formulas

In 1937, Kuder and Richardson provided results pertaining to the "theoretically best estimate" of the reliability coefficient stated in terms of a definition of equivalence of two forms of a test. They rendered the following definition of "equivalence": For a test consisting of items *a, b, ..., n* and a second hypothetical test consisting of corresponding items *A, B, ..., N*, the two tests are (operationally) defined as "equivalent" if 1) items *a* and *A*, *b* and *B*, etc. may be considered interchangeable, 2) the members of each pair are equal in difficulty (i.e., have the same mean) and are correlated to the extent of their respective reliabilities, and 3) the inter-item correlations for each test are equal (cf. Kuder and Richardson, 1937).

Kuder and Richardson were critical of the split-half "method" of obtaining reliability estimates. In particular, they highlighted the "pertinent observation" that such split-half coefficients do not produce unique values, as there are $\frac{k!}{2(\frac{k}{2}!)^2}$ ways of dividing a given test with *k* items into two halves, each of which could be used to produce a potentially different estimate of the reliability of the test. They also noted that this problem could not be ameliorated by simply obtaining an estimate of the reliability of a test by correlating two equivalent forms of the same test (as was Spearman's original conception) as, for two such equivalent

forms, a shift of items from one to another would produce $\frac{(2k!)}{2(k!)^2}$ such pairings of presumably equivalent forms, each pair of which, once again, could be correlated to give a (possibly different) estimate of reliability.

Kuder and Richardson offered a solution to the problem of obtaining two equivalent sets of measures of the same thing, from which an estimate of reliability could be obtained, by showing that an estimate of the reliability of a test composed of k (dichotomous) items is given by

$$r_{tt} = \frac{\sigma_t^2 - \sum_{j=1}^{k} pq + \sum_{j=1}^{k} r_{jj} pq}{\sigma_t^2}, \tag{2.13}$$

in which σ_t^2 is the observed variance of the test scores, $\sum_{j=1}^{k} pq$ is the sum of the item variances, and $\sum_{j=1}^{k} r_{jj} pq$ is the sum of the product of the item reliabilities and their variances (cf. Kuder and Richardson, 1937). However, equation (2.13), they noted, is not calculable due to the fact that the item reliabilities, i.e., the r_{jj}'s, are "not operationally determinable except by use of certain assumptions" (p. 154). They then presented a number of modifications of the basic formula, modifications which would enable one to estimate the correlation between two equivalent forms of a test from item statistics computed on a single form. Two of these coefficients became well-known as the *Kuder-Richardson Formulas 20* and *21*, respectively

(2.14) $$KR_{20} = \left(\frac{k}{k-1}\right)\left(\frac{\sigma_t^2 - k\overline{pq}}{\sigma_t^2}\right),$$

in which $\overline{pq}$ is the average of the observed item variances and σ_t^2 is defined as above, and

(2.15) $$KR_{21} = \left(\frac{k}{k-1}\right)\left(\frac{\sigma_t^2 - k\,\overline{p}\,\overline{q}}{\sigma_t^2}\right),$$

in which $\overline{p}$ is the average of the item averages (or "difficulties"), and $\overline{q} = 1 - \overline{p}$ (cf. Kuder and Richarson, 1937). KR_{20} will equal KR_{21}, they noted, for tests in which all items have the same means; otherwise, KR_{20} will be greater than KR_{21}.

Guttman's Lower Bounds to Reliability and the Birth of Coefficient Alpha

In 1945, Guttman distinguished three sources of variation in observed test scores, viz., variation due to trials, persons, and items. He conceptualized error as being defined for each person on each item over a universe of trials. His conception emphasized the notion of the "propensity distribution", which is the distribution of test scores, over an infinity of experimentally independent trials, conditional on a particular individual's true score. Guttman then defined the reliability of a test, T, for a given population of individuals as

(2.16) $$\rho_T^2 = 1 - \frac{E_i\,\sigma_{\varepsilon_i}^2}{\sigma_T^2} = \frac{\sigma_\tau^2}{\sigma_T^2},$$

in which $E_i \sigma^2_{\varepsilon_i}$ the expected value (over a population of individuals) of the (conditional) variances of the propensity distributions, σ^2_t is the unconditional variance (over a population of individuals) of the individuals' expected test scores over a population of trials, and σ^2_T is the unconditional variance of the test over trials and over the population of individuals. Guttman proved that (2.16) was equal to the correlation between two experimentally independent trials, claiming that "if it is possible to make two independent trials of a test in practice, on a *large* population,...the correlation between the two trials may be taken as equal to the reliability coefficient" (p. 268). He further noted that this definition, although still lending itself to the notion that the total variance of the test is equal to the sum of the true score and error variances, does so without the assumption, required by the conventional approach to reliability estimation, of uncorrelated true and error scores. However, Guttman also recognized that the attempt to obtain responses that were independent would inevitably be plagued with ineluctable practical difficulties, and that, practically speaking, only estimates of *lower bounds* to reliability could be obtained. He derived six such bounds, each of which required only the assumption that the errors are independent between items and between persons over the universe of trials.

One of Guttman's lower bounds, λ_3, a generalization of Kuder and Richardson's KR_{20} for items not necessarily restricted to dichotomous scoring, he defined as

(2.17) $$\lambda_3 = \frac{k}{k-1}\left(1 - \frac{\sum_{j=1}^{k}\sigma_j^2}{\sigma_T^2}\right),$$

in which σ_j^2 is the variance of responses over persons to the j^{th} item, for $j = 1, 2,\ldots, k$, and σ_T^2 is defined as in (2.16). In 1951, Cronbach defined a quantity, which he denoted α, that is equivalent to Guttman's λ_3. He proved that α is equal to the mean of the split-half coefficients that would be obtained from the $\frac{k!}{2(\frac{k}{2}!)^2}$ splittings of the test into two halves. This coefficient was subsequently developed and popularized by Cronbach and, hence, came to be known as "Cronbach's alpha".

Axiomatizations of Classical Test Theory

Early Attempts

In his text *Statistical Method* (1923), Kelley included in a section on the reliability coefficient a number of the basic results of what would come to be known as *classical test theory.* It is here that Kelley offered up the aforementioned definition of reliability as the correlation between "comparable" tests. In 1931, Thurstone published *The Reliability and Validity of Tests,* in which he expanded on Kelley's treatment, including, among other things, additional sections on

different methods for determining the reliability of a test, the effect of test length on validity and on the relations between reliability and validity, and particular methods for scoring tests. This piece, which he developed out of his lecture notes on psychological measurement theory, is the first of its kind to include a relatively comprehensive summary of the then 30 year history of test theory.

Gulliksen

In 1950, Gulliksen provided a formal summary of the first half-century of test theory in his *Theory of Mental Tests*. In this classic work, Gulliksen defined the quantities, and relations between them, represented in the classical true score model, which describes the observed test score for individual *i* as the sum of the individual's "true score" and an "error component", i.e., the difference between the individual's observed score and true score. He defined "random errors" as errors of measurement that will average to zero over a large number of cases, which, by this definition, are distinguished from "systematic errors" that might come about, for example, from some bias associated with the measurement instrument. To call Gulliksen's work a thorough summary of the theory of mental testing would be a gross understatement. This monumental work provides a comprehensive account of the first 50 years of technical developments pertaining to psychological testing, and includes derivations of the basic formulas of the (classical) true score model. On the latter issue, Gulliksen described two equivalent approaches to the problem of determining the

characteristics of "true" and "error" random variates, on which no realizations could be taken, on the basis of a single observed test score. He also provided, among other things, a formal definition of parallel tests in terms of true score and error variance, a discussion of the various interpretations which may be given to the error of measurement, descriptions of the effects of test length on reliability, validity, and other test parameters such as observed mean and variance, as well as sections pertaining to topics ranging from how reliability estimates may be obtained to methods for standardizing and equating test scores. Gulliksen's work was the first to constitute an exhaustive treatment of the issues to be considered by both test constructors and users alike when pronouncing on the quality of particular testing materials.

Lord & Novick

In 1959, Lord published an article in *Psychometrika* in which he presented three possible approaches to making inferences about true scores, the last of which he claimed lies at the heart of mental test theory. He presented five different true score models. The simplest of these, which he called the *matched-forms model*, involves only two assumptions, viz., that 1) that the expected value of the error of measurement is always zero, and 2) the true score of each individual testee is presumed to be equal for each of the k administered tests.

Seven years later Novick (1966) presented what he took to be the core axioms and principal results of *classical test theory*, which he there defined as "that

theory which postulates the existence of a true score, that error scores are uncorrelated with each other and with true scores and that observed, true and error scores are linearly related" (pp. 1-2). He added that classical test theory "is the simplest case of *weak true score theory*, by which we mean that collection of models that make no specific assumptions concerning the functional form of observed score, true score, or error score distributions" (p. 2). Many of the results presented by Novick echo Gulliksen's treatment; in fact, Novick admitted that the motivation behind the piece was not primarily to derive new results, but, rather, "to explicate the conditions under which old results are valid" (p. 1). However, whereas Gulliksen worked under the assumption that the mean error is equal to zero in focal populations of respondents, Novick's treatment, following Guttman, was founded on the claim that the expected value of the error random variable *conditional on the i^{th} examinee* was equal to zero, from which he derived consequences for unconditional distributions (i.e., in populations of respondents):

> Now suppose that a measurement g is taken on a randomly selected experimental unit generating the observed score random variable $X_{g\bullet}$ taking values $x_{g\bullet}$. Let $T_{g\bullet}$ be the random variable (the true score random variable corresponding to the true score values $\tau_{g\bullet}$ that might be generated (though not observable) and let $E_{g\bullet}$ be a random variable (the error random variable) corresponding to the values $e_{g\bullet}$ thus obtainable. Then clearly
>
> $$e_{g\bullet} = x_{g\bullet} - \tau_{g\bullet}$$

> ...The axioms of classical test theory may then be *derived* from the following theorem.[12]
>
> THEOREM 2.1
>
> (a) $Ee_{g\bullet} = 0$
> (b) $\rho(e_{g\bullet}, \tau_{g\bullet}) = 0$.
> *If* X_{ga} *and* X_{ha} *are independent then* E_{ga} *and* E_{ha} *are independent and* [in which a denotes individual examinees]
> (c) $\rho(e_{g\bullet}, e_{h\bullet}) = 0$. (pp. 2-3)

In the proofs Novick provided, he was able to show that the true score and error random variables are uncorrelated *by construction* rather than by definition (cf. Novick, 1966, p. 3).[13]

These individual efforts by Lord and Novick laid the groundwork for their joint accomplishment in writing *Statistical Theories of Mental Test Scores* (1968), which would become, and remains today, the single most influential treatise on test theory. In the first three parts of the book, Lord and Novick recapitulated many of the topics and results presented by Gulliksen, but with a number of amendments, most notably: 1) drawing the distinction between tests generally and "composite tests", the latter of which consist of a set of component measures whose individual properties determine the statistical characteristics of composites constructed of them; 2) extending the pre-existing definition of

12 Rather, the theorem *defines* the axioms of the classical true score model.

13 Novick noted that his technique for presenting results pertaining to the classical true score model was novel with respect to the psychometric literature, save Guttman's treatment in 1945 , the latter of which he recognized had been largely ignored by subsequent writers.

parallel measures to include the condition that such measures are not only equal in both their true scores and error variances, but are so in every subpopulation of population **P**; and 3) providing a definition for measures having the same true scores, but possibly different error variances, viz., so-called "τ-equivalent" measures. The fourth part of *Statistical Theories* expanded considerably on the previously received, and decidedly narrow, conception of validity, and emphasized the construct validation approach to assessing validity, which Lord and Novick claimed consists of two components: 1) showing that a test correlates appreciably with other tests with which theory suggests it should correlate, and 2) showing that the test does not correlate appreciably with all other tests with which theory suggests it should not correlate. They further noted that the difficulty in establishing the construct validity of a test "is that the criterion, the construct, is not directly measurable" (p. 278).

The final sections of *Statistical Theories* cover topics pertaining to latent trait models and strong true-score theory. An introduction by the authors to some general notions regarding latent trait theory and latent variable modelling is followed by a section, contributed by Alan Birbaum, in which Birbaum described the "logistic test model" and other latent trait models and their uses in making inferences about an examinee's level of ability with regard to some latent trait. The inclusion of such topics represents a significant departure of Lord and Novick's treatment of test theory from that previously codified in Gulliksen, and

a shift in focus within the test theory literature generally from a presentation of *classical* to *modern* test theory results.

Generalizability Theory: An Extension of the Classical True Score Model

In the early 1970's, Cronbach, along with Gleser, Nanda, and Rajaratnam (1972), developed *generalizability theory* as an extension to the classical treatment of reliability. According to generalizability theory, an individual's observed test score is but a sample of size 1 from a universe of scores, each of which could have been observed as an index of a particular trait. The observed score is conceived as the sum of the individual's "universe score" and one or more sources of error. A counterpart of the classically defined reliability coefficient, the *coefficient of generalizability* is defined as the ratio of universe-score variance to the expected observed score variance; it "expresses, on a 0-to-1 scale, how well the observation is likely to locate individuals [with regard to the attribute in question], relative to other members of the population" (Cronbach et al., 1972). A stated advantage of generalizability theory is that it distinguishes between studies in which measurement procedures can be developed and refined (so-called "G studies") and those which employ such measurement procedures in order to make decisions about individuals' standings on the attribute under study (i.e., "D studies").

The Birth of Modern Test Theory

Setting the Stage: Lawley, Tucker, and Lazarsfeld

In a 1943 paper, Lawley addressed problems associated with item selection and test construction (cf. Lawley, 1943). In this work, he presented a formula which expresses the probability of an individual passing a (dichotomously scored) test item as a function of both the individual's ability on some trait of interest and of two item parameters. One of the parameters quantifies the level of difficulty of the item and the other its power to discriminate between individuals of different abilities on the attribute of interest. Importantly, Lawley noted that the method he proposed assumes that the items of which a given test is composed are measuring the same ability; however, he did not suggest an explicit method for testing whether or not such a condition holds for a given set of item responses. A year later, he published a paper in which he aimed to extend his method for selecting items by employing factor analysis in order to reveal the relations of item responses to underlying abilities (cf. Lawley, 1944). In this work he showed how items with unequal difficulties could introduce a spurious factor due mainly to the differences in the item difficulties, rather than indicating the presence of a second "true" factor.

In a 1946 paper concerning the maximum validity of a test composed of equivalent (i.e., parallel) items, Tucker proposed a mathematical model of the relationship between the probability of "success" on an item (i.e., endorsement of

a dichotomously scored item) to true scores on the underlying ability of interest. He called these item/ability regressions *item characteristic curves*, and claimed that they "can be thought of, in the simplest case, as depending on two parameters, one for the general level of difficulty and one for the discriminative power of the item" (p. 2). The item difficulty he defined as "the score on the scale of ability where the probability of correct responses is one half" and described the item discriminative power "in terms of a coefficient of the amount of spread of the item curve" (p. 2). He proposed that the mathematical function that would produce such curves was the normal ogive,

$$p_{js} = \int_{u_j=-\infty}^{u_j=(s-s_j)/\sigma_j} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u_j^2} du_j , \tag{2.18}$$

in which p_{js} = the probability of a correct response to the j^{th} item by an individual with ability score s, s_j is the difficulty of the j^{th} item, σ_j is the item discrimination power, and u_j is an arbitrary variable, used in order not to confuse this integration with other operations on the s scale (cf. Tucker, 1946, p. 4).

In 1950, Lazarsfeld contributed two chapters on latent structure analysis to Stouffer, Guttman, Suchman, Lazarsfeld, Star, and Clausen's *Measurement and Prediction*. There he described "the latent structure approach to the treatment of itemized tests" (p. 362). He introduced the concepts of "manifest" and "latent" to describe respectively the observed test "response patterns" and the underlying continuum, about which inferences must be made. In addition, Lazarsfeld

explicated the notion of "trace lines" as the function relating the probability of "correct" responding to items to location on the latent continuum. Formally, he defined the trace line for a joint "positive" answer to the full set of (dichotomously scored) items on the test as

$$f_{ijk\ldots}(x) = f_i(x) f_j(x) f_k(x) \cdots, \tag{2.19}$$

in which $f_i(x)$, $f_j(x)$, $f_k(x)$, etc., are the trace lines for the individual items.

Lazarsfeld defined a *pure test* as a test in which any interrelationships among items is completely explained by the existence of one underlying continuum, or, as "an aggregate of items such that the joint positive answers to any number of items have themselves a trace line which is the product of the original trace lines for the items viewed separately" (p. 369). The above given mathematical formulation, according to Lazarsfeld, "leads to a model from which...rather important mathematical inferences can be drawn", inferences that "can be judged as to whether they are right or wrong" (p. 366).

The Birth of Modern Test Theory: Lord and Lord & Novick (and Birnbaum)

In 1952, and again in 1953, Lord presented a set of results which are considered by many to signal the inception of what would come to be known variously as "latent trait theory", "item response theory", and "modern test theory". Building on aspects of the work of those individuals just mentioned as well as Carrol (1950), Guilford (1936), Brogden (1946), and others, Lord presented

an alternative theory of mental test scores to the at the time standard "true score" theory. What follows is a summary of the key components of this new test theory, as presented by Lord (1952, 1953):

The Latent Trait, or "Ability"

By "ability" Lord meant the particular "mental trait" for which a given test is a measure. He further claimed that "The ability itself is not a directly observable variable; hence its magnitude, in terms of whatever metric may be chosen, can only be inferred from the examinee's responses to the test items" (1952, p. 1), and, as such "any operational definition of ability...must consist of a statement of a relationship between ability and item responses" (1952, p. 4). So-called "abilities" would eventually become interchangeable with "latent traits", which Lord and Novick (1968) would describe as "the psychological dimensions necessary for the psychological description of individuals" (p. 359). In a more technical sense, abilities, or latent traits, are random variates for which realizations are not possible (i.e., they cannot be measured "directly"), a consequence of which is that their scales may be arbitrarily set. Despite the fact that observed values for these random variates cannot be obtained, their effects are thought to be *manifest* in observed responses to test items. Moreover, Lord (1953) noted that the relation between ability and the classically-defined true score (i.e., the expected value of an infinity of test scores for a given individual) is

in general curvilinear and, further, that there exists a perfect curvilinear correlation between the two.

1. The "single assumption": Homogeneous items

Lord claimed that the "single assumption" of his theory of mental test scores is that the "trait or ability under discussion...be thought of as an ordered variable represented numerically in a *single* dimension" (1953, p. 518; emphasis added), which he described as one of several restrictions that would need to be imposed for his theory to be applicable. Specifically, Lord (1953) claimed that

> Consideration will be restricted to tests that are homogeneous in the following sense: A homogeneous test is for present purposes defined as a test composed of items such that, *within any group of examinees all of whom are at the same ability level,* the responses given to any item are statistically independent of the responses given to the remaining items. (p.521)

Lord (1952) gave a formal specification of homogeneity in the following characterization of the frequency distribution of test scores for examinees at a given level of ability: For a test consisting of n dichotomously scored (i.e., endorsement of "correct" response = 1, lack of endorsement of "correct" response = 0) items, the distribution of test score, s, conditional on fixed ability, c, is defined as

$$f_{s.c} = \Sigma^{*}\Pi_{s}P_{i}\Pi_{n-s}Q_{i} \quad (s = 0,1,...,n), \tag{2.20}$$

in which P_i is the probability of "success" (i.e. endorsement) of item i, Q_i is $(1-P_i)$, $\Pi_s P_i$ is the product of the values of P_i for any s values of i, $\Pi_{n-s}Q_i$ is the

product of the values of Q_i for the remaining $n-s$ values of i, and Σ^* is the sum of $\binom{n}{s} = \frac{n!}{s!(n-s)!}$ such possible products. Lord noted further that (2.20) will have a binomial form, i.e.,

(2.21) $$f_{s.c} = \binom{n}{s} P^s Q^{n-s},$$

when all the items are "equivalent", that is, when all $P_i = P$ (and hence, all $Q_i = Q$).

2. Item characteristic curves

As aforementioned, according to Lord, any "operational definition" of the underlying ability measured by a test must consist in a statement of the *relationship* between that ability and item responses. More specifically, he claimed that the relationship between the two may be stated as follows:

> the probability that an examinee will answer an item correctly is a normal-ogive function of his ability. Denoting this probability for the i-th item by P_i, this relationship may be stated more explicitly:
>
> $$P_i = \int_{-\infty}^{\frac{c-a_i}{b_i}} N(y)dy,$$
>
> where c is the measure of ability, a_i and b_i are values characterizing the item, y is simply a variable of integration, and $N(y)$ is the normal frequency function,
>
> $$N(y) = \frac{1}{\sqrt{2\pi}} e^{\frac{-y^2}{2}}.$$ (1952, pp. 4-5)

In addition, Lord claimed that plotting P_i as a function of c would produce curves corresponding to Tucker's *item characteristic curves* and Lazarsfeld's *trace lines*. He further specified that the regression of *test score* (usually taken to be a simple unweighted sum of the individual test items) on ability would also in general be curvilinear, at least over a large range of ability level, and, in fact, that such *test characteristic curves* will be identical in shape to the average of the item characteristic curves (cf. Lord, 1953).

3. A latent trait test theory model

Lord (1952) described the "fundamental objective" of his theory of mental test scores as finding the "bivariate frequency distribution of test score and ability" (p. 10), and pointed out the general result from statistical theory that this desired distribution is simply the product of the conditional distribution of test score on ability and the marginal distribution of ability, viz.,

$$f_{cs} = f_c f_{s.c} \quad (s = 0, 1, ..., n), \tag{2.22}$$

in which $f_{s.c}$ is as defined in (2.20) and f_c is the (arbitrarily defined) distribution of a population of examinees with respect to the ability (or "latent variable") of interest.

Lord specified a number of restrictions which would need to be imposed in order to apply his proposed test theory They were that: 1) the tests considered be composed of items that are scored either 0 or 1; 2) the "test score" would be defined for each examinee as the simple unweighted sum of the item scores; 3)

the tests considered would consist in items whose ICC's are i) monotonic increasing functions of ability level, ii) bounded below by 0 and above by 1, iii) smooth and having only one inflection point; and 4) the tests considered are homogeneous (in the sense specified in (2.20)). These restrictions would determine in large part the specifics of the particular latent trait model he proposed (cf. in particular Lord, 1952) for modelling responding to test items.

4. Errors of measurement

Lord (cf. 1952, 1953) presented a number of results regarding errors of measurement that departed from the then received classical conception. First, given that the general form of the distribution of test score at fixed ability level specified by Lord is binomial (see (2.21)), errors of measurement need not, and clearly should not in certain cases, be conceived as being normally distributed. Second, Lord (1952) showed that the correlation ratio of test score on ability is equal to the curvilinear correlation of test score on ability and is defined as

$$\eta_{sc} = \sqrt{1 - \frac{E(\sigma^2_{s.c})}{\sigma^2_s}}, \tag{2.23}$$

in which $E(\sigma^2_{s.c})$ denotes the expected value of variance of test scores conditional on ability and σ^2_s denotes the observed test score variance. He showed, however, that since true score and ability have a perfect curvilinear relationship, true scores are equal to the conditional mean of test score given ability, and, hence, the standard deviation of test scores for a given ability equals the standard

deviation of test scores conditional on true score, $\sigma_{s.c} = \sigma_{s.t}$. But the average value of $\sigma^2_{s.t}$ is analogous to the square of what is known in classical test theory as "the standard error of measurement", which is defined as

(2.24) $$S.E._{meas} = \sigma_s \sqrt{1 - r_{ss}} ,$$

in which r_{ss} denotes the test reliability. In substituting (2.24) into (2.23) and making use of the well-known result that $r_{st} = \sqrt{r_{ss}}$, Lord showed that

(2.25) $$\eta_{sc} = \sqrt{r_{ss}} ,$$

i.e., that the curvilinear correlation between test score and ability equals the index of reliability (cf. Lord, 1952).

A third result presented by Lord pertaining to errors of measurement is that for circumstances in which the item (and, hence, test) characteristic curves can be reasonably assumed to be curvilinear in the manner specified in Lord's restrictions, the standard error of measurement will be different at different ability levels, and will in general be smallest at the extremes of the ability continuum (cf. Lord, 1953).

In their aforementioned treatise on test theory, Lord and Novick (1968) dedicated five chapters (four of which were contributed by Allan Birnbaum) to the topic of latent trait models. They provided a summary of almost two decades of work on latent trait theory, emphasizing concepts central to the theory such as *local independence* (i.e., statistical independence) as a formal definition of item

homogeneity, the normal ogive as one particular example of *item characteristic curves*, the binomial form of conditional distributions of test score given level of the latent trait (which they denoted θ), and the relation of latent trait to true score.

Birnbaum's contribution consisted of sections pertaining to topics such as the sufficiency of certain formulas used to create composites of items, classification by ability level, estimation of ability, and a more detailed explication of different latent trait models, with an emphasis on a "logistic test model", which he claimed "very nearly coincides with the normal ogive model" (p. 399). However, he pointed out that the former "has advantages of mathematical convenience in several areas of application" (p. 399). Hence, he specified that the item characteristic curve for the g^{th} dichotomously scored item could be described by a logistic cumulative distribution function, viz.,

$$f_g(u_g \mid \theta) = \frac{\exp[Da_g(\theta - b_g)u_g]}{1 + \exp[Da_g(\theta - b_g)]}, \tag{2.26}$$

in which u_g denotes the item response for the item (either "0" or "1"), θ denotes a fixed value of the latent trait, a_g and b_g are item parameters, and D is a scaling factor, usually set to D=1.7. (2.26) gives the conditional probability distribution function for the g^{th} item for a set of individuals who are invariant with respect to their position on the latent trait.

In addition, Birnbaum included some important results pertaining to the "information structure" of items and tests. Birnbaum defined "information" as a quantity that is inversely proportional to the width of the confidence interval of an estimate of a given examinee's ability (Hambleton and Cook, 1977). He defined the following quantity,

$$I(\theta_1, u_g) = \frac{P'(\theta_1)^2}{P(\theta_1)Q(\theta_1)}, \tag{2.27}$$

"as a measure of *information* per item having ICC of the form $P(\theta)$, that can be used to discriminate abilities in a neighborhood of θ_1" (p. 449; emphasis in original), in which u_g is defined as in (2.26) and $Q(\theta)$ denotes, as usual, $1-P(\theta)$. For a compositing rule that is of the weighted-sum form, $x(v) = \sum_{g=1}^{n} w_g u_g$, Birbaum derived the information function of the compositing rule,[14] $x(v)$, as

$$I(\theta, x) = \frac{\left(\sum_{g=1}^{n} w_g P_g'(\theta)\right)^2}{\sum_{g=1}^{n} w_g^2 P_g(\theta) Q_g(\theta)}, \tag{2.28}$$

[14] Birnbaum referred to this as the "information of the *scoring* rule"; here, "compositing rule" is employed so as to differentiate between the rule that converts the responses of a given respondent to the stimulus materials of a test into real numbers (i.e., what is here termed the "scoring rule") and the rule that produces a scalar function of X_j, $j = 1,2,\ldots, k$, for each respondent (i.e., what is here termed the "compositing rule").

in which the w_g are any positive numbers. He also showed that the maximum value of (2.28) for a particular compositing rule is given by the *information function of a test,* which is defined as

$$I(\theta) = \sum_{g=1}^{n} I(\theta, u_g) \equiv \sum_{g=1}^{n} \frac{P'(\theta_1)^2}{P(\theta_1)Q(\theta_1)}. \tag{2.29}$$

In addition, he noted that (2.29), as the sum of the item information functions, is determined by the particular statistical model being employed, and, further, that it is not dependent on the particular choice with regard to the compositing rule.

A Brief Word on Rasch

In 1960, Rasch developed, autonomously from other latent trait models, and along quite different lines, a probabilistic model which could be (and has been) viewed as a nonparametric latent variable model in which the ICC's are one-parameter logistic functions (Hambleton, Swaminathan, Cook, Eignor, and Gifford, 1978). The "Rasch" model, one of the more commonly employed 1-parameter IRT models, is an appropriate mathematical model for dichotomous test items which are conceptualized as measuring one attribute, as having equal discriminating power, but possibly differing difficulties.

A New Conception of "Validity of the Test": Construct Validation Theory

Peak

In 1953, Peak contributed a chapter in Festinger and Katz' *Research Methods in the Behavioral Sciences* entitled "Problems of objective observation". Peak introduced the notion of "functional unities" to describe certain common characteristics that are shared by a set of processes, behavioural events, or objects. To say that processes, events, or objects have functional unity, Peak argued, means that their shared characteristics go beyond mere "superficial similarities", and, rather, that 1) they change concomitantly, 2) they are dynamically interdependent, or 3) one is causally dependent on the others. She further claimed that most of the methods that are employed in the discovery of the functional unity among observed processes, etc. (e.g., analysis of the "internal consistency" of a set of measured variables) are able to reveal merely the presence of concomitant variation.

In her chapter, Peak included a section on validity, in which she described the role of validity as the interpretation of functional unities. She noted limitations associated with merely considering the "face validity" of a set of "observed processes", or the traditional approach of defining validity in terms of the correlation of a test with some criterion, claiming instead that a broader

conception of validity is required if it is "to have use in a scientific system" (p. 283). Specifically, she asserted that

> to establish the validity of a construct and of the defining measures is to conduct experimental investigations. This involves all the problems of formulating theory, deducing consequences, and testing the deductions under conditions of controlled observation...If behavior theory leads to deductions about conditions of change in process *A* and the effects of *A* on other processes, ways must be found to determine the accuracy of these deductions. When predictions prove to be correct, both the theory and the construct as measured are validated to some degree. (pp. 288-289)

And, furthermore, that

> validation of theory and of instruments of observation tend to proceed simultaneously and...can be separated only in so far as experience has accumulated to suggest that predictions made from a given theoretical structure tend to work out well when the events involved are measured by one set of instruments and badly with another set. (p. 289)

In addition, Peak underscored Steven's distinction between an "indicant" and a "measure", the former of which is a presumed effect that has an unknown (but usually monotonic) relationship with some underlying phenomena, the latter of which is merely a scaled value of the phenomena itself (1951; cited in Peak, 1953). She claimed that the concept of validity should ideally be restricted to examinations of the relationship between the measure and the process measured, but noted that the problem (at least with regard to the measurement of psychological entities) is that "there is no direct access to the underlying phenomena" (p. 291), and, hence, we shall always be left to observe only

indicants of the phenomena. However, "The hope is that we shall approximate more and more closely the law which relates indicant and the thing we want to measure" (p. 291).

Cronbach & Meehl

In 1955, on the heels of the publication of the APA "Technical recommendations for psychological tests and diagnostic techniques" (APA, 1954), Cronbach and Meehl published what would become *the* seminal work concerning the validity of psychological tests, entitled "Construct Validity in Psychological Tests".[15] In the paper they distinguished between three types of validity: *criterion-related (predictive* and *concurrent) validity*—which involves estimating the correlation between a test and a given criterion score, the latter of which may be obtained either subsequently or concurrently with the test score; *content validity*—which is established by demonstrating that the test items are a sample from the behavioural domain under study; and *construct validity*—which is involved whenever the test is to be interpreted as a measure of some attribute which is not "operationally defined" (cf. Cronbach and Meehl, 1967, p. 57). Building on the work of Peak (1953) and others, Cronbach and Meehl laid the groundwork for a completely novel approach to the validation of psychological

[15] Note that the version of Cronbach and Meehl's 1955 *Psychological Bulletin* article which is cited here is a reprint of the original article which appeared in Jackson and Messick (1967).

measures, the ripple effects of which continue to be felt throughout, and beyond, the discipline. The keys features of their paper may be summarized as follows:

1. A New Conception of Validity

Echoing Peak, Cronbach and Meehl distinguished between different types of validity, emphasizing the conditions under which conventional definitions (i.e., either criterion-oriented or content validities) are *inapplicable* for certain tests, viz., tests in which "no criterion or universe of content is accepted as entirely adequate to define the quality to be measured" (p. 58), as is the case, they claimed, for every sort of psychological test at some point or another. Specifically, construct validation comes into play when the tester is interested not in the test behaviour per se, or in being able to predict from test scores certain non-test behaviours. Rather, it features in investigations in which concern is with making inferences about some unobservable quality ("trait", "attribute", "ability") which is thought to underlie test behaviour, i.e., is hypothesized to be responsible, at least to some extent, for variation in test scores. The aim of construct validation then, is to, via an ongoing and progressive program of research with regard to any particular attribute or set of attributes, *determine* "what psychological constructs account for test performance" on a given test (p. 58).

2. The Theoretical Structure of a Test

A construct valid test is a test whose behaviour (i.e., responses to test items) is in keeping with the theory, that is, the "interlocking system of laws" or "nomological network", about the particular attribute which the test is purported to measure. In other words, a test has construct validity if, and only if, the responses to test items are consistent with the "theoretical structure" of the test, that is, some statement of the relationship between the unobservable attribute for which the test is purported a measure and responding to the items of the test. Cronbach and Meehl (1967) paraphrase this notion throughout their seminal article: "Construct validation takes place when an investigator believes that his instrument reflects a particular construct, to which are attached certain meanings. The proposed interpretation generates specific testable hypotheses, which are a means of confirming or disconfirming the claim" (p. 65); "We can say that 'operations' which are qualitatively very different 'overlap' or 'measure the same thing' if their positions in the [theoretical] net tie them to the same construct variable" (p. 66); "To validate a claim that a test measures a construct, a nomological net surrounding the concept must exist" (p. 66), "Hence, the investigator who proposes to establish a test as a measure of a construct must specify his network or theory, sufficiently clearly that others can accept or reject it" (MacFarlane, 1942; cited in Cronbach and Meehl, 1967, p. 67); "A test should not be used to measure a trait until its proponent establishes that predictions

made from such measures are consistent with the best available theory of the trait" (p. 71).

The nomological network surrounding a construct is characterized as relating 1) observable properties or quantities to each other, 2) theoretical constructs to observables, and 3) different theoretical constructs to one another (cf. Cronbach and Meehl, 1967). Since, within the context of test validation, the aim is to assess whether scores for a given test are, in fact, measures of the attribute under study, the tester must be able to deduce behavioural consequences pertaining to the test (including particular relations with other tests and/or criteria of other sorts) from the nomological network, and then assess whether such consequences do, in fact, hold empirically in a set of test data. If the deduced consequences hold, then this may be taken as "evidence" supporting the construct validity of the test; if the empirical outcomes are not in line with predicted consequences, then this may be taken as evidence against the construct validity of the test, but may well (and some would say should) lead to an altering of the nomological network. The key element of this aspect of construct validation theory is that in order to claim that particular observed test behaviours comprise evidence in support of the validity of the test presupposes certain premises with regard to how the test "should" behave, premises which are given by the currently received theory with regard to the construct of interest. For example, as Cronbach and Meehl (1967) note, "Only if the

underlying theory of the trait being measured calls for high item inter-correlations do the correlations support construct validity" (p. 63), and "Whether a high degree of stability is encouraging or discouraging for the proposed interpretation depends upon the theory defining the construct" (p. 64). Importantly, "unless the network makes contact with observations, and exhibits explicit, public steps of inference, construct validation cannot be claimed. An admissible psychological construct must be behavior-relevant" (APA, 1954; cited in Cronbach and Meehl, 1967, p. 66) if there is to be any justification for claims that a test purported to measure the construct is construct valid.

3. *A Program of Construct Validation*

Cronbach and Meehl's explication of construct validity goes well beyond merely determining in a single instance whether a test measures what it is purported to measure. Rather, construct validation is conceived of as a progressive scientific enterprise, in which different sources of evidence with regard to a given construct are continually integrated into the nomological network in which the construct is embedded, such that, over time a scientific community gains a firmer hold on the meaning of the construct. As such, test evaluation is thought to play but one role in a broader endeavour. Deductions regarding how a test of an attribute should perform are given by the relevant theoretical network, as it stands at a fixed point in time. However, the failure to support empirically such propositions may indicate that the received theory

about the attribute is incorrect, and hence, may lead to changes in the relations specified in the network. Hence, construct validation "is not to be identified by particular investigative procedures" (p. 58), but, rather, as an ongoing process wherein the validity of a particular test is assessed in relation to deductions from existing theory, but in which the observed "behaviour" of the test may lead to modifications of the theory. It is believed that, via this process, a science comes progressively closer to a full articulation of *what exactly* is being measured by the test.

Loevinger

In 1957, Loevinger published an article entitled "Objective tests as instruments of psychological theory",[16] in which she claimed that classical validity (i.e., criterion-oriented validity) is not a suitable basic concept for test theory, and does not provide an adequate basis for test construction (cf. Loevinger, 1967). In particular, Loevinger argued that since criterion-oriented and content validities are essentially *ad hoc*, construct validity is the whole of validity from a scientific point of view, and her aim was "to develop a coherent view of psychometrics, a mutually implicative test theory and method of test construction", which would consist in a "radical reformulation of the validity problem" (p. 79).

[16] Note that the version of Loevinger's 1957 *Psychological Reports* article which is cited here is a reprint of the original article which appeared in Jackson and Messick (1967).

Although Loevinger's perspective on the issue of validity was informed in large part by the standards set forth in the "Technical Recommendations" (APA, 1954) and by Cronbach and Meehl (1955), in particular with regard to the inclusion of construct validation as an essential component of the process of validating both tests and psychological constructs, she diverged from these treatments in several notable ways. First, Loevinger eschewed the classification of validity into types, i.e., content, criterion-oriented, and construct validities, and, instead, identified two distinct *contexts* for considering validity: administrative and scientific. The former, she claimed, could be further divided into content and criterion-oriented validities; the latter, construct validity, exhibits, according to Loevinger, "the property of transposability or invariance under changes in administrative setting which is the touchstone of scientific usefulness" (p. 83).

Second, whereas Cronbach and Meehl used the terms "construct" and "trait" interchangeably, Loevinger challenged that "Traits exist in people; constructs...exist in the minds and magazines of psychologists"; "the trait is what we aim to understand, and the corresponding construct represents our current understanding of it" (1967, p. 83). She further contended that what lies at the centre of the validity issue is "exactly what the psychologist does not construct: the validity of the test as a measure of traits which exist prior to and independently of the psychologist's act of measuring" (1967, p. 83).

Third, Loevinger repackaged validity into what she considered to be three mutually exclusive, exhaustive, and mandatory components of construct validation: substantive, structural, and external. The substantive component of construct validity involves determining the extent to which the content of the items of a test can be accounted for in terms of the trait believed to be measured by the test along with the context of measurement. The structural component refers to the extent to which the structural relations among test items are consistent with the structural relations of other (i.e., non-test) manifestations of the trait being measured. She further notes the existence of various structural models which may be used in assessing the structural validity of a test, and that the particular choice of model for test construction (or evaluation) should be given by the existing theory pertaining to the trait in question. Finally, external validity refers to the relation of the test to non-test behaviours, usually in the form of correlations between the test score and certain external criteria as is the case with predictive and concurrent validities.

In sum, Loevinger staunchly advocated a psychometrics driven by a construct validation-oriented approach, wherein

> Only construct validity, which aims at measuring real traits, promises tests which will both draw from and contribute to psychology...The lines of evidence which together establish the construct validity of a test refer to its content, its internal structure, and relation to outside variables. A single explanation or theory must encompass all evidence, for construct validation to be approximated. (1967, p. 119)

The legacy of her advocacy of a construct-validation approach to both conceptualizing and evaluating measures is clearly apparent in the current test analytic practices of psychological scientists.

Other Developments

Multitrait-multimethod Matrices: Campbell & Fiske

In 1959, Campbell and Fiske contributed to the efforts of other proponents of construct validation by introducing the *multitrait-multimethod* approach to examining validity. They contended that the establishment of construct validation requires both *convergent validation* and *discriminant validation,* the former of which refers to the extent of consistency among independent measures of the same trait, the latter referring to the absence of large, positive correlations among independent measures of distinct traits. The multitrait-multimethod approach is based on a *methods* (1,2,...,*p*) by *constructs* (1,2,...,*q*) factorial design of item types in which the ij^{th} item is a measure of the i^{th} construct measured by the j^{th} method. The multitrait-multimethod analysis then rests on an examination of the elements of the resulting $pq \times pq$ inter-item correlation matrix (the MTMM matrix).

Campbell and Fiske (1959) described four features of a multitrait-multimethod *matrix* which support construct validity. They are: 1) when there is evidence of convergent validity, i.e., when the magnitudes of correlations

between different measures of the same trait are large; 2) when the convergent validities are larger than corresponding divergent validities. For example, given two traits, A and B, each measured by two particular methods, 1 and 2, $r_{A_1A_2}$ and $r_{B_1B_2}$ should both be larger than either $r_{A_1B_2}$ or $r_{A_2B_1}$; 3) when a variable is more highly correlated with an independent measure of the same trait than with measures of a different trait with the same method of measurement, e.g., $r_{A_1A_2} > r_{A_1B_1}$; 4) when the same *patterns* of trait interrelationships hold among crossings of traits with either the same or different methods. For example, for three traits, A, B, and C, each measured by three methods, 1, 2, and 3, the pattern characteristic of the set $r_{A_1B_1}$, $r_{A_1C_1}$, and $r_{B_1C_1}$ should be consistent with that of the set $r_{A_2B_2}$, $r_{A_2C_2}$, and $r_{B_2C_2}$, as should the pattern seen in the set $r_{A_1B_2}$, $r_{A_1C_2}$, and $r_{B_1C_2}$ be consistent with that of the set $r_{A_1B_3}$, $r_{A_1C_3}$, and $r_{B_1C_3}$. The authors noted, however, that as a group, these desired conditions are rarely met, as method and apparatus factors often make a substantial contribution to psychological measures, a contribution which, in their view, could not be overlooked by researchers with a genuine interest in establishing construct validation.

Covariance Structure Analysis: Joreskog

In the late 1960's, while working on the rotational problem of factor analysis, Joreskog developed *confirmatory factor analysis* out of a more general

model for testing specific hypotheses about relationships between measured variates and a set of latent random variates for which the measured variates are purportedly measures (cf. Joreskog, 1966; Joreskog, 1969). Joreskog (1969) developed a general procedure by which any number of parameters of a given latent variate model could be held constant and the remaining parameters estimated by maximum likelihood methods. In 1966, he demonstrated how a confirmatory factor analytic model could be used to test a *simple structure*[17] hypothesis. In 1971, Joreskog presented various models applicable to *congeneric tests*, which he defined as tests which measure the same trait, but whose true scores, in contrast to parallel and tau-equivalent tests, are not identical, but are linearly related.[18] He noted an advantage of congeneric tests that they need not be "directly comparable", in the sense that the latent variable which they measure in common need not be measured on the same scale. All the models presented by Joreskog in the paper are special cases of a general model which, he claimed, could be used for handling all estimation and testing problems.

[17] In which simple structure is defined in terms of Thurstone's criteria: 1) each row of the loading matrix must contain at least one zero, 2) each column of the loading matrix must have at least as many zeros as there are factors in the model, 3) for every pair of columns of the loading matrix there should be some rows in which one loading is zero and the other is nonzero, 4) if the number of factors in the model exceeds four, then, for every pair of columns of the loading matrix, a large proportion of rows should have two loading of magnitude zero, and 5) for every pair of columns of the loading matrix there should be only a small number of rows with two nonzero loadings (Thurstone, 1947; cited in Joreskog, 1966).

[18] In fact, Joreskog (1971) noted that both parallel and tau-equivalent tests are special cases of congeneric tests.

In particular, with regard to the latter, Joreskog presented a number of interesting results: First, he described the classical test theory model for congeneric test scores: Let $x_1, x_2, ..., x_m$ be a set of m random variates, the j^{th} representing the set of scores on item j, in some focal population. The classical true-score decomposition of each x_i is then

(2.30) $$x_i = t_i + e_i \quad i = 1, 2, ...m,$$ [19]

for which the usual assumptions of the classical model are presumed to hold (i.e., uncorrelatedness of true and error components for the same test and of error components for different tests, and that the expected value for the error component is zero). If the x_i are congeneric, then

(2.31) $$t_i = \mu_i + \beta_i \tau \quad i = 1, 2, ..., m,$$

with $E(\tau) = 0$ and $Var(\tau) = 1$. Thus,

(2.32) $$x_i = \mu_i + \beta_i \tau + e_i \quad i = 1, 2, ..., m,$$

in which $E(X_i) = \mu_i$ and β_i is the covariance between x_i and τ. The reliability of x_i, $i = 1, 2, ..., m$, is then equal to

(2.33) $$\rho_i = \frac{\beta_i^2}{\beta_i^2 + \theta_i^2},$$

[19] In Joreskog (1971) the equation appears as $x_i + t_i + e_i$. It is assumed that this is a typing error given that the model as presented here is well established.

in which, predictably, β_i^2 is the true score variance and θ_i^2 the variance of e_i.

Second, Joreskog deduced from (2.32) that a set of m congeneric variates could be represented in linear factor analytic terms:

(2.34) $$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\beta}\tau + \mathbf{e},$$

in which $\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\beta}, \text{and } \mathbf{e}$ are column vectors of order m consisting of the x_i, μ_i, β_i, and e_i, respectively. It follows from (2.34) that the population variance-covariance matrix of $\mathbf{x}$, Σ, is equal to

(2.35) $$\Sigma = \boldsymbol{\beta}\boldsymbol{\beta}' + \Theta^2,$$

in which Θ^2 is a diagonal matrix whose elements are the θ_i^2. Thus, as Joreskog claimed, if the x_i have a multivariate normal distribution: 1) the hypothesis of congenerism (and also parallelism and tau-equivalence) can be tested statistically; 2) maximum likelihood estimates of the parameters can be obtained, and, hence, so too an estimate of a lower bound to the reliability of each x_i (via (2.33)).

Unidimensionality

Green, Lissitz, & Mulaik

In 1977, in response to a growing number of instances in which coefficient alpha was being misused as an index of item homogeneity, Green, Lissetz, and

Mulaik attempted to clarify the relation between the concepts of *internal consistency* and *homogeneity*. Although these authors themselves did not give clear definitions of either internal consistency or homogeneity, they cited a number of historically relevant instances of a conflating of the two terms (cf. Green et al., 1977, p. 829-831). Furthermore, it is clear that, in the context of test items, they considered the term "homogeneous" to be synonymous with "unidimensional", but not so for the term "internally consistent". Furthermore, they pointed out an important feature of the relationship between the properties of homogeneity and internal consistency, viz., although homogeneity implies high internal consistency, high internal consistency does not necessarily imply homogeneity. Hence, to use alpha, an estimate of the lower bound to the reliability of an unweighted sum of a set of items, as an index of unidimensionality is simply a misuse of the statistic.

Green et al. (1977) provided a number of numerical counterexamples, in which they constructed artificial data sets, each of which could be represented in terms of a common factor model, but for which they varied parameters such as the number of factors in the model (m), the number of factors influencing individual items (p), the communalities, and the number of repetitions of a basic set of items (a set of items representing the distinct combinations of m factors taken p at a time) in the model. Their observations included the following findings: 1) Coefficient alpha increases as a) the number of items k increases,

regardless of the number of factors in the model, b) the number of parallel repetitions of each type of item increases, c) the number of factors on which each item depends increases; 2) coefficient alpha approaches and exceeds .80 when the number of factors on which each item depends is two or greater and p is moderately large; and 3) coefficient alpha decreases moderately as the items' communalities decrease. The obvious implication of these results is that the magnitude of coefficient alpha can be large when the dimensionality of the items is greater than 1. Hence, it cannot be used as an index of homogeneity (a synonym of unidimensionality according to these authors). Green et al. recommended that coefficient alpha be abandoned as a measure of unidimensionality, concluding that "Perhaps the test constructor would be better advised to look both at the raw correlations among his items and perform a factor analysis of these correlations to see how tenable the notion of homogeneity is for his items" (p. 837). Moreover, they added that developments in confirmatory factor analytic techniques "would allow the test constructor to test hypotheses about the unidimensionality of a set of items" (p. 837).

McDonald

In response to increased interest in latent trait theory and the use of latent trait models for analyzing sets of (binary) test items, and the growing recognition of the need to verify that a set of items are "unidimensional" if certain of these models are to apply, McDonald (1981) provided an explication of the notion of

dimensionality as it applies to both tests and items. The general thrust of his paper is as follows: First, he noted that although the concepts *unidimensionality, homogeneity,* and *internal consistency* are referenced throughout the test theory literature, there is no general agreement as to what these concepts signify, or whether they signify the same property, and, hence, may justifiably be used synonymously. In particular, McDonald argued that the concept of *homogeneity* is used sometimes as a synonym for *unidimensionality* and at other times it is used interchangeably with *internal consistency,* which itself does not possess a clear and universally accepted definition. This, he argued, leads to a logical contradiction in that both *homogeneity* and *internal consistency* are treated as measurable properties, while *unidimensionality* is a property which either does or does not hold and, as such, is necessarily an integer-valued concept. He contended that

> We can never say of two unidimensional tests that one is more unidimensional than the other, as the contradiction would be plain. By substituting *homogeneous* or *internally consistent,* both of which are English language labels that on some definitions or in the absence of clear definition might be taken to denote quantitative concepts—more or less homogeneous, and more or less internally consistent—we might hide the contradiction. Indeed, just such a contradiction seems to have crept into the literature. (1981, p. 103)

Use of these terms in the context of deciding whether or not a set of tests/items is unidimensional, he concluded, will as a rule "add confusion to psychometric discourse, without serving any distinct purpose" (p. 112).

Second, McDonald clarified that the issue of verification (i.e., that a set of *k* items is unidimensional, or, more generally, *m*-dimensional) is separate from, and *presupposes* that definitional matters are settled, as there exists no unitary definition of unidimensionality, but, rather, only specific (and distinct) *senses* of the concept – the point being that one cannot *verify* that a set of items is unidimensional without *first* having the particular sense in which the term is being used firmly in place. McDonald noted three different conceptualizations of the notion of unidimensionality in the context of tests and items, viz.: 1) unidimensional sets of quantitative (i.e., continuous) test scores, 2) unidimensional sets of binary test items, and 3) *homogeneous/internally consistent* items.

In regard to (1), McDonald described how the linear common factor model with one common factor gives a clear (mathematical) definition of a unidimensional set of quantitative tests, specifically, in terms of the principle of local independence, or the statistical independence of a set of test scores at a fixed level of a single common factor. Hence, he claimed, "if the tests fit the single-factor model, we can say that the entire battery is unidimensional" (1981, p. 101). Regarding (2), he described two different conceptualizations of unidimensionality for binary test items, the first of which pertains to whether such items conform to a Guttman perfect scale, with specific examples given in the work of Guttman (1950), Loevinger (1947; cited in McDonald, 1981) and

others. The second conceptualization comes from latent trait theory (or latent structure analysis more generally), in which a set of items is considered to be unidimensional if and only if just one latent trait accounts for the distribution of response patterns of the items. McDonald further described how common factor theory and latent trait theory may be united under *nonlinear factor analysis* (cf. McDonald, 1967), supplying, among other things, a single conceptualization of undimensional sets of items and unidimensional sets of tests, viz., that "a set of n tests or a set of n binary items is unidimensional if and only if it fits a non-linear factor model with one common factor" (p. 104).

Third, McDonald considered the suitability of using coefficient alpha as a quantitative measure of homogeneity. He concluded that, for the same reasons cited above, if homogeneity is taken to be a synonym for unidimensionality, then coefficient alpha, which can take on values ranging from zero to one, cannot be used as a decision criterion for claiming unidimensionality. A reasonable alternative to coefficient alpha as a criterion of unidimensionality, he claimed, might be given by loss functions used in fitting competing models, such as the likelihood ratio function when fitting linear common factor models, with the usual assumptions regarding normality. He claimed that, in general, "Suitable decision criteria could be based on the loss functions used in fitting the models on the basis of which unidimensionality is defined" (1981, p. 113).

The Sequential Component of Test Analysis

Thissen, Steinberg, Pyszczynski, and Greenberg

In 1983, Thissen, Steinberg, Pyszczynski, and Greenberg alleged that, despite the prevalent use of personality and attitude measures in psychology, there has been little if any increase in the sophistication of statistical techniques used in the construction of such measures. They claimed that, for the most part, psychological researchers generally continue to predominantly employ variants of classical test theory to evaluate the quality of measures, but that this use typically "consists of piecemeal computation of the statistics of classical theory with little consideration of their meaning for the scale as a whole" (p. 211). The motivation behind Thissen et al.'s paper was in large part a recognition of the limitations inherent to the classical true score model for addressing issues pertaining to how a test should be optimally scored, or what a "high" reliability coefficient says about the quality of a test, etc. They proposed a "superior alternative" to the classical analysis whereby the logic underlying item response theory, which had been traditionally applied chiefly to "cognitive" (i.e., dichotomous) items, is applied to the Likert style response scales typically seen in personality measures. The basic features of their prescription for how the quality of personality and attitude measures should be assessed are as follows:

1. Modelling Item Responses

In the tradition of IRT, Thissen et al. begin by specifying a statistical model for each item response in which the response is described as a function of the trait being measured. Specifically, random variate x_j, containing the scored responses of the individuals in some focal population to item *j*, is modelled as

(2.36) $$x_j = \mu_j + \lambda_j\theta + \varepsilon_j,$$

with

(2.37) $$\varepsilon_j \sim N(0,\sigma_j^2),$$

and in which θ, an unobservable latent variate, has a mean zero and variance unity. The item parameters are μ_j and σ_j^2, the mean and variance, respectively, of x_j, and λ_j, the slope of the regression of x_j on θ. This is easily recognizable as a classical unidimensional common factor decomposition and its applicability to test analysis comes from identifying the attribute that the items were designed to measure with θ.

Thissen et al. describe two sub-models, the first in which $\lambda_j = \lambda$ and $\sigma_j^2 = \sigma^2$, for all *j*, thus implying the optimal compositing rule

(2.38) $$\hat{\theta}_i = \left(\frac{1}{\lambda n}\right)\left(\sum x_{ij} - \sum \mu_j\right).$$

This optimal compositing rule is the maximum likelihood estimator (MLE) of θ_i, xindividual *i*'s score on the latent variate. This result demonstrates that a

sufficient condition for the sum or average of the items to be proportional to the MLE for θ is that the data may be described with a one-factor common factor model in which the loadings are constrained to be equal (cf. Thissen et al., 1983).

The second sub-model presented by Thissen et al. considers the case in which neither the λ_j nor the σ_j^2 may be reasonably constrained to be equal, and, hence, both are allowed to vary over the items. In this case the MLE (optimal compositing rule) for θ is

$$\hat{\theta}_i = \frac{\left[\sum\left(\frac{\lambda_j}{\sigma_j^2}\right)\left(x_{ij} - \mu_j\right)\right]}{\left[\sum\left(\frac{\lambda_j^2}{\sigma_j^2}\right)\right]}, \tag{2.39}$$

a weighted sum (or average) of the items (cf. Thissen et al., 1983).

2. *Dimensionality*

Thissen et al. observed that three factor analytic outcomes are possible with Likert-type data, which they describe as follows.

Condition 1:

This condition is met when a unidimensional common factor model with equal loadings fits the observed item responses. If this condition holds, then the (unweighted) sum of the item responses for the i^{th} individual is a linear function of the MLE of the trait value for the i^{th} individual, and, hence, an optimal composite is the sum of the individual item responses. A lower bound to the

reliability of an item from a test for which this scenario holds may then be estimated by

$$\frac{\hat{\lambda}^2}{\hat{\lambda}^2 + \hat{\sigma}^2}, \tag{2.40}$$

in which $\hat{\lambda}^2$ and $\hat{\sigma}^2$ are respectively the square of the MLE for λ and the MLE for σ^2. This quantity may then be "stepped-up" by the classical Spearman-Brown formula in order to obtain an estimate of the (lower bound to) reliability for the composite.

Condition 2:

This condition is satisfied if a unidimensional common factor model with unequal regression parameters provides a fit to the set of item responses. Here, the best estimate of the trait value (i.e., $\hat{\theta}_i$) is a weighted sum of the item responses, in which the weights must be determined from the data. Condition 2 entails more complicated formulae for estimating item reliability than does condition 1; however, the general result holds that, if the items fit a one-factor common factor model (with or without equality constraints), then the items may be composited into a test score, the reliability of which (or lower bound to) may be estimated by some means, the latter of which are explicitly tied to the particular model.

Condition 3:

Here, a unidimensional common factor model does not fit the observed set of item responses, which constitutes statistical evidence that more than one source of variation among the individuals contributes to the items responses, or, in other words, that the test measures more than one trait. Thissen et al. claimed that, if this condition holds, "It is nearly impossible to score such a test so that the score represents a single conceptual entity" (p. 215), and, without some test score (i.e., composite), there is no sense in estimating the reliability of the test. Thissen et al. further suggested that a preliminary analysis of unselected item pools will typically result in condition 3. If the goal is construction of a new inventory, they recommend the use of restricted factor analysis as a mechanism for identifying and removing items which seem to be measuring dimensions other than the primary dimension (i.e., trait) of interest and possibly adding new items until either condition 1 or 2 may be shown to hold.

3. *A sequential approach*

Although there is certainly value in demonstrating how one might use the principles of IRT to model continuous item responses, the novel feature of Thissen et al.'s paper is its emphasis on the inherently sequential nature of test analysis. Here, test analyses are conceived of as a set of sequentially related steps, with passage to a given step in the sequence justified by the test's satisfaction of the union of all of the requirements associated with all previous

steps. Thus, to estimate "the reliability of a test" is, in fact, to estimate the reliability of some composite of the test's items, the justification for compositing a test's items is that these items are unidimensional in some particular sense, and so on.

3. A SUMMARY OF CLASSICAL AND MODERN TEST THEORIES

Over the past 100 years an enormous body of work concerning the development and evaluation of tests has accumulated, only a very small portion of which has been presented here. Broadly speaking, test theory results are subsumed under one or the other of *classical* or *modern* test theory, and the two are frequently contrasted, usually in the context of promoting the latter as a better, more sophisticated, and/or more versatile conceptualization of responding to tests than the former. However, the distinctions that are made between CTT and MTT are generally relatively superficial, and are rarely backed up by formal, axiomatic descriptions of these distinct theoretical paradigms. Hence, here, I begin by giving summaries, including mathematical foundations, of what I take to be classical and modern test theories respectively.

A Summary of Classical Test Theory

The mathematical foundations of classical test theory lie in the classical true score model, the central idea behind which is that an individual's observed score on a given measure is the sum of two components, a "true score", τ, and an

error component, ε. Let X be defined as a measure of γ, the attribute of interest. Imagine that for a given individual there exists an infinity of measures of γ, i.e., an infinity of X scores, each of which is defined as the sum of a true score and an error component. The resulting distribution is the distribution of X conditional on a particular individual p,[20] which is also known as the "propensity distribution" of the infinity of measures of γ for p. It takes the following form:

$$X \mid \boldsymbol{\tau} = \tau_p \sim f_{x_p}(\tau_p, \sigma_p^2), \tag{3.1}$$

in which $f_{x_p}(a,b)$ is a density with mean a and variance b, $\tau_p = E(X_p \mid \boldsymbol{\tau} = \tau_p)$, i.e., the mean of the propensity distribution for person p is equal to the "true score" for person p, and $\sigma_p^2 = E[(X - \tau_p)^2 \mid \tau = \tau_p] = E(\varepsilon^2 \mid \tau = \tau_p)$. Note that the latter quantity, the variance of the propensity distribution for person p, is equal to the variance of the error random variate $(\varepsilon \mid \tau = \tau_p) = [(X - \tau_p) \mid \tau = \tau_p]$ for person p. That is, it is the variance of the deviations, for a given individual, of the observed test scores from the individual's true score over an infinity of (hypothetical) replications of the test.

Consider a population, **P**, of individuals, each with a propensity distribution with mean τ_p and variance σ_p^2. The unconditional distribution of X in **P** is

[20] Note that in the remainder of the work, "p", and not the previously employed "i", will be employed to denote a particular individual.

(3.2) $$X \sim f_x(\mu_\tau,\ \sigma_\tau^2 + \sigma_\varepsilon^2)$$

in which $\mu_\tau = \underset{p}{E}\,\tau_p = \underset{p}{E}\,E(X \mid \tau = \tau_p) = E(X)$ and $\sigma_\varepsilon^2 + \sigma_\tau^2 = \underset{p}{E}(\sigma_p^2) + V(\tau_p) =$

$\underset{p}{E}(V(X \mid p) + V(E(X \mid p)) = V(X)$. Although the σ_p^2 can vary over individuals, they are not individually estimable, and, hence, under the classical scheme, estimates of the reliability for a given test are not sought for individual respondents. Instead, the reliability of the "test" is defined as the proportion of observed score variance on the *population* of individuals that is due to true score variation, which is given by the ratio of true score variance to observed score variance. This is equivalent to the squared coefficient of correlation between observed and true scores,

(3.3) $$\rho_{X\tau}^2 \equiv \frac{\sigma_\tau^2}{\sigma_X^2} = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_\varepsilon^2}\,.$$

Because true scores and errors cannot be derived from an observed test score, X, for each member of **P**, (3.3) cannot be estimated from a random sample of test scores. This indeterminacy may, however, be circumvented by defining two measures of the same attribute, X_1 and X_2, as *parallel* if, and only if the following two conditions hold:

(3.4) $\tau_{1p} = \tau_{2p}\ \forall\, p$, and

(3.5) $\sigma_{\varepsilon_1}^2 = \sigma_{\varepsilon_2}^2 = \sigma_\varepsilon^2$.

It then follows from (3.4) and (3.5) that

(3.6) $$\mu_{\tau_1} = \mu_{\tau_2} = \mu_\tau ,$$

(3.7) $$\sigma^2_{X_1} = \sigma^2_{X_2} = \sigma^2_X , \text{ and}$$

(3.8) $$\sigma_{X_1X_2} = \sigma_{\tau_1\tau_2} = \sigma^2_{\tau_1} = \sigma^2_{\tau_2} = \sigma^2_\tau .$$

Hence

(3.9) $$\rho_{X_1X_2} = \frac{\sigma_{X_1X_2}}{\sigma_{X_1}\sigma_{X_2}} = \frac{\sigma_{\tau_1\tau_2}}{\sigma_{X_1}\sigma_{X_2}} = \frac{\sigma^2_{\tau_1}}{\sigma^2_{X_1}} = \frac{\sigma^2_{\tau_2}}{\sigma^2_{X_2}} = \frac{\sigma^2_\tau}{\sigma^2_X} = \rho^2_{X\tau} .$$

That is, the correlation between two parallel tests is equal to the reliability of one or the other test. Hence, the reliability of either test can be estimated by

(3.10) $$\hat{\rho}_{X_1X_2} = \frac{\hat{\sigma}_{X_1X_2}}{\hat{\sigma}_{X_1}\hat{\sigma}_{X_2}} .$$

A number of different strategies have been proposed for producing parallel tests, some of which require testing on more than one occasion, others of which require testing on only one occasion.[21] With regard to the former there are the *split-half, alternate forms*, and *test-retest* methods, each of which aims to provide an estimate of the reliability of a test by correlating two versions of the test.[22] The most well known of the split-half coefficients is the Spearman-Brown "correction" (cf. Brown, 1910; Spearman, 1910),

[21] Unfortunately, the different methods for producing parallel tests have come, wrongly, to be viewed as different *types* of reliability. However, under CTT there exists only one "type" of reliability, and that is, as expressed in (3.3), the ratio of true score variance to observed score variance.

[22] And, each of which introduces systematic variation that might bias the estimate.

(3.11) $$\hat{\rho}_{TT'} = \frac{2\hat{\rho}_{X_1X_2}}{1+\hat{\rho}_{X_1X_2}},$$

which gives an estimate of the reliability of a test, T, consisting of two parallel halves, X_1 and X_2. The general form of (3.11) for k tests is given by

(3.12) $$\hat{\rho}_{CC'} = \frac{k\hat{\rho}_{CC'}}{1+(k-1)\hat{\rho}_{CC'}},$$

which is an estimate of the reliability of a composite, $C = \sum_{j=1}^{k} x_j$, i.e., the sum of the k parallel parts, $X_1, X_2, ..., X_k$. With the alternate forms method, the aim is to create two parallel tests by producing two equivalent *forms* of the same test, estimating the reliability of each form by taking their correlation, and then estimating the reliability of the total test with Spearman-Brown. With the test-retest[23] method, the same test is administered at two different time points, and then the scores from the two administrations are correlated in order to give an estimate of the reliability of the total score.

Internal consistency methods, such as the commonly employed coefficients KR_{20}, and α (the latter of which, as previously mentioned, was first given by

[23] Even though it is commonly used as a coefficient of reliability, it has long been recognized that it is a misnomer to call the test-retest method a method of estimating reliability, as any changes in observed scores over time may be reflective of either inconsistency with regard to the measure, or fluctuations in true scores. Hence, test-retest coefficients, properly characterized, are estimates of both the reliability (i.e. precision of measurement) and stability of the attribute measured. These two components can be estimated separately with certain statistical software programs, e.g., LISREL.

Guttman), treat each item as an alternate form and, hence, require only one administration of a test. A computational formula for KR_{20} is given by

$$KR_{20} = \left(\frac{k}{k-1}\right)\left(1 - \frac{k\overline{pq}}{\hat{\sigma}_t^2}\right), \tag{3.13}$$

in which $\overline{pq}$ is the average of the observed item variances, and $\hat{\sigma}_t^2$ is an estimate of the variance of the total scores. This can be shown to be equal to the average of all the possible "split-half reliabilities" of a test consisting of k dichotomously scored items, and is equal to the reliability of the unweighted sum of the items if the items are parallel. The formula for obtaining estimates of coefficient α is

$$\hat{\alpha} = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum_{j=1}^{k}\hat{\sigma}_j^2}{\hat{\sigma}_t^2}\right), \tag{3.14}$$

in which the $\hat{\sigma}_j^2$'s are the observed item variances and $\hat{\sigma}_t^2$ is defined as above; (3.14) is the analogue of (3.13) for tests consisting of continuous items.

It is of note that despite the fact that these methods were all designed to serve the purpose of creating parallel tests, and will, if parallelism does in fact hold, give estimates of the reliability of a test (or, rather, of the unweighted composite of the items), parallelism has long been recognized as a property which is seldom realized in practice. However, as Guttman (1945) showed, (3.11) – (3.14) will give *lower bounds* to reliability when parallelism does not hold,

bounds which, in practice, "will often be usefully greater than zero" (Guttman, 1945, p. 258).

Classical test theory is concerned not only with the precision of measures, but also with the "validity" of measures, by which is meant (despite some recognition of the more superficial notions of *face validity* and *content validity*) concomitant variation of the test with some chosen *criterion*.[24] The aim of testing is to predict this criterion, and, hence, the validity of a test is judged simply by the accuracy of this prediction. Since a test can be correlated with potentially an infinity of different criteria, it has potentially an infinity of different validities; however, the general formula for estimating the validity (or validities) of a test from sample data is

$$\left|\hat{\rho}_{XY} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X \hat{\sigma}_Y}\right|, \tag{3.15}$$

in which Y is the criterion (whatever it may be) and X is the test score (usually an unweighted sum of the items) from which one would like to make predictions about Y. Hence, under the CTT framework, a test is a "valid" measure of the attribute it was designed to measure if the scores realized by testees are highly predictive of any one criterion or set of criteria, deemed by the individual

[24] Which, for example, could be "a record of outcome" (e.g., a rating, or grade) (Cronbach, 1960, p. 103), "some important form of behaviour that is external the measuring instrument" (Nunnally, 1967, p. 87), "a measure of the property [of interest] which is taken to be perfect" (Ghiselli, 1964, p. 338), or, more generally, "some other observable measurement" (Lord & Novick, 1968, p. 261).

researcher to be relevant given the particular context in which the attribute of interest is being measured.

In essence, classical test theory consists chiefly in the classical true score model (as described on pages 62-63) plus various indices of reliability and validity. Implicit to the set of theoretical results that comprise CTT is that, in practice, the job of the test analyst is (merely) to produce total scores for a sample of n examinees, estimate (a lower bound to) the reliability of the total score by calculating one of the classical indices, and then correlate the total score with some chosen criterion (or criteria) of relevance in order to assess validity.

A Summary of Modern Test Theory

Modern test theory, which is also commonly identified as "latent trait theory", or sometimes "item response theory", is anchored by the notion that the observed relationships among a set of items of which a test consists may be accounted for by some underlying, but (typically) essentially unmeasurable, attribute. In other words, the observations taken on a set of test items are seen to be imperfect reflections of the attribute for which the test is a measure, which itself is accessible only through its influence on those items of which the test is comprised. In particular, a great deal of what has come to be known as modern test theory has as its foundation the application of latent variable modelling to the theory of latent traits, or, more generally, to the theory of item responding,

the origins of which were described briefly in Chapter Two. In general, whether one is applying modern test theory principles to the construction of new measures or to the evaluation of the worth of pre-existing ones, MTT has, over the past five decades, undeniably changed the face of test theory. The theoretical results that have accumulated under the banner of "modern test theory" are far too vast to summarize in full here. However, one might view MTT as contributing the following features to general test theory: 1) the addition of more, and increasingly complex, models with which to analyze responding to test items; 2) a theory of compositing items into test scores; 3) a broader conceptualization of estimation of precision of measurement; and 4) the addition of a variety of techniques which may be used in the construction of tests. Each will now be described briefly in turn.

Modelling Item Responses

The mathematical foundations of MTT are firmly steeped in the theory of latent variable modelling, the general premise of which is that knowledge about unobservable variates of interest may be derived from sets of observable "indicators" of such variates. These are, in the jargon of latent variable modelling, the *latent* and *manifest* variates respectively. In other words, it is thought that since the latent variates are not observable they cannot be measured "directly", but only "indirectly" via a set of measured variates, each of which measures the latent variates in question with some degree of error. Latent

variable modelling, hence, offers a means by which researchers may gain insight into that which they wish to study, i.e., the unobservable variates, by linking, via a statistical model, the distribution of the manifest variates to the distribution of the latent variates. Then, certain implications may be derived and empirically tested on a sample of data drawn from a particular population. The general steps involved in building a latent variable model are as follows:[25]

1. The Joint Distribution of $\underline{X}$ and $\underline{\theta}$

Let $\underline{X}$ denote a random vector that contains the k items on some test T, and let all densities refer to some focal population **P**. Thus, the density $f_{\underline{X}}$ describes the joint distribution of the scores on the k test items when the members of **P** respond to these items. Let $\underline{\theta}$ denote the $m\times1$ vector containing the (random) latent variates. The first step in building a latent variable model involves a statement of the joint distribution of $\underline{X}$ and $\underline{\theta}$, the general form of which is

$$f_{\underline{X}\cap\underline{\theta}} = f_{\underline{X}|\underline{\theta}} f_{\underline{\theta}} \tag{3.16}$$

in which $f_{\underline{X}|\underline{\theta}}$ is the distribution of $\underline{X}$ conditional on $\underline{\theta}$, and $f_{\underline{\theta}}$ is the "prior distribution", or the unconditional distribution of $\underline{\theta}$.

[25] This description of the basic features of latent variable models borrows elements from Bartholomew and Knott's (1999) treatment.

2. *The Unconditional Distribution of $\underline{X}$*

It is the latent variates contained in $\underline{\theta}$ that are of interest to the researcher, as they are thought to represent the attributes under study; however, because $\underline{\theta}$ is unobservable ("unmeasurable", "imperceptible", etc.), and the researcher is only in possession of indicators, contained in $\underline{X}$, of the latent variates, the particular model-implied unconditional distribution of $\underline{X}$ must be derived by integrating with respect to $\underline{\theta}$ over the product of $f_{\underline{X}|\underline{\theta}}$ and $f_{\underline{\theta}}$, i.e.,

$$f_{\underline{X}} = \int_{-\infty}^{\infty} f_{\underline{X}|\underline{\theta}} f_{\underline{\theta}} d\theta, \tag{3.17}$$

in which $f_{\underline{X}|\underline{\theta}}$ and $f_{\underline{\theta}}$ are defined as above.

3. *Specific Latent Variable Models*

A given latent variable model is brought about when the modeler makes particular choices about the densities $f_{\underline{X}|\underline{\theta}}$ and $f_{\underline{\theta}}$, and (for parametric latent variable models) places certain restrictions on the parameters of these densities. Specifically, the researcher must decide on the following:

3a. The form of $f_{\underline{\theta}}$

For nonparametric models the form of the prior distribution can remain unspecified; however, as Bartholomew and Knott (1999) point out, there are two purposes for making some assumptions about the form of $f_{\underline{\theta}}$: 1) in order to use particular estimation methods (e.g., maximum likelihood) which require that the

prior distribution has a particular form such that stable estimates will be produced; and 2) in order to determine the form of the "posterior distribution", i.e., the distribution of $\underline{\theta}$ conditional on $\underline{X}$, $f_{\underline{\theta}|\underline{X}}$, and certain quantities derived from it, such as an estimate of $\underline{\theta}_p$, i.e., an estimate of the p^{th} individual's location with respect to the latent dimensions. Bartholomew and Knott (1999) note that two common choices for the form of the prior are the multivariate standard normal and multivariate uniform distributions.

3b. The form of $f_{\underline{X}|\underline{\theta}}$

It is generally supposed that any observed associations among the X_j are due strictly to the latent variates and, therefore, when conditioning on $\underline{\theta}$, those observed associations vanish. In formal terms this means that the multivariate distribution $\underline{X}$ conditional on $\underline{\theta}$ is equal to the product of the individual conditional distributions, i.e.,

$$f_{\underline{X}|\underline{\theta}} = \prod_{j=1}^{k} f_{X_j|\underline{\theta}} \,. \tag{3.18}$$

This specification is known as the *principle of local independence*, and has often been cited as the hallmark of latent variable modelling.[26]

[26] Although (3.18) is the usual choice for the distribution of $\underline{X}$ conditional on $\underline{\theta}$ among most users of latent variable modelling methods, Stout (1990, 2002) provides examples of latent variable models which do not define the density $f_{\underline{X}|\underline{\theta}}$ in terms of the principle of local independence.

3c. Specification of $f_{X_j|\underline{\theta}}$

Once the form of the multivariate distribution of $\underline{X}$ conditional on $\underline{\theta}$ has been specified, then a choice regarding the forms of the univariate distributions, X_j conditional on $\underline{\theta}$, needs to be made. The specific choice will depend in large part on the measurement properties of the manifest variates. For example, if the X_j are dichotomous, $f_{\underline{X}|\underline{\theta}}$ will be a discrete mass function, and the $f_{X_j|\theta}$ will be Bernoulli distributed. If $\underline{X}$ contains variates with metric properties, however, then the choice regarding the form of $f_{X_j|\underline{\theta}}$ may be less straightforward. For the present purposes, two points on this issue are noteworthy: 1) that the researcher must choose *some* "reasonable" form for $f_{X_j|\underline{\theta}}$,[27] and 2) that the $f_{X_j|\underline{\theta}}$ often takes, but need not take, the same form for all j.

3ci. Item response functions

Once the forms of the univariate conditional distributions, $f_{X_j|\theta}$, have been chosen, the researcher is then left with a choice regarding the forms of the conditional mean functions, $E(X_j|\underline{\theta})$, or *item response functions (IRF's)*. As above, decisions with regard to the choice of the form of the IRF's will be constrained by the measurement properties of the X_j and $\underline{\theta}$, but, once again,

[27] Bartholomew and Knott (1999) propose a unified treatment of latent variable models, in which they claim that the $f_{X_j|\underline{\theta}}$ can always be represented as a member of the one-parameter exponential family, and, hence, as some linear function of the latent variates.

generally the researcher 1) must choose among some set of reasonable candidates for $E(X_j \mid \underline{\theta})$, and 2) the choice is typically, but need not be, consistent across the X_j.

4. Restrictions on Parameters

A given latent variable model is characterized by a set of q model parameters. A final step in specifying a particular latent variable model involves, at least for parametric models, decisions regarding which, if any, constraints will be imposed on the parameters of the chosen $f_{\underline{\theta}}$ and $f_{\underline{X}|\underline{\theta}}$. Some parameters may be fixed to particular values, others constrained to be equal, and others left free to vary, in which case their values will need to be estimated.

Ultimately, a latent variable model is a claim about a set of s parameters of $f_{\underline{X}}$. Once the researcher has specified the particular latent variable model, and ensured that it is identified (i.e., established that there exist unique values for the free parameters of the model), then the researcher may employ any of a number of different fit statistics, or other indices of fit, in order to assess whether the data are in keeping with the model. If the data do, in fact, conform to the model, the researcher then can make inferences about the latent variates, and, hence, about the attribute(s) of interest, in reference to the population under study.

Latent Variable Test Theory Models

As aforementioned, the contours of MTT are shaped in large part by latent variable modelling theory, the general features of which have just been outlined. In particular, a great deal of what has come to be known as "modern test theory" has as its foundation the application of latent variable modelling to the theory of latent traits, or, more generally, to the theory of item responding, the origins of which were described briefly in Chapter Two. The theoretical results that have accumulated in this domain are far too vast to summarize here. However, the essence of modern test theory may be adequately captured in a description of the following key components:

The latent trait

Latent trait theory is premised on the notion that all individuals may be characterized as "possessing" more or less of a single trait (attribute, ability, property, etc.) of interest. However, the possibility of quantifying the amount possessed by each individual is precluded by the fact that the trait is unobservable, and, hence, unmeasurable. Nevertheless, even though the trait is not amenable to "direct" measurement, it is thought to have predictable (but not perfectly so) relationships with other variates, which can be measured. Hence, any collection of such variates, each of which is an error-laden measure of the trait, may be thought to constitute a test of the trait. Therefore, test "behaviour", i.e., the observed responses from a random sample of testees to the items of the

test, is thought to be "generated" by the latent trait, and the issue then becomes one of ascertaining whether this proposition is tenable, and, if so, with what degree of precision do the test items measure the trait of interest.

Latent trait models

Latent trait models[28] constitute a class of latent variable models with the following features:

1. Unidimensionality – since it is generally supposed that a single trait[29] underlies test performance, $\underline{\theta}$, which represents the set of m latent traits in any given latent trait model, is often presumed to have only one element, and hence, θ is scalar-valued.
2. Local independence – as with latent variable models in general, latent trait models typically specify the condition of local independence, i.e., that observed relationships among the measured variates (in this case, test items) are due strictly to the (single) latent trait. This means that, for a given latent trait model, the multivariate distribution of the test

[28] It is generally supposed that the expression "trait" is reserved for variates which are continuously distributed and latent trait theory generally adopts this interpretation. The present work, however, does not exclude, at least potentially, the use of latent variable test theory models for which the measurement properties of the latent variate are not strictly interval- or ratio-level. Here, by "latent trait model" is meant any latent variable model used for the purpose of pronouncing on the quality of a test. Hence, the expression "latent trait model" will be used synonymously with "test theory model".

[29] Although multidimensional latent trait models, i.e., those that are applicable to situations in which there is more than one latent variable, have been considered (cf. Mulaik, 1972, Samejima, 1974, Stout, 2002), here the description of latent trait models will be limited to the subclass of *unidimensional* models for reasons that will be elaborated in Chapter 5.

items conditional on θ is given by the product of the individual distributions for any fixed θ.

3. Item characteristic curves—each latent trait model involves a specification of the form of the IRF's, with the particulars of the specification being constrained by both the item response formats and the distributional assumptions as regards θ.

The application of latent variable modelling to test evaluation has had a number of notable implications for test theory. One of the greatest gifts of MTT, to which almost all other advantages are tied in some way, is the inclusion of more precise statistical models for describing item responding, which explicitly link the amount of attribute (trait, ability, property, etc.) possessed by individuals to their responding to a set of items which make up a test of the attribute in question. The primary implication of employing a latent variable model as a test theory model is the ability to specify particular relations between test items and the underlying attribute (for which the latent variate is a proxy) based on theoretically-derived expectations of how the items should perform given these relations, and then formally test whether the data are or are not in keeping with the particular model at hand. This means that any latent variable *test theory* model may be employed to formally test that a given set of items "measure but one thing", presumably the trait purportedly measured by the test. Of course, a technical paraphrase of "the items measure but one thing" is that the

responses to the items of a test from a random sample of examinees are "unidimensional" in a particular sense, the latter of which is specific to the model under consideration. Hence, MTT has provided a formal treatment for justifying that a set of responses to the items of a test may be conceptualized as measuring a single attribute (trait, ability, property, etc.) in common.

A Theory of Compositing

Whereas in the classical treatment it was merely presumed that the best means of compositing the responses to the individual items of a test into a test score was to simply sum the item responses for an individual together,[30] a substantial portion of modern test theory has been concerned with the study of composites and the various criteria for defining a particular composite as "optimal". The rationale underlying this body of theoretical results is as follows:

1. If a set of items has been shown to conform to a (unidimensional) test theory model, then the items can be justifiability composited.
2. A given composite, ϕ^*, is simply a function of the test items, say, $f^*(X_1, X_2, ..., X_k)$.
3. There exists an infinity of possible composites of the items of a given test.

30 Or, in certain cases, produce a weighted sum.

4. The class of composites of the items of a given test may conveniently be divided into two sub-classes: the linear composites and the nonlinear composites.

   i. Linear composites: These composites are formed as weighted sums of a set of test items, i.e.,

(3.19) $$\phi = \sum_{j=1}^{k} w_j X_j \, .$$

   ii. Nonlinear composites: These composites are formed as nonlinear functions of a set of test items. Two commonly employed examples of such composites are the *maximum likelihood estimators* (MLE)[31] and the *expectation a posteriori* (EAP) *predictors* of θ. As regards the former, the maximum likelihood estimator $\hat{\theta}_p$ for the p^{th} examinee with a given response profile is obtained by maximizing the log likelihood function. For example, for a set of dichotomous items, the maximum likelihood estimator $mle(\hat{\theta}^*)$ for a particular response pattern $\underline{X} = \underline{x}^*$ is the solution to the equation

(3.20) $$\frac{\partial}{\partial \theta} \log P(\underline{X} = \underline{x}^* \mid \theta) = 0$$

[31] Bartholomew (1981) notes that to call these "estimators" is to use a misnomer, as estimators are typically taken to be "best" guesses, based on sample data, of the value of some population *parameter* of interest; θ, however, is a random variate, and not a parameter.

(cf. Birbaum, 1968), in which $P(\underline{X} = \underline{x}^* \mid \theta)$, the likelihood of the data given θ, is equal to

$$\prod_{j=1}^{k} P(X_j = x_j^* \mid \theta)^{X_j} [1 - P(X_j = x_j^* \mid \theta)]^{1-X_j} . \tag{3.21}$$

The corresponding EAP estimate $eap(\theta)$ is given by

$$E(\theta \mid \underline{X} = \underline{x}^*) = \int_{-\infty}^{\infty} \theta f(\theta \mid \underline{X} = \underline{x}^*) d\theta = \int_{-\infty}^{\infty} \frac{\theta P(\underline{X} = \underline{x}^* \mid \theta) f(\theta)}{h(\underline{X} = \underline{x}^*)} d\theta , \tag{3.22}$$

in which $P(\underline{X} = \underline{x}^* \mid \theta)$ is defined as in (3.21), $f(\theta)$ is the unconditional density of θ, and $h(\underline{X} = \underline{x}^*)$ is the unconditional density $\underline{X}$ evaluated at $\underline{x}^*$ (cf. Muraki and Engelhard, 1985).

5. The fact that there are an infinity of possible composites of a given set of test items implies a need for criteria for singling out one particular composite as preferred. Only given antecedently specified criteria that define senses of optimality can a particular composite be judged as "optimal" in a given context of test application.
6. Optimality criteria may usefully be divided into classes, *mathematical optimality* and *practical optimality.*

   i. Mathematical optimality:

   In general, there are two different sub-classes of mathematically-based criteria for choosing a given compositing rule as preferred. The first consists in a comparison of the measurement precision delivered by each

resulting composite. The second consists in specifying a general statistical principle of prediction or estimation, for example, the principle of maximum likelihood, under which a specific composite can be singled out as optimal in the particular sense defined by the principle. A comparison of the measurement precision delivered by a set of candidate composites goes under the general heading of *efficiency* considerations, and rests on a consideration of *theoretical information*,[32] which quantifies the amount of information that an observable random variate, $\underline{X}$, contains about an unobservable parameter, γ, upon which the distribution of $\underline{X}$, depends. Information is proportional to the reciprocal of the width of the confidence interval for a sufficient and efficient estimator of γ, and may be defined generally as

$$I(\gamma) = \sum_{j=1}^{k} \frac{E'(X_j \mid \gamma)^2}{V(X_j \mid \gamma)}, \tag{3.23}$$

for any set of k observable random variates, each of which is considered to be a measure of γ. $I(\gamma)$ is a function γ and, hence, its values may vary along the γ-dimension.

In the context of the item response modelling that is the hallmark of modern test theory, the attribute that is purportedly measured by a set of

[32] Also known as "Fisher information" after R.A. Fisher, who invented the concept.

test items is represented by a random, latent variate θ, and the (test) information function may be represented as

$$I(\theta) = \sum_{j=1}^{k} \frac{E'(X_j \mid \theta)^2}{V(X_j \mid \theta)}, \tag{3.24}$$

a single-peaked function for which larger values are associated with greater precision of measurement of the attribute represented by θ. The test information function gives an upper bound to the information that can be delivered by *any* possible composite of the test items. The information function for a given (linear or nonlinear) composite, ϕ^*, may be defined generally as

$$I(\theta, \phi^*) = \frac{\left[E'(\phi^* \mid \theta)\right]^2}{V(\phi^* \mid \theta)}, \tag{3.25}$$

which is also a single-peaked function of θ.

Two candidate composites, ϕ_1 and ϕ_2, can be compared by calculating the *relative efficiency function*

$$RE(\theta, \phi_1, \phi_2) = \frac{I(\theta, \phi_1)}{I(\theta, \phi_2)}, \tag{3.26}$$

this also a function of θ.[33] Values of (3.26) that are greater than unity indicate that ϕ_1 has greater precision of measurement than does ϕ_2, for those values of θ. Composite ϕ_1 should then be preferred to ϕ_2 if the relative efficiency function is greater than unity in a range of θ that corresponds to the relative attribute levels of the population of individuals whose measurement is of interest.

ii. Practical optimality:

Practical optimality refers to non-statistical criteria typically associated with the local details of test usage. A prime example is the ease of calculation of a composite. Nonlinear composites, for example, are more difficult to calculate than weighted sums. If, as is sometimes the case, weighted sums have similar statistical properties to their nonlinear brethren, then little may be lost in terms of *mathematical optimality*, while much is gained in convenience.

A Broader Conception of Precision of Measurement

An often cited advantage of modern test theory is its emphasis on the local nature of reliability estimation, specifically for composites which are functions of items conforming to models in which the IRF's are nonlinear. Hence, a

[33] Note, equation (3.26) also has application in comparing the relative efficiency of two different tests,

composite may be a more or less "reliable" measure of θ for a given examinee, depending on the "amount of attribute possessed" by the individual examinee. There are three notable implications to this fact: 1) Assessing the precision of a given composite will be contingent on specifying the general expected range of θ in the population from which the sample, on which an estimate of precision will be based, is drawn; 2) a given composite will be a precise measure of the attribute of interest (as represented by θ) for some test takers and not for others; and 3) comparisons can be made with regard to the relative precision with which two different composites of the same set of test items (for a given population) measure the attribute in question.[34]

As indicated above, for a given compositing rule (and, hence, the corresponding composite score it produces), the precision with which the composite[35] measures a fixed $\theta = \theta_0$ (a proxy for a specific "amount of attribute possessed") is quantified by the amount of information contained in the composite, which may be obtained by estimating (3.25) for $\theta = \theta_0$. As implied by (3.25), the precision of the composite will be greatest for values of θ at which the slope of the "test response function" (i.e., $E(\phi \mid \theta)$) is steep relative to the variance

each of which is scored into composites of some form, for measuring θ, and, hence, a given attribute of interest.

34 And, also, two different tests of a given attribute, each of which is scored with a particular, but possibly different, composite.

35 With the proviso that compositing in such a manner is justified, a topic which will be emphasized in Chapter 5.

of ϕ at θ. However, for models in which the IRF's are linear, precision will not vary across different values of θ, i.e., $I(\phi \mid \theta)$ will be constant across the θ-range.

Alternatively, rather than conceiving of precision of measurement in terms of information, it may be defined in terms of nonlinear reliability. The following equation

$$\rho^2_{\phi\theta} = 1 - \frac{E[V(\phi \mid \theta)]}{V(\phi)}, \tag{3.27}$$

gives a lower bound to the reliability of ϕ (lower bound because θ is not the true score variate of ϕ). There exist specific forms for both (3.25) and (3.27) which are commonly used in conjunction with particular test theory models.

Modern test theory essentially consists of a set of latent variable models, each of which is founded upon a common core definition of unidimensionality, but with each tied to a particular conceptualization of the relationship between the test items and θ. In the hands of MTT, these models become "measurement models" through the identification of θ with the attribute for which the items were designed to be indicators. MTT employs these measurement models in support of its attempt to: 1) mathematically model item/attribute regressions; 2) test the hypothesis that a set of items measure but one thing; 3) determine a composite of a set of items that optimally estimates (predicts) the standing of individuals on the attribute for which the items were designed to be indicators; and 4) quantify the local precision (i.e., conditional on values of the attribute)

delivered by a composite conceptualized as yielding measurements of the attribute for which the items were designed to be indicators. In essence, MTT has added a sophisticated and complex set of mathematical tools to the test analytic game, however, it has done so without drastically changing the structure of test analysis as a whole, a point to which I will return at the end of the current chapter.

Two Commonly Employed Latent Variable Test Theory Models

1. A Two-parameter item response model

Let $\underline{X}$ denote a $k\times1$ vector of random variates whose distributions contain the scored responses of the members of focal population **P** to the k dichotomous items of which test T consists. The X_j are called manifest variates and each may assume two values: 1=endorsed, 0=not endorsed. Also, let θ denote an unobservable (random) latent variate. Let $\underline{x} = (x_1, x_2, ..., x_k)$ stand for a particular realization of $\underline{X}$, there being 2^k such "response patterns". Finally, let $P(\underline{\boldsymbol{X}} = \underline{x})$ be the proportion of objects in a population **P** with a given response pattern. It is the 2^k proportions that are modelled by a 2-parameter IR model.

As with all latent variable models, a 2-parameter IR model begins with a statement about the joint distribution of $\underline{X}$ and θ as the product of the distribution of $\underline{X}$ conditional on θ, $f_{\underline{X}|\theta}$, the prior distribution, f_θ, and a specification of the forms of f_θ and $f_{\underline{X}|\theta}$. Then the unconditional distribution of

$\underline{X}$ is derived by integrating, with respect to θ, over the product of $f_{\underline{X}|\theta}$ and f_θ. The particular choices of f_θ and $f_{\underline{X}|\theta}$, along with certain restrictions on the parameters of these distributions will generate a 2-parameter IR model.

1.1 Specification of the form of f_θ

Typically the prior distribution f_θ is specified to be

(3.28) $$N(0,1).$$

1.2 Specification of $f_{\underline{X}|\theta}$

As a consequence of the dichotomous [0/1] response format for each X_j,

(3.29) $$f_{X_j|\theta} = P(X_j = x_j \mid \theta).$$

Since the items of T are presumed to measure in common one thing, the usual property of local independence is specified for $f_{\underline{X}|\theta}$ with θ scalar-valued, i.e.,

(3.30) $$f_{\underline{X}|\theta} = \prod_{j=1}^{k} P(X_j = x_j \mid \theta) = \prod_{j=1}^{k} P(X_j = 1 \mid \theta)^{x_j} (1 - P(X_j = 1 \mid \theta)^{1-x_j}.$$

1.3 Conditional first and second moments of $f_{\underline{X}|\theta}$

For 2-parameter (and certain other) IR models for dichotomous, [0/1] variates, the conditional mean function (i.e., the IRF) for X_j is equal to the probability of an "endorsement" conditional on θ, i.e.,

(3.31) $$E(X_j \mid \theta) = P(X_j = x_j \mid \theta),$$

and the conditional variance is equal to

(3.32) $$P(X_j = x_j \mid \theta)\left[1 - P(X_j = x_j \mid \theta)\right].$$

The IRF's are typically modelled with one of two possibilities: either with the two-parameter logistic function (cf. Birnbaum, 1968),

(3.33a) $$P(X_j = x_j \mid \theta) = \Phi(a_j(\theta - b_j)),$$

or with the normal ogive (cf. Lord, 1952),

(3.33b) $$P(X_j = x_j \mid \theta) = \frac{e^{a_j(\theta - b_j)}}{1 + e^{a_j(\theta - b_j)}},$$

in which the a_j and b_j are the item parameters. For a given item, a_j is the value of θ at which the probability of an endorsement is .50, and b_j is the slope of the IRF at its inflexion point (i.e., the value of θ at which the probability of an endorsement is .50).

1.4 The model parameters of the 2p IRT model

The model-implied probabilities for the 2^k response patterns, $\underline{x}$, are then given by

(3.34) $$\widetilde{P}(\underline{X} = \underline{x}) = \int_{-\infty}^{\infty} \prod_{j=1}^{k} \mathrm{P}(X_j = 1 \mid \theta)^{x_j} \left[1 - P(X_j = 1 \mid \theta)\right]^{1 - x_j} f(\theta) d\theta.$$

The 2^k response pattern proportions can be estimated by their sample counterparts, $\widehat{P}(\underline{X} = \underline{x})$, and based on these estimates, the 2^k model parameters estimated. If a set of model parameters can be found such that the model-implied proportions are "close" to their sample counterparts (with the sense of

"close" defined by some fit function), then the k items may be said to be unidimensional (in a 2-parameter IR sense).

Because for 2-parameter (and other) IR models, the IRF's are nonlinear, precision of measurement will vary over the θ-range, and, hence, may be assessed locally, i.e., for particular values of θ. For any weighted-sum composite, i.e., $\phi = \sum_{j=1}^{k} a_j X_j$, then the precision of ϕ may be quantified by the information function:

$$I(\phi|\theta) = D^2 \sum_{j=1}^{k} a_j P(X_j = 1|\theta)[1 - P(X_j = 1|\theta)], \tag{3.35}$$

in which D=1.71, the weights, a_j, are defined as in (3.33a) and (3.33b), and $P(X_j = 1|\theta)[1 - P(X_j = 1|\theta)]$ is the variance of X_j conditional on θ.

2. The unidimensional, linear common factor model

Let $\underline{X}$ denote a $k \times 1$ random vector that contains random variates whose distributions contain the scored responses of the members of a focal population **P** to the k (continuous or pseudo-continuous) items of which test T is comprised and let θ denote a continuous, unobservable (random) latent variate. As with

the 2-parameter IR model, the linear common factor (ULCF) model[36] begins with a specification of the forms of f_θ and $f_{\underline{X}|\theta}$.

2.1 Specification of the form of f_θ

Although a general formulation of the ULCF model may be given without a complete statement of the forms of f_θ and $f_{\underline{X}|\theta}$, for the present purposes it will be noted that a common choice for f_θ is

(3.36) $$N(0,1).$$

2.2 Specification of $f_{\underline{X}|\theta}$

Application of the ULCF model typically involves placing the following constraints on the parameters of $f_{\underline{X}|\theta}$: 1) that the conditional mean functions be equal to $\underline{\Lambda}\theta$; and 2) that the conditional covariance matrix be diagonal and positive definite. If conditional normality is further required,

(3.37) $$f_{\underline{X}|\theta} \sim N_k(\underline{\Lambda}\theta, \Psi),$$

in which $\underline{\Lambda}$ is a $k\times 1$ vector of linear regression weights of the X_j on θ, the "common factor", and Ψ is a $k\times k$ diagonal, positive definite matrix of variances of the "specific"[37] factors. It follows under conditional normality that

[36] Note that for parsimony the ULCF model is presented here as one particular latent variable model, when, in fact, there exist a number of distinct ULCF models from which one may chose, all of which, however, share certain features, such as the property of local independence with θ scalar-valued and linearity of item response functions.

[37] In test analytic applications, the "specific" factors are usually taken to be some mix of measurement error and variation specific to each of the observed test items.

(3.38) $$f_{\underline{X}|\theta} = \prod_{j=1}^{k} f_{X_j|\theta} .$$

2.3 Conditional first and second moments of $f_{\underline{X}|\theta}$

From (3.37), the conditional first and second moments of $f_{\underline{X}|\theta}$ are respectively

(3.39) $$E(\underline{X} \mid \theta) = \underline{\Lambda}\theta$$

and

(3.40) $$C(\underline{X} \mid \theta) = E[(\underline{X} - \underline{\Lambda}\theta)(\underline{X} - \underline{\Lambda}\theta)' \mid \theta) = \Psi ,$$

and, from (3.38) and (3.39), the univariate IRF's are given by

(3.41) $$E(X_j \mid \theta) = \lambda_j \theta \quad \forall j,$$

in which λ_j is the linear regression weight of the j^{th} item on the latent factor, θ. In other words, the conditional mean functions of each X_j are linear in θ.

2.4 The model parameters of the ULCF model

The parameters of the ULCF model are contained in $\underline{\Lambda}$ and Ψ. However, as with all applications of latent variable models, the researcher is in possession only of $\underline{X}$, and is typically interested in deriving implications with regard to some element of its distribution. In the ULCF model case, the parameter of the unconditional distribution of $\underline{X}$ that is of interest is its covariance matrix, Σ. However, in order to link the distribution of $\underline{X}$ to the distribution of θ, Σ must

be represented in terms of the model parameters. Specifically, from (3.36) and (3.37) it follows that

$$\Sigma = C(E(\underline{X} \mid \theta)) + E(C(\underline{X} \mid \theta)) = \underline{\Lambda}\underline{\Lambda}' + \Psi. \tag{3.42}$$

As with the 2-parameter IR model presented above, in order for the test analyst to claim that her test data are in keeping with a ULCF model, she must assess, by some appropriate criterion or set of criteria, whether the model "fits" the test data. Specifically, she must ascertain whether there exist values for the elements of $\underline{\Lambda}$ and Ψ (i.e., the parameters of the model) which will reproduce a model-implied covariance matrix of $\underline{X}$, $\tilde{\Sigma}$, that is "close" to Σ, the population covariance matrix of $\underline{X}$, the latter of which is estimated by test data.

If it can be shown that the ULCF model is a good fit to the observed responses to the k items of T, then, once again, the test analyst is justified in compositing the items, and estimating the "reliability" of the resulting composite, ϕ. The exact nature of that estimate will depend on which specific ULCF model was employed; however, a general formula for estimating a lower bound to the reliability of $\phi = \sum_j w_j x_j$, in fact, a special case of (3.27), is

$$\hat{\Omega}_\phi = 1 - \frac{\underline{w}'\Psi\underline{w}}{\underline{w}'\hat{\Sigma}\underline{w}} = \frac{\underline{w}'\hat{\underline{\Lambda}}\hat{\underline{\Lambda}}'\underline{w}}{\underline{w}'\hat{\Sigma}\underline{w}}, \tag{3.43}$$

in which $\underline{w}$ is a vector of weights, and $\hat{\underline{\Lambda}}$ and $\hat{\Sigma}$ are sample estimates of $\underline{\Lambda}$ and Σ respectively (cf. Heise and Bohrnstedt, 1971).

Additional Contributions of Modern Test Theory

In the domain of test construction MTT has contributed substantially to areas involving such topics as item selection, test equating, tailored testing, and so on. Although these topics are considered to be somewhat tangential to the focus of the present work, which is an analysis of test theory as it applies to previously existing tests, a brief mention of a number of common applications of modern test theory principles and techniques is warranted given the breadth of theoretical results which have accumulated in the past couple of decades on such topics.

Item calibration

In classical item analysis, the characteristics of the individual item are of interest only through the effect of the item on the total test score (Lord and Novick, 1968). In addition, classical techniques of item analysis and selection typically focus on strategies for choosing sets of items which will produce parallel measures. The problem with such procedures is that they are not invariant across populations of examinees. The modern test theoretic notion of information, however, makes way for an approach to item analysis which allows for characterizing items in terms of latent trait parameters, and, thereby, provides the test developer with sample-invariant item parameters (Hambleton et al., 1978). Specifically, the test developer may examine item information functions, and then select particular items according to the amount of

information they contribute (for a specific range on the latent variate) to the total amount of information provided by the test (Hambleton et al., 1978). Lord (1977; cited in Hambleton et al., 1978) outlined an item selection framework which involved 1) specifying the shape of a desired test information curve (which he called the "target information curve"); 2) selecting items with item information curves that "fill up" the hard-to-fill areas of the target information curve; 3) after the addition of each item, calculating the test information curve for the selected items; and 4) continuing to select items until the calculated test information curve approximates the target information curve to a satisfactory degree.

Item banking

An item bank consists of a (usually large) collection of items having particular characteristics (i.e., items for which the item parameter estimates are known), which may be made available to test constructors looking for items with particular properties. The invariance property of the latent trait item parameters makes it possible to compare item statistics for samples coming from dissimilar populations (Hambleton et al., 1978). Ideally, an item bank should contain a sufficient number of highly discriminating items with difficulty parameters spread throughout a large range on the latent trait dimension. Two further practices which arise from the existence of item banks are *tailored testing* and *test equating*.

Test score equating

Test score equating refers to the process of matching individuals' test scores (each of which results from the application of some compositing rule) on two (or more) tests of the same attribute. A distinction is made between *horizontal* and *vertical* equating methods.[38] In horizontal equating, the test forms are expected to be comparable with respect to difficulty level (i.e., the expected values of the test scores with respect to some target population); here, the equating adjusts for unintended differences in difficulty or in the latent trait distributions underlying each of the forms (Slinde and Linn, 1977; cited in Hambleton et al., 1978). Vertical equating refers to the process of equating test forms which are constructed to differ in difficulty. McDonald (1999) summarized three different methods of test equating: true-score equating, linear equating, and equipercentile equating, and claimed that true-score equating is the most precise and most informative of the three.

Test tailoring

The motivation behind tailored testing is that testing is typically done in settings in which a group of individuals take the same (or a "parallel") test; however, since individuals will typically differ with respect to the attribute underlying the test (and, hence, also with respect to their position on the latent

[38] Although McDonald (1999) claims that this distinction might be better represented by a continuum of degree of difference in difficulty.

trait dimension), in certain circumstances the test giver would like a test tailored to each individual examinee's position with regard to the latent trait dimension. Specifically, individuals would ideally be presented with items such that the probability that the individual answers the item is .50, for each and every item. Latent trait modelling is especially suited to developing such tests because θ estimates are independent of the particular set of items administered, and, hence, examinees can be compared despite having responded to items of different average difficulties.

A common application of tailored testing is *computerized adaptive testing* (CAT). Computerized adaptive tests are tests which are administered by computer and are adapted specifically to the individual examinee such that the items presented to the examinee are neither too difficult nor too easy (Embretson and Reise, 2000). The goal of CAT is to administer a set of items that are in some sense maximally informative and efficient for each individual examinee (Embertson and Reise, 2000).

Analysis of differential item functioning

Differential item functioning (DIF) describes a situation in which item responding differs between members of two (or more) groups despite the fact that the two groups are equally matched with respect the attribute of interest, an admittedly undesirable property, and one which ideally needs to be detected and identified prior to widespread use of a set of items with this property. DIF is

said to occur when a given item does not have the same relationship with the latent trait under study across two or more populations of examinees. Specifically, an item is defined as showing DIF if its IRF differs for two or more populations; hence, examinees who are equal with respect to the latent trait do not have the same probability of giving a particular response on the item (Embretson and Reise, 2000). Modern test theory principles may be used to detect and model DIF, typically with the aim of eliminating those items from a pool of potential items (or, if the set of items in which the dubious item occurs already constitutes a measure, eliminating the contribution of the item when scoring the test).

What has been presented thus far is a description of the theoretical results and mathematical foundations of classical and modern test theories. It has been concluded that: 1) *classical test theory* consists primarily in the classical true-score model plus a host of indices of the reliability and validity, values which can be estimated from of a set of test scores from a sample of examinees drawn from a particular population; and 2) *modern test theory* has offered to the practice of test analysis the employment of latent variable models that can be taken as representing the relationship between test item responding and the attributes whose measurement is of interest, precise model-based definitions of item-set homogeneity centring on the concept of unidimensionality, the ability to

efficiently test hypotheses of unidimensionality, a theory of model-based compositing rules, and of model-based estimates of precision of measurement.

The following chapter consists in a summary of the findings from a systematic examination of research studies published in a sample of peer-reviewed journals in which test analyses frequently appear, over a specified time period. The aim in painting a picture of how researchers use test theory tools is to show that test analytic practices are not in general guided by a sound logic, and, furthermore, that the employment of the implements born out of CTT and MTT are frequently misunderstood and misused. Ultimately the goal is to contrast the current state of affairs with regard to test analytic practices with a proposal for a logical, sequential framework for analyzing tests, the latter of which is elaborated in full in Chapter Five. Chapter Six consists in a discussion of certain of the more prominent confusions inherent in current applied test analytic practice as contrasted to the proposed test analytic framework.

4. CURRENT TEST ANALYTIC PRACTICES

Method

One of the goals of the present work is to examine current practices pertaining to how researchers (as opposed to test constructors) evaluate the soundness of the measures they employ. To this aim, a sample of articles in peer-reviewed journals with a high probability of containing articles in which test analyses have been conducted were reviewed. Specifically, I reviewed a subset of the articles appearing in five peer-reviewed journals, for the time period spanning January 2003 through December 2004 (see Appendix 1).[39] The intended aim was not to catalogue *all* practices in the domain of test analysis. Rather, it was to obtain a *picture* of current test analytic practices that could be used, in light of the logical framework that is developed in Chapter Five, to both exemplify bad test analytic practice, and provide an indication of its prevalence.

[39] The author recognizes the non-randomness of such a sample, but defends the approach on the basis of the fact that one of the substantive aims of the proposed work is to exemplify certain common practices among researchers conducting test analyses, rather than merely generalizing findings to a larger population.

Procedure

Articles appearing in five peer-reviewed journals between January 2003 and December 2004 were reviewed (see Appendix 1). These particular journals were chosen for review due to the high likelihood that they would contain studies involving some aspect of test analysis. The articles in all volumes published in 2004 were examined for three of the journals, the articles for the half the volumes published in 2004 were examined for one of the journals, and the articles in all volumes published between 2003 and 2004 were examined in the fifth journal (see Appendix 1).[40]

Initially, a total of 416 articles were examined; each was assessed for whether or not it addressed any aspect of test analysis, e.g., item analysis, examination of association structure (i.e., assessment of the fit of particular statistical models), reporting of precision estimates, or analysis of validity. Announcements, editorials, book reviews, biographies, non-English articles, and articles which did not involve any test analysis were not examined. Of the 251 remaining articles, studies were divided into those that involved some examination of the structure of the test items, those that provided reliability estimates but no examination of structure, and those that referenced previous test evaluation findings, but neither examined the structure of the items, nor

[40] The rationale for examining only half the volumes of *Personality and Individual* Differences and an additional year of volumes of the *Canadian Journal of Behavioural Science* was to keep the number articles examined per journal roughly the same.

provided reliability estimates. Those studies which evaluated the structure of the items and/or provided reliability estimates from the data generated in the study were more thoroughly examined for details regarding (see Appendix 2):

1. Whether the aim of the analysis (i.e., exploratory or confirmatory) was identified;
2. whether the researcher(s) was explicit with regard to how many attributes the test is expected to measure, and whether this feature of the test was assessed (i.e., the "dimensionality" or "structure" of the test is examined);
3. the nature of the statistical model employed, and whether a formal test of model fit was conducted (if applicable);
4. if and how the items were composited (i.e., what was the nature of the compositing rule employed to produce test scores), and whether this was appropriate for situation at hand;
5. what, if any, indices were used to estimate the "reliability" (i.e. precision) of the test;
6. if and how the issue of validity was handled (if applicable);
7. whether the analyses appeared to be guided by an explicit logic.[41]

[41] This, admittedly subjective, component of the review was intended as an opportunity to assess whether there exists in applied test analyses in general a consistent, logically sound framework according to which analyses are conducted.

Results

Of the 251 articles subject to a more thorough review, 213 (85%) assessed some aspect of the measure(s) employed (e.g., estimated precision, examined association structure, estimated validity coefficients). Of these, 79 (37%) involved assessment of both structure and precision; 12 (6%) examined structure, but did not analyze precision; 109 (51%) analyzed precision, but did not examine structure, 9 (4%) neither examined structure nor analyzed precision, and 4 (1 %) articles did not provide enough information to determine whether either structure or precision was assessed (Appendix 3 provides a summary of some of the key findings described below).

A Summary of Current Test Analytic Practices

1. Aim of Analysis

Of the 251 articles subjected to thorough review, 85 (34%) made an explicit statement as to the aim of the study, and the analyses conducted therein. The stated aims could be classified as being either "exploratory" or "confirmatory" in nature: In 32 (38%) of the studies, researchers indicated that the motivation of the study was to *confirm* that the test measures a given attribute (or set of attributes) in a manner consistent with some expectation (e.g., based on previous research, or, more generally, received theory regarding the attribute(s) which the test is designed to measure); 6 (7%) of the studies indicated an exploratory aim, i.e., to

"discover" what the test measures. In the remaining 47 (55%) studies, the motives of the researchers could be best described as consisting in a blend of confirmatory and exploratory aims, the latter of which was evidenced by some statement indicating that the aim of the study was, in part, to *explore*, i.e., to "discover", "investigate", "examine", or "determine", what and how the test measures.

The remaining 166 articles contained no explicit statement as to the aim of the analysis. Of these, 43 (26%) included neither evaluation of structure, nor assessment of the precision of the measures employed; 104 (63%) included assessment of precision, but did not examine structure; in 3 (2%) studies, structure was examined, but no estimates of precision were reported; 14 (8%) studies examined structure, as well as estimated, by some means, the precision of the measures employed; 2 (1%) studies did not provide enough information in order to ascertain whether any aspect of structure was assessed or precision estimated.

a. A blend of confirmatory and exploratory approaches

In a number of test analyses examined in the current study, researchers were concerned with confirming that their test conformed to a particular pre-specified structure, but were, at the same time, willing to let the data guide them with regard to making decisions about how many attributes actually underlie a given test, and, so, did not test the fit of particular restricted m-dimensional

models. Rather, these researchers typically employed procedures such as principal component analysis (PCA) or exploratory factor analysis (EFA), through which the number of factors (presumably taken to be proxies for the attributes measured by a test) were "extracted" ("determined", "uncovered") by some method and compared with expectations. For example, Harvey, Pallant, and Harvey (2004) stated that the purpose of their study was to "investigate whether the six-factor structure of the Frost Multidimensional Perfectionism Scale could be *replicated*" (p. 1007; emphasis added), yet they employed a PCA in order to "*explore* the underlying structure of the scale" (p. 1011; emphasis added). Shaw and Joseph (2004) stated that the aim of their study was to "*replicate* Maltby and Day's...study", but conducted a PCA and concluded, consistent with their expectations, that the "Eigenvalue and Scree test criteria both *suggested* a three component solution" (p. 1427; emphasis added). Campbell-Sills, Liverant, and Brown (2004) used what they referred to as "an exploratory factor analysis with a CFA [i.e., confirmatory factor analysis] framework" because of a lack of a "strong empirical basis for [using] CFA...i.e., *consistent evidence* with regard to the appropriate number of factors" (p. 246; emphasis added); Williams and Paulhus demonstrated a mix of exploratory and confirmatory aims in their 2004 study which they claimed was conducted in order to "*uncover* the factor structure of the Self-Report Psychopathy (SRP-II) Scale", and in which they employed

"Exploratory factor analyses ...to *determine* whether the two-factor structure could be uncovered" (p. 768; emphasis added).

Rather than remaining strictly confirmatory in their approaches, some researchers sought to make sense of results which did not square with initial expectations, and reinterpreted the factor structure of the test in light of what they considered to be new evidence about what the pertinent test might really measure. Harvey et al. (2004), for example, ultimately decided to retain only four factors on the basis of their examination of a scree plot, and rather than concluding that their hypothesized (i.e., six-factor) structure did not appear to hold, they reinterpreted the factor structure of the test, and reorganized the items into four, rather than the original six, subscales.

b. Using data as a basis for making claims about what a test measures

Another indication of the presence of exploratory and confirmatory aims was researchers' employment of first a procedure, such as EFA, for *identifying* or *discovering* which and how many attributes were measured by the test (i.e., the underlying or latent "structure"), followed by a test (sometimes on the same, and sometimes on different, samples) of whether the observed test scores were in keeping with the identified structure. In test analyses falling into this category, researchers frequently made some initial claim as to what the test measures, but refrained from making a commitment to an a priori specified structure until after the data were examined, such that the data might inform as to which specific

hypothesis (i.e., regarding the structure of the items) should be formally tested. For example, researchers might have, on the basis of commonly employed criteria, such as the "number of eigenvalues greater than 1" criterion, or examination of scree plots, decided to "retain" a certain number of factors, and, if those factors could be reasonably interpreted, the test was thought to constitute a measure of those attributes represented by the interpreted factors. Then, a procedure such as confirmatory factor analysis (CFA) would be employed in order to formally test whether the previously identified structure could be replicated, and, hence, the test could be said to have "factorial validity".

For example, Currie, Cunningham, and Findlay (2004), following on the recommendations of Gerbing and Hamilton (1996; cited in Currie, et al., 2004), claimed that EFA is a "useful initial strategy to *determine the underlying* dimensional model", and then "Confirmatory factor analysis...is...used to evaluate the model *derived from EFA*" (p. 1057; emphasis added). In Woolley, Benjamin, and Woolley's (2004) study, it was decided on the basis of an "examination of rotated factor solutions, combined with...theoretical judgement" that "a four-factor structure offered the most *parsimonious solution* for explaining the interrelationships of the items". Then a CFA was performed "using the validation sample data to *validate* the four factors the [test] was hypothesized to contain" (p. 323; emphasis added). In del Barrio, Aluja, and Spielberger (2004), "Both the eigenvalue one criterion..., and the Scree test...were used for factor

extraction. Additionally, a confirmatory factor analysis...was used...in order to *obtain* factor structures well adjusted to the data" (p. 231; emphasis added).

c. Specification of multidimensional structures

In the reviewed articles, some researchers were prepared to declare that their test had a "multidimensional structure" (i.e., that more than one attribute, or more than one facet of a higher order attribute was being measured by the test), but did not always make a clear statement as to exactly how many dimensions "underlie" the test. For some researchers this meant simply concluding that the test was "multidimensional", with no firm commitment to precisely how many attributes (or facets) were being measured. For example, after conducting a PCA, Miller, Joseph, and Tudway (2004) concluded that the "results support the evidence that impulsivity can be viewed as a *multi-dimensional* construct" (p. 355; emphasis added); in their study, Marsh, Parada, and Ayotte (2004) concluded that "the results...provided strong support for...the *multidimensional* perspective that is a particular strength of the SDQII" (p. 37; emphasis added).

d. Testing multiple models

Another approach used by researchers to evaluate the structure of a given test was to test multiple models (usually a number of competing restricted models), and then choose the model that had the "best" fit to the data. For instance, despite the fact that O'Connor, Colder, and Hawk (2004) claimed that the "goal of [their] study was to confirm a two-factor structure" (p. 987), they

tested not only a two-, but also three-, four-, and five-factor models; Grégoire (2004) compared two-, three-, and four-factor models, and judged that the "four-factor solution fitted the data much better than did the two- and three-factor solutions" (p. 463). In examining the psychometric properties of the Achievement Goal Questionnaire, Finney, Pieper, and Barron (2004) espoused that one of the features of CFA is that it "provides the opportunity to examine the extent to which alternative models might explain the interrelationships among the items" (p. 373); and DuHamel et al. (2004) alleged that one of the strengths of their study was that it was the first "to investigate multiple models for the symptom structure of PTSD as measure by the PCL" (p. 257).

e. Accommodating results

As aforementioned, some researchers employed "exploratory techniques"[42] with the aim of confirming that their test measured a certain number of attributes, but, in the face of contradictory results, reinterpreted the test as actually measuring something more or less than it was originally designed to measure. Another strategy by which researchers tried to make sense of unexpected results was to re-specify an initial model in which the fit to the data

[42] Here, by "exploratory technique" is meant a statistical model for which the association structure it implies is not restricted to a particular form, e.g., exploratory factor analysis (EFA) or principal component analysis (PCA). Note, however, that the current author rejects the notion that a *technique* (procedure, method, etc.) is by nature "exploratory" or "confirmatory"; rather, techniques may be employed with either an exploratory or confirmatory *aim* (or some combination thereof).

was poor, and then assess the fit of the new model to the test items. In their evaluation of the Beck Depression Inventory-II (BDI-II), Osman, Kopper, Gutierrez, and Bagge (2004) tested the fit of three different models using CFA and, when "None of the models...met all the pre-established initial and final adequacy-of-fit criteria" (p. 124), they employed an EFA "to explore alternate solutions", claiming that the exploratory analysis "suggested the extraction of a one-factor solution" (p. 125).

A number of researchers also re-specified certain aspects of a particular model in order to improve the fit of the model when it failed to be supported by data. For example, when two of the scales for the Perfectionism Inventory (PI) did not show acceptable fit to unidimensional confirmatory factor models, Hill et al. (2004) freed up one parameter for each scale, thereby improving fit indices. Beck et al. (2004), whose explicit aim was to examine whether the three-factor structure of a pre-existing measure would replicate in an independent sample, eliminated cross-loading items, resulting "in a model that approached an adequate fit to the data" (p. 292).

2. *Examining the "Structure" of a Test*

a. Why and when structure is examined

This category overlaps to a fair degree with category 1(b) above, which described the practice of exploring the data in order to *determine*, or *discover*, what is the structure of the test. In the articles examined in the present work,

there were many instances in which researchers treated the examination of the "underlying" structure of the test as an exploratory exercise, however, one which tended to be informed at least to some extent by theoretical considerations, i.e., by what the experts in the field believe in regards to which and how many attributes are being measured by the test. In such cases, researchers "explored" the structure of the test and then interpreted which attributes were being measured by which items: "Exploratory factor analysis...*revealed* the emergence of three factors" (Zweig and Webster, 2004, p. 239; emphasis added); "An exploratory factor analysis of the...items *produced* two factors" (Finley and Schwartz, 2004, p. 150; emphasis added); "the purpose is...to *determine* the factor structure of the AMAS-A" (Lowe and Reynolds, 2004, p. 663; emphasis added);"...we approached factor analysis with an open mind as to the number and nature of the factors" (Kohn, O'Brien-Wood, Pickering, and Decicco 2004, p. 114). From this, it would seem that some researchers reserved the examination of structure as an opportunity to *establish* what is being measured by the test rather than to confirm that the test is performing according to researchers' expectations for the populations under consideration.

b. Assessing the dimensionality of test items

Researchers approached the assessment of the "dimensionality" of test items in a number of different ways. First, from 1(b) above, the dimensionality of a test was sometimes considered something to be discovered via the employment

of certain exploratory procedures: Krueger et al. (2004) contended that "Unidimensionality [could] be *assessed* via exploratory factor analysis" (p. 110; emphasis added), and that "unidimensionality [was] *established* by demonstrating that a one-factor model provides a *parsimonious* fit to the data" (p. 113; emphasis added); Ullstadius, Carlstedt, and Gustafsson (2004) claimed that the "question" of dimensionality could be addressed at the item level. Some researchers tested the fit of particular restricted models, but, as was discussed in 1(c) above, hypothesized a number of competing models, and then the model that constituted the "best" fit (by whichever criteria the researchers employed) was seen in a sense to *determine* the dimensionality of the test. For example, after testing three competing initial models, and making modifications to the model with the best fit, Davis, Capobianco, and Kraus (2004) concluded that, based on "CFAs and internal reliability estimates", the items on the Conflict Dynamics Profile, a assessment tool they developed for measuring responses to conflict, "tap[ped] 15 unique constructs" (p. 728). Moneta and Yip (2004) claimed that "the *identification* of dimensionality [could] be conducted using confirmatory factor analysis, which allow[ed] the comparison of alternative models" (p. 540; emphasis added).

Second, dimensionality was sometimes treated as something that could be *imposed* on a test according to researchers' desires or needs, rather than something that could be deduced from the theoretical structure of the test (e.g.,

defined by the developer of the test, or implicit in the theoretical framework within which the test was constructed): "we calculated factor scores according to the hierarchical three-factor model...and a total score of the 13 items contained within the three-factor model. This 13-item score provides a more coherent or unidimensional estimate of the construct of psychopathy" (Cooke, Hart, and Michie, 2004, p. 336); "the primary purpose of the present investigation was to evaluate the proposal that relations between self-concept and different components of mental health can be *better understood* from a multidimensional perspective of self-concept than from a unidimensional perspective" (Marsh et al., 2004, p. 27; emphasis added); "the use of the three-factor or seven-factor *representations* appears to be a matter of *preference*" (Roesch, Rowley, and Vaughn, 2004; emphasis added); "not only tests *can be considered* multidimensional, but items as well" (Ullstadius et al., 2004, p. 1004; emphasis added); "Neuberg, Judice, and colleagues...recommended *using* the NFCS *as* a two-factor instrument" (Moneta and Yip, 2004, p. 531; emphasis added); "Some researchers have argued in *favor* of a three-factor solution" (Williams and Paulhus, 2004, p. 767; emphasis added).

Third, some researchers interpreted differing empirical findings with regard to the dimensionality of test scores drawn from different populations as an indication that they could use the test as a measure of say one particular attribute for one population, and a measure of more than one attribute for

another population. That is, researchers interpreted the fact that dimensionality is function of item content and population as an opportunity to "discover" what the test measures (for a particular population), as opposed to testing *whether* the test measures a particular attribute for a given population. Hence, a test once considered to be unidimensional could be reconceived as multidimensional for a given population and vice versa. Roesch et al. (2004) found that "Contrary to expectations, a principal components analysis...did not support the three-factor representation that was hypothesized", and, hence, commenced with attempting to establish the validity of the "now unidimensional measure" (p. 282). However, these researchers also compared a number of different confirmatory factor models in order to "explore the dimensionality of [the] *construct*" (p. 285; emphasis added) underlying the measure, and concluded that "The dimensionality of the SRGS was shown to be highly unstable" (p. 287).

Fourth, some researchers treated unidimensionality as an *ideal toward* which the test analyst strives, accompanied with varying degrees of satisfaction that a measure is unidimensional *enough.* In the literature reviewed in the present study there were references to "*relatively* unidimensional" constructs (Marsh et al., 2004; emphasis added), items which are "*reasonably* unidimensional" (Mungas, Reed, and Crane, 2004; emphasis added), items which "*primarily reflect* a single dimension" (Krueger et al., 2004; emphasis added). Wolfe, Ray, and Harris (2004) "performed dimensionality analyses to determine

the *degree* to which each instrument exhibits *sufficient* internal consistency to support an assumption of unidimensionality" (p. 847; emphasis added).

Finally, researchers used a variety of different "exploratory" criteria for deciding whether the scores from a given test did or did not conform to a particular structure. Many tested multiple models, each of which specified an, often different, *m*-dimensional structure; others employed criteria such as the number of eigenvalues greater than one, or scree plots, the ratio of first and second largest eigenvalues (cf. Bolt, Hare, Vitale, and Neumann, 2004), or proportion of variance accounted for by the first *m* factors/components, from either an EFA or PCA, as an indication of the number of attributes that were being "tapped", "represented", or "assessed" by a given measure. In addition, some researchers incorrectly employed certain test theory indices as measures of unidimensionality, such as inter-item correlations, inter-subscale correlations, or coefficient alpha: "Pairwise correlations among the three scales were high...In short, the measures were overlapping but not identical (Mantler, Shellenberg, and Page 2003, p. 146); "Thus, the mean inter-item correlation may give us a more useful index [of reliability] than alpha since it shows the *homogeneity* of items" (Alexopoulos and Kalaitzidis, 2004, p. 1211; emphasis added); "The average interscale correlation is 0.10, suggesting that these scales are relatively *orthogonal*..." (Leach and Lark, 2004, p. 150; emphasis added); Blackburn, Renwick, Donnelly, and Logan (2004) tested the significance of the bivariate

correlations among the scales in order to assess the unidimensionality of their measure; Blair et al. (2004) concluded that "High inter-rater reliability coefficients for total scores...and high Cronbach alpha coefficients and inter-item correlations provide confirmation that the PCL-R is a *homogeneous* scale tapping into a *unitary* construct" (p. 114; emphasis added); Bogels and van Melick (2004) claimed that the "*homogeneity* of the total score in the present population was high" (p. 1587) in reference to reported estimates of alpha; Finley and Schwarz (2004) warned that the apparent two-factor structure of their test, although consistent with the theory, could possibly be "artificial in light of the high correlations between...subscales and in light of the extremely high Cronbach's alpha value for the overall...score" (p. 155).

Another aspect of researchers' treatment of dimensionality within test analysis concerned how multidimensionality was handled. Some researchers treated a "multidimensional" measure in a vague, non-specific manner (e.g., without formally testing a particular hypothesis about the number of dimensions underlying responses to test items). Others set out to test that a specific multidimensional structure underlies the data from a given measure, but often without use of either specific multidimensional confirmatory (i.e., "restricted") models, or unidimensional models for individual subscales, each of which, presumably has been designed to measure but a single attribute, or facet of a higher order attribute.

The points just outlined speak to researchers' general practices with regard to assessing dimensionality (presumably as a indication of the number of attributes being measured by a given test) in a test analytic context. It is important to note, however, that many of the researchers whose work is reviewed here did not explicitly address the issue of dimensionality at all, or they merely gave it lip service. In the present work, of the 213 studies which assessed some property of the test(s) employed (i.e., empirical assessment of structure, precision, validity), 94 (44%) examined structure, but only 26 (28%) of these unequivocally addressed expectations regarding the dimensionality of the test, and only 4 (5%) employed models which were unidimensional in some sense. However, most of these studies (79 of 94, or 84%) analyzed the reliability of the tests that were being evaluated, and reported reliability estimates of either total or subscale composites of items.

c. Assessing fit

Generally those researchers assessing dimensionality via exploratory techniques, if they used any criteria for fit at all, typically tested an unrestricted *m*-factor model with a chi-square test in which the null hypothesis was of the form $H_0 : \Sigma = \Sigma_m$, for $m = 1, 2, \ldots$ against the alternative hypothesis $H_A : \Sigma =$ any positive definite matrix. However, the issue of fit in the context of exploratory techniques seemed not to be the primary focus for many of the researchers employing such procedures; rather, often the goal appeared to be to *determine*

how many and which attributes were being measured ("tapped", "represented", "picked up") by the test of interest, and not to test how well a particular *m*-factor model fit the data.

Many of the studies reviewed in the present work employed restricted models such as CFA, or IRT models, which have built into them formal tests of fit with respect to specific hypothesized latent structures. In such cases, these researchers typically used multiple criteria for assessing model fit, including some combination of some version of the chi-square test of fit along with goodness-of-fit estimates such as the Goodness-of-fit Index (GFI), the Comparative Fit Index (CFI), the Normed Fit Index (NFI), the Tucker-Lewis Index (TLI), the Root Mean Residual (RMR), or the Root-mean-square Error of Approximation (RMSEA). For the most part, the employment of criteria such as these was in keeping with the general standards set forth by APA (American Psychological Association, 1999). However, how individual researchers assessed *relative fit* when comparing "competing" models was not consistent from study to study; some researchers simply compared the observed chi-square values and/or the observed goodness-of-fit estimates (cf. for e.g., Cole, Rabin, Smith, and Kaufman, 2004; Wallace, 2004). Others employed chi-square difference tests in order to determine in a set of models which constituted the best fit to the data. Some researchers were interested in testing the difference between models which differed only in terms the number of factors (cf. for e.g., Bishop and Hertenstein,

2004; DuHamel et al, 2004); others still were concerned with whether particular *m*-factor models would constitute a better fit once constraints were placed on certain of the parameters in the model (cf. for e.g., Campbell-Sills et al., 2004).

d. Attribute, factors, and subscales

In some of the test analyses reviewed in the present work, the distinctions between "attribute" ("facet", "trait", etc.), "factor", and "subscale" were not always clear. In certain cases "factor" was used interchangeably with "subscale": Spence, Oades, and Caputi (2004) noted that "An exploratory factor analysis conducted by Petrides and Furnham (2000) revealed that the 33-items *load* onto four *sub-scales*" (p. 455; emphasis added); Rodebaugh et al. (2004) claimed to "investigate the possibility that reverse-worded *items* on the FNE *form* a distinct *factor*" (p. 170; emphasis added) and, because the two test theory models they employed both "assume unidimensionality", they "analyzed one *factor* at a time" (p. 171; emphasis added); Finley and Schwarz (2004) used factor-analytic methods in order "to identify the underlying latent components (*i.e. subscales*)" (p. 148; emphasis added); Marsh et al. (2004) claimed that a confirmatory factor analysis "demonstrated a well-defined multidimensional factor structure of *reliable*, highly differentiated self-concept *factors*" (p. 27; emphasis added); Knyazev, Slobodskaya, and Wilson (2004) claimed that a "Confirmatory factor analysis...showed that a four-factor model best fitted the data but [that] the three...*subscales* should be treated as *sub-factors* of a second-order factor" (p. 1565;

emphasis added); Beck et al. (2004) stated that "Excellent *internal consistency* was noted for each *factor* (α = 0.86-0.97)" (p. 289; emphasis added); in their test analysis, Roesch et al. (2004) concluded that "some of the *factors* of the 7-factor model had questionable *internal consistency*" (p. 281; emphasis added).

In other studies, "attribute" and "factor", and "attribute" and "subscale" were not well distinguished. Connor, Zhong, and Duberstein (2004) alleged that "The *domain* N is comprised of six lower-order *facets*..." which "are important to study because they have differing *reliability*" (p. 75; emphasis added); Roesch et al. (2004) claimed that the answer to the question of how many "dimensions" a particular test has "might be as simple as the number and type of *items* that each validation study used" (p. 287; emphasis added) and that "Growth was originally conceptualized as a multidimensional *construct* consisting of three *dimensions*...Items that compose these *dimensions* were generated based on the theoretical and empirical literature" (p. 282; emphasis added). As mentioned in 1(c), on the basis of results from examination of structure, a number a researchers interpreted a "good" fitting multidimensional model (i.e., an *m*-factor solution with $m > 1$) as indicating that the "construct" ("attribute", "trait") being measured by the test was "multidimensional" (cf. Marsh et al., 2004, Miller et al., 2004).

In addition, some researchers made statements pertaining to what factors "do" in comparison to what a test (or subtest) does: "The scree plot *called for* a three factor solution...the first factor *tapped* 'plausibility of the defendant's

claim'...the second factor...*assessed* the extent to which the defendant intended to kill her husband...The final factor...*tapped* the defendant's general psychological instability" (Schuller, Wells, Rzepa, and Klippenstine, 2004, p. 131; emphasis added).

3. *Practices pertaining to compositing test items*

Of the 251 articles that were subject to thorough review, 189 (75%) included reliability analyses of some kind. Of these, approximately 25% were explicit with respect to the nature of the composite for which reliability was estimated, e.g., the unweighted sum of the items for the test/subscale, the unweighted mean of the items for test/subscale, etc. For those studies in which the nature of the composite was explicit, the majority employed as a compositing rule the unweighted sum of the items. Typically there was no justification or rationale given for choosing a particular compositing rule over others, and it was not always clear whether the employed compositing rule was applied at the level of the subtest, for the total set of items, or both.

4. *Practices pertaining to analysis of reliability*

In the studies reviewed here, there was a fair degree of variety in terms of how researchers handled the issue of precision estimation. First, many of the studies surveyed (119 or 47%) reported reliability estimates from previous studies (and/or test manuals) instead of calculating estimates from the data on

the measures of interest that were generated in those specific studies. In some of the studies in which more than one measure was used, reliability estimates were computed for some, but not all, of the measures, generally with no rationale offered for doing so (cf. Barrett et al., 2004; Gagne, Lyndon, and Bartz, 2004; Grano, Virtanen, Vahtera, Elovainio, and Kivimaki, 2004). Other studies did not address the reliability of the measures employed at all (cf. Nagtegaal and Rossin, 2004; Ullstadius et al., 2004). A number of studies examined the structure of the items of the test(s) of interest, but did not report on reliability, even if the structure was deemed to be in keeping with expectations (cf. Hewitt, Foxcroft, and MacDonald, 2004; Hill, Neumann, and Rogers, 2004; Karademas and Kalantzi-Azizi, 2004; Maller and French, 2004).

Second, some researchers did not specify the nature of the reliability indices they employed (cf. Miller and Bichsel, 2004; Piper, Ogrodniczuk, and Joyce, 2004) and virtually none gave justification for employing the particular indices they did. In the studies reviewed here, many (64 %) employed at least one of the "classical" coefficients of reliability, with coefficient alpha being the most frequently employed estimate in these cases (94%), followed by "test-retest" estimates (23%). Furthermore, with regard to the latter, only 53% of the researchers who used a test-retest coefficient did so with the explicit aim of assessing the *stability* of the measure.

Third, there was no consistent method for reporting reliability estimates. Some researchers reported reliability coefficients only for composites formed from the entire set of items comprising a test, even for tests composed of mutually disjoint subsets (or "subscales) of items (cf. Muris, deJong, and Engelen, 2004, Williams and Paulhus, 2004); others reported the range of observed reliability estimates for a set of subscale composites, rather than explicitly reporting the estimated reliability for the composites formed from the items in *each* subscale (cf. Blumentritt, and van Voorhis, 2004); some researchers reported some aggregate of the reliabilities of subscale composites, such as the mean of alpha coefficients across subscales (cf. Sears and Rowe, 2003), or coefficient alpha for the mean of subscale items (cf. Sirois, 2004). In addition, what was considered the threshold for "acceptable" observed reliability varied widely across the surveyed studies: Values for observed reliability estimates that were deemed acceptable by researchers ranged from .46 (cf. Egan, Kroll, Carey, Johnson, and Erickson, 2004) to .98. In fact, some researchers interpreted low observed reliability as being acceptable for reasons such as consistency with previous studies (cf. Egan et al., 2004; Zhang, 2004), the fact that the measure of interest had few items (cf. Francis and Jackson, 2004), or that low reliability was offset by good validity (cf. Vittengl, Clark, and Jarrett, 2004).

Fourth, there existed amongst researchers different conceptions of what interpretations could be given to various estimates of reliability. For example, as

was addressed in 2(b), coefficient alpha (as well as other "internal consistency" coefficients) was used by some as a measure of unidimensionality. In addition, researchers did not always distinguish between different estimates of reliability: Goulding (2004) reports that "Psychometric evaluation of the O-LIFE has shown it to have good *test-retest* reliability (coefficient *alpha*=0.80)" (p. 161; emphasis added). Some also failed to distinguish between the fit of a model to the data and estimates of precision: Finley and Schwarz make reference to the "*unreliability* of the two-factor solution" (2004, p. 151; emphasis added). In addition, researchers held varying conceptions of the relationship between reliability and validity: "The low level of test-retest reliability raises concerns about the validity of the CSA. It is probable that the test items do not measure the W-A dimension or the V-I dimension with sufficient precision" (Parkinson, Mullally, and Redmond, 2004, p. 1277; Oliver and Simons (2004) claimed that "While the ALS has been shown to have good internal reliability and discriminant validity..., there is limited empirical support for its hypothesized factor structure" (p. 1280); Lowe and Reynolds (2004) state that "Besides internal consistency..., *another index of construct validity* of a measure's scores is the pattern of moderate correlations exhibited between subscale scores" (p. 675; emphasis added).

Finally, the most striking feature of researchers' practices with regard to analyzing precision had to do with *where* estimation of precision occurred in a

test analysis. In the studies that examined the structure of the test items, 58% gave some estimate of precision *prior to* evaluating the structure of the test. In sum, most researchers either did not evaluate structure at all before calculating reliability estimates, or did so only after estimating reliability.

5. *Practices pertaining to analysis of validity*

The issue of validity was addressed in one way or another in 38% of the articles reviewed; some sort of quantitative assessment of validity (including correlations between measures, correlations between factor scores, multiple regression coefficients, exploratory factor analyses, and formal tests of confirmatory factor models) was attempted in 74% of these studies.

Researchers varied in terms of the approaches they took to validity assessment. Some researchers estimated some combination of "convergent", "concurrent", "divergent", "discriminant", and "criterion-related" validity (cf. Fiorentino and Howe, 2004; Jay and John, 2004). Other researchers evaluated the structure of some measure(s) of interest in order to investigate relationships among attributes, using either exploratory (cf. Leach and Lark, 2004; Williams and Paulhus, 2004; Zhang, 2004) or confirmatory (cf. Currie, el-Guebaly, Coulson, Hodgins, and Mansley, 2004) factor analysis. Some researchers examined structure as well as estimated relatively straightforward indices of validity such as convergent and discriminant coefficients (cf. Lowe and Reynolds, 2004; van der Ploeg, Mooren, Kleber, van der Velden, and Brom, 2004).

There were also researchers that reported validity estimates, but did not describe the nature of these estimates (cf. Jeyakumar, Warriner, Raval, and Ahmad, 2004).

In some studies, a particular aspect of validity was assessed, that of "factorial invariance", an expression used to describe when test scores drawn from different populations may be said to have the same factor structure, i.e., the test data from samples taken from all populations under consideration can be said to fit a particular statistical model. For example, Keogh (2004) investigated whether the factorial structure of the Anxiety Sensitivity Index was invariant across gender; in their 2004 study, Taub, MeGrew, and Witta sought to establish the factorial invariance of the WAIS-III across different age groups; Utsey, Brown, and Bolden (2004) tested "structural invariance" of the Africultural Coping Systems Inventory across three independent and ethnically distinct samples.

6. The logic of current test analytic practices

One of the primary aims of the present study was to examine current practices of test evaluation. Of the studies in which some aspect of the test(s) employed was being evaluated (i.e., involving either assessment of structure, estimation of precision, and/or validity), approximately 37% involved both the assessment of the fit of one or more statistical models and estimation of the precision of at least one function of the test items; in 70% of these studies, precision estimates were given prior to testing model fit.

Some aspect of the validity of the test was examined in 38% of the studies, sometimes prior to and sometimes subsequent to the assessment of precision. Approximately 39% employed, at least to some extent, a logic for analyzing the test, i.e., some rationale for judging the quality of the test, or tests, employed, 37% contained some, either explicit or implicit, statement regarding expectations as to what the test measures, most typically, with respect to the number of attributes which the test was presumed to measure. In this group of studies, 80% involved both assessment of model fit and estimation of precision; in 62% of these, precision estimates were given prior to assessment of model fit. In addition, if there was explicit mention of the formation of a composite of the items, rarely was an indication given as to the nature of the compositing rule employed in producing the test score.

5. A PROPOSED LOGIC FOR TEST ANALYSIS

In the introductory chapter it was noted that although much has been made of the advances and sophistication brought to the test analytic game by MTT, a key distinction must be made between *test theory* and *test analytic frameworks*. As opposed to being competitors with regard to *how* test analyses are done, CTT and MTT constitute quantitative "pictures" of the relationship between the responding of respondents to test items and the attribute (trait, ability, property, etc.) for which the items are designed to be indicators. Each of the CTT and MTT orientations has spawned quantitative tools that a test analyst can enlist to assess the performance of a test. However, neither constitutes a framework that specifies *how* these tools should be used to pass judgment on the quality of a given test. A test is a test *of something* and, as such, can only be judged as being "good" or "bad" in reference to how good or bad it is *as a test* of that thing. Hence, test analysis is an *evaluative practice* by virtue of the fact that it is the job of the test analyst to pronounce on the worth of the test, and he can only do so by reference to criteria, or rules, that fix what it means for a *test to behave as it should behave*, or *as it was designed to behave*. However, clearly one cannot look to *test theory* for an elucidation of such rules, because test theory is

merely a set of mathematical tools, as opposed to a set of rules that fix *how* these tools are to be used.

If the aim of test analysis is to make decisions about whether the performances of tests are "satisfactory", "adequate", "passable", "dreadful", "abysmal", etc., then the aim is to make evaluative judgments. However, evaluative judgments will only have meanings if founded on clear senses of the evaluative terms ("satisfactory", "unsatisfactory", etc.) of which they are comprised. The senses of evaluative terms in general, and the evaluative notions particular to test analytic practice such as *adequate test performance, the items measured what they were designed to measure, a poor test of attribute* γ, etc., are, as with all terms, fixed by rules. Thus, the brand of test analysis that has the power to pronounce upon test performance, i.e., true *evaluative* test analysis, is, by its nature, part of a *rule-guided practice*. This brand of test analysis is founded on clear, antecedently specified *standards* for all of the key components employed in passing judgment on a test. It is not sufficient for the test analyst to have a "hunch", or an "intuition", about how the test items measure the attribute of interest; rather, he must be able to state the properties that must be possessed by the joint distribution of the test items in order that the test be judged as performing *as it should perform*, or *as it was designed to perform*. This makes evaluative test analysis a brand of confirmatory analysis, for to make such

judgments about a test is to compare the *actual behaviour* of a test to an antecedently specified *standard of correctness.*

There is, of course, the possibility of engaging in purely exploratory analyses into the statistical properties of a set of test items, but, in the absence of antecedently stated rules that fix the senses of evaluative terms, this brand of analysis lacks the force to justify pronouncements about the quality of tests, and, as such, is *not* test analysis proper. It will later be shown that a fatal, but common, flaw of current test analytic practice has been the failure to distinguish between exploratory and evaluative/confirmatory aims, this resulting, time and again, in test analyses that feature the unworkable pairing of the desire to make evaluative claims and the absence of antecedently specified standards of adequate test performance.

A test analytic *framework* is, then, a set of rules that fixes what it *means* for a test to perform "adequately" (and "inadequately"), and the steps that the test analyst must take in order to *justifiably* pronounce upon the quality of a given test. A test analytic framework is a *prescription* for *how* the mathematical products (or other methodological accoutrements) of test theory should be used to pass judgment on a test. Clearly, neither CTT nor MTT are frameworks in this sense. In fact, despite the impressive developments that have characterized theoretical test *theory*, the practice of test *analysis* has apparently received little attention. Hence, it remains unstructured, unsystematic, and piecemeal. It is,

then, precisely the *absence* of a true test analytic framework that has plagued test analytic practice within the behavioural and social sciences with the frequent production of confused and contradictory results.

The absence of a clearly stipulated logical framework for test analysis is both undesirable and unnecessary, for a careful consideration of the literature (in particular the work of Cronbach and Meehl, 1955, Embretson, 1983, Loevinger, 1967, Lord, 1952, 1953, Peak, 1953, and Thissen et al., 1983) reveals all of the necessary ingredients for creating such a framework. In brief, the framework elucidated herein consists in the following sequentially structured components:[43] 1) specification of the *theoretical structure* (TS) of the test to be analyzed; 2) the translation of the TS into a set of quantitative requirements for the joint distribution of the test items, the resulting translation called the *quantitative characterization* (QC)—this component will standardly consist in the choice of a (unidimensional) test theory model that squares, or is in keeping, with the TS; 3) a test of conformity of the joint distribution of the items of the test to the chosen test theory model; 4) conditional on the test items having been judged as conforming to the QC, the derivation of an optimal, model-implied, rule for the compositing of the test items—this step is equivalent to choosing a function that maps a respondent's scored responses to the test items into a number that can

[43] Elements of this framework, in particular with regard to steps 1 and 2, were initially sketched out in Maraun, Jackson, Luccock, Belfer, and Chrisjohn (1998).

justifiably be seen as an estimate of her value on the attribute for which the items are indicators; 5) the estimation of the reliability (or, more generally, the precision) of the resulting composite of test items; 6) conditional on the composite having been shown to possess adequate reliability, the entering of the composite into "external" construct validation studies (e.g., multi-trait, multi-method analyses, general explorations of the composite's place in the broader nomological network), the aim being to assess whether the scores on this composite can rightly be called (error-laden) measurements of the attribute for which the items were designed to be indicators.

Steps one to five constitute *internal* facets of test analysis, and step six, the *external* facet. A test, then, may rightly, but provisionally, be judged to be performing adequately in a focal population **P** of respondents if it 1) conforms to its TS and the reliability of the resulting composite is adequate, and 2) has behaved, to date, "as it should behave" in external construct validation studies. Clearly, the internal facets of test analysis are a mix of what were classically known as reliability and validity concerns, while step six is what was classically seen as construct validation proper. More will be said about these distinctions shortly.

A Proposed Framework for Test Analysis

One of the primary concerns of the current work is to explicate a logically sound framework for the conducting of test analyses. To render coherent evaluative decisions about the performance of a test, a test analytic framework must involve A) a clear articulation of what it *means* for a test to perform *satisfactorily*; B) a clear articulation of what it *means* to say that a test has been "appropriately" analyzed; C) criteria for identifying the quantitative features of a set of test items that are *relevant* to a judgment of a test's performance; D) a specification of the conditions under which a set of test items can justifiably be composited; E) a specification of the conditions under which the reliability of a test is defined and can be coherently estimated; F) a specification of the conditions under which the validity of a test can justifiably be investigated. This set of requirements motivates the logic proposed.

Once again, the components of the framework, presented in the order in which they should be addressed, are as follows:

1. The TS of the test is specified.
2. A QC that is in keeping with the TS is chosen.
3. The conformity of the joint distribution of the test items (in a focal population **P**) to the QC is assessed.

4. Conditional on the conformity of the joint distribution of the items to the QC, an optimal, model-implied compositing rule is derived and employed in order to scale the respondents in **P**.
5. The reliability (or, more generally, the precision) of the resulting composite is estimated.
6. Conditional on the composite possessing an adequate degree of precision, the composite is entered into "external" construct validation studies, the aim being to assess whether the scores on this composite can rightly be called (error-laden) measurements of the attribute for which the items were designed to be indicators.

It must be emphasized that these features of the proposed framework do not constitute merely a list, but, rather, a list of *sequentially ordered* steps—whether or not one moves on to a given step is *contingent* upon what occurs at an immediately preceding step. The logic underlying this particular ordering is based on the observation that a test (i.e., a set of items designed to measure some particular attribute) cannot justifiably be composited unless it has been shown to conform with its theoretical structure, its degree of precision may not be estimated unless it has been shown to be compositable, and it cannot be "validated" unless there can legitimately be created a composite, and this composite possesses an adequate degree of precision in focal population **P**. However, the sequential nature of test analysis does not preclude the making of

certain provisional claims, such as, for example, that the test has been shown to conform to a chosen test theory model (and, hence, to its TS), or that the test, having been shown to conform to an appropriate test theory model, is compositable, the resulting composite having been shown to possess adequate reliability. In fact, because there is no limit to the number of external validity-related analyses that can be conducted, claims about a test's (adequate) performance are inherently provisional. Each component of the framework will now be elaborated in turn.

1. Specification of the Theoretical Structure of the Test

If a test is to be meaningfully judged as performing "adequately", or "inadequately", or "passably", in a focal population **P**, then senses must be assigned antecedently to these, and like, evaluative terms. The specification of the TS is the first step in the fixing of the senses of these terms. The TS provides, on the linguistic plane, one component of a standard of correctness for the notion of "adequate test performance". A TS must be worked out or deduced for each test individually, because what constitutes adequate performance for a test comprised of dichotomous items that were designed to measure anxiety, for example, may well be different from that for a set of dominance items with 10-point Likert response scales. Being a linguistic specification, the TS, despite being the foundation of any true, evaluative, test analysis, must be converted

into quantitative terms before it can have implications for the features of the joint distribution of the test items.

A TS can be defined as a loose linguistic specification of how the items of a test, T, were designed to measure a given attribute, γ, of interest, including how the items are linked to γ, and whether they are viewed as "fallible" indicators of γ. A clear specification of the TS of T, whose performance is to be analyzed is the starting point for any true, evaluative test analysis. Analyses for which this specification is absent are not founded on unambiguous senses of "the test is behaving as it should behave" and similar notions, and, hence, yield evaluative claims that are at best ambiguous, and, at worst, vacuous.

Although the components that should properly comprise the TS of a test are open to debate, it would seem that, minimally, the TS of any test is representable as a four-tuple, TS(I,D,R,E). Element I stands for item type, D, the number of attributes that the items were designed to measure, R, the (theoretical) form of the regressions of the items on the attribute that they were designed to measure, and E, the error characteristics of the items. As regards item type, commonly occurring response scale formats are continuous (C), x-point Likert (xPL), and dichotomous (DI). Although, in theory, D can take on any positive

integer, in practice, its value will typically be unity.[44, 45] The sense of "regression" as concerns R is non-mathematical (or, more accurately, *pre*-mathematical), and refers to how the items are conceptualized as varying with the level of the attribute. These regressions are pre-mathematical because the attribute is not a variate, but, rather, the unobservable property that the items were designed to measure. Frequently encountered values of R are *monotone increasing* (MI), *linear increasing* (LI), *S-shaped* (S), and inverted *U-shaped* (U) item/γ regressions. Finally, for the sake of generality, E will be allowed to assume two values, error-free (EF) or error-in-variables (EIV), even though modern test analytic practice essentially always assumes that the items that comprise a test are error-laden indicators of the attribute that they were designed to measure.

The following are examples of theoretical structures commonly encountered in psychological research:

i. **TS(C,1,MI,EIV):** Tests with this theoretical structure are comprised of k continuous (i.e., I=C) items which were designed to be indicators of a single attribute, γ, of interest (i.e., D=1). Assuming that the items have been recoded so that they are keyed in the same direction, if the items are

[44] As, even those tests comprised of many items designed to measure many attributes virtually always can be decomposed into mutually disjoint subsets of items, each subset designed to measure but a single attribute. Each subset, then, may be considered to be a "test".

[45] This premise that shall be assumed throughout the remainder of the current chapter unless otherwise specified.

functioning according to expectations, the conceptualization of the item/γ regressions is as follows: as the amount of γ possessed increases, the values of the response options endorsed also increase (i.e., R=MI). Finally, the items are conceptualized as fallible indicators of γ (i.e., E=EIV).

ii. **TS(C,1,LI,EIV):** A test with this theoretical structure is comprised of *k* continuous items which were designed to be indicators of a single attribute γ. Once the items have been recoded such that they "point in the same direction", the theoretical regressions are conceptualized to be linear increasing (a sub-class of the larger class of monotone increasing functions). This TS also specifies that the items are to be conceived as "imperfect" indicators of γ.

iii. **TS(DI,1,S,EIV):** A test having this theoretical structure consists in a set of *k* dichotomous [0/1] items which were designed to be indicators of a single attribute, γ. The item/γ regressions are seen as monotone increasing, but the dichotomous response option format of these items necessitates that these regression have both an upper and lower asymptote. These regressions are, then, taken to be S-shaped. The items are, as is usual, considered to be error-laden indicators of γ.

iv. **TS(C,1,U,EF):** As above, the *k* continuous items of the test are thought to jointly measure a single attribute, γ. However, each item is conceived as having an inverted U-shaped regression on γ: For a particular item,

responses increase with increases in γ up to a certain point, beyond which, responses decrease as the level of γ increases. Here, each of the items is conceived as having a deterministic relationship to γ, i.e., each is thought to measure the attribute without error.[46]

When the members of focal population **P** respond to the items that comprise a test, T, whose items are designed to measure an attribute γ, and the items are scored, a random vector $\underline{X}$ is induced. The j^{th} element of $\underline{X}$ is the random variate X_j whose distribution contains the scored responses to item j, of the objects contained in **P**. The distribution of $\underline{X}$ in **P**, and, in particular, the association structure of the elements contained in $\underline{X}$, is fully determined by $f_{\underline{X}}$. The aim of a test analysis is to pass judgment on the performance of test T, in population **P**, on the basis of features of $f_{\underline{X}}$. Obviously, because the TS is a linguistic specification of how the items were designed to measure γ, the TS does not imply any specific empirical requirements of $f_{\underline{X}}$, the fulfilment of which would justify claims about TS/T conformity. Testable requirements for $f_{\underline{X}}$ (for the empirical "structure" of the items) must be generated through the translation of the components of a TS into quantitative counterparts.

[46] This theoretical structure is unlikely to be encountered (at least with regard to measures employed within psychology); however, a TS such as this is, at the very least, conceptually possible.

2. Derivation of the QC

The generation of empirical test analytic requirements for $f_{\underline{X}}$ comes about by developing a quantitative characterization, or translation, of the components of a given TS. A quantitative characterization (QC) of a given TS consists in a set of empirical requirements for $f_{\underline{X}}$ that is also consistent with the TS. The QC is, in other words, the quantitative embodiment of the TS, and specifies the properties that must be possessed by $f_{\underline{X}}$ in order that test T be correctly judged as conforming to its TS. There is, at least in theory, the possibility of constructing many sound (and many unsound) quantitative characterizations of a given TS. Proper evaluative test analyses can only begin with a choice of a QC that is isomorphic to the TS of the test to be analyzed. Isomorphism of TS and QC is essential, because, in the absence of such an isomorphism, the failure of $f_{\underline{X}}$ to satisfy the requirements implied by a QC cannot be taken as evidence that the test does not square with its TS and, hence, cannot be taken as legitimate grounds for indicting the test's performance in population **P**.

As was shown in Chapter Four, it is commonplace for analysts doing "test analyses" to "fit models" to test data. It is, however, clear from the fact that the meaning of "the test performs adequately in population **P**" is fixed in part by an isomorphic pairing of TS and QC, that this rampant model fitting is misguided (a case that will be taken up in the following chapter). The requirement that QC be isomorphic to TS *defines* what it means for a model-based result to be *relevant* to

passing judgment on a test. There exist countless test theoretic models, and, undoubtedly, more yet to be invented. It is, therefore, a virtual certainty that there will exist at least one such model that describes, at least reasonably well, $f_{\underline{X}}$. Roskam and Ellis (1992), for example, prove that for any $f_{\underline{X}}$ for which Σ contains only positive elements, there exists a unidimensional, monotone, latent variable model that describes $f_{\underline{X}}$. Mere conformity, then, of $f_{\underline{X}}$ to *some* model can have no necessary implications for the judgment of the performance of T. To put this differently, if "adequate test performance" were to be equated with merely finding a test theory model that happened to describe $f_{\underline{X}}$, then tests would, as a matter of course, be judged as performing adequately, and, conversely, there would exist no grounds for indicting a test.[47] What allows for the justifiable indictment of a test's performance is the lack of conformity of $f_{\underline{X}}$ to a QC that is *isomorphic* to the TS of the test. Thus, a model-based result is *relevant* to the passing of judgment on a test's performance just in the case in which the model in question is isomorphic to the test's TS (i.e., is an isomorphic QC).

The derivation of a TS/QC isomorphism is, moreover, the missing link of construct validation theory, a component whose importance was never made clear by Cronbach and Meehl (1955), and whose general mishandling by applied test analysts has time and again trivialized applied test analytic output. The TS

[47] It will be suggested in the following chapter that the endemic failure of test analysts to found their analyses on an isomorphic TS/QC pairing has resulted in just such a trivial state of affairs.

is recognizably the "internal" component of theory in a construct validation program centring on a test T designed to measure an unobservable attribute, γ. It is the theory that describes the relationship between unobservable γ and the observable indicators that are the test items. The QC is then, obviously, the deduced, testable empirical consequences of TS. Applied test analysts frequently bypass this first, and essential, stage of the construct validation program, compute a default composite whose existence they fail to justify, and enter it directly into the contingent, "external" facet of construct validation. Such violations of sequence (to be discussed in the following chapter) produce equivocal and vacuous conclusions.

To construct a quantitative characterization of a given TS, one maps the components of the TS into quantitative counterparts. Because all tests of practical interest describe the situation in which a set of k items were designed to be indicators of a single (unobservable) attribute, i.e., TS(.,1,.,.), all QC's of interest are founded on particular conceptualizations of unidimensionality.[48] Thus, in practice, the task of paraphrasing a particular TS(.,1,.,.) virtually always reduces to the task of choosing an appropriate existing *unidimensional* test theory model. When such models are employed as QC's, the correspondence relations are as follows:

[48] Because, as aforementioned, all subtests may be treated as "tests", i.e., as sets of items which are presumed to measure a single attribute.

i. The attribute, γ, for which the test items were designed to be indicators is represented by a synthetic random variate defined on focal population **P**. When latent variable models are employed as QC's, then this random variate is a latent (unobservable) variate, θ; when component models are employed, it is some composite of the items X_j, $j = 1, 2,..., k$. Thus, random variates of various sorts stand in as proxies for unobservable attributes.

ii. The notion that the k test items measure but one thing in common, the attribute, γ, is paraphrased as the claim that the k test items are unidimensional in a sense that is dependent upon the other components of the TS, i.e., the item/construct regressions and the error characteristics of the items. When latent variable models are used as QC's, the unidimensionality principles in play are those of strong or weak local independence, although a fully specified model is required before these principles have testable implications for $f_{\underline{X}}$. In the strong local independence case, the items are claimed to be statistically independent conditional on the latent variate, θ, while in the case of weak local independence, they are said to be conditionally uncorrelated.

iii. The item/attribute regressions referred to in the TS are paraphrased as analogous item/synthetic variate regressions. When QC's happen to be

latent variable models, the latter regressions are called item characteristic curves.

iv. Latent variable models, and their close cousin the classical true score model, were invented to model the situation of fallible indicators. Thus, it is natural to paraphrase TS's with the EIV component as latent variable models. The EIV component itself is represented within these models as the random vector $\underline{X}$ having a non-point distribution conditional on θ. That is, the variance of each item, conditional on θ, is non-zero. This variance is usually referred to as the "error variance".

Although, in theory, there can be created any number of different QC's for a given TS, to date, the test analyst's options are exhausted by the standard latent variable and component models. Regardless, the onus is on the test analyst to choose as a QC the model that is the best available match to the TS.

Examples of Some Quantitative Characterizations

1. A QC for TS(C,1,MI,EIV): Unidimensional monotone latent variable models

TS		QC
γ	→	Random, unobservable, latent variate θ
D=1: The items measure one attribute, γ	→	$f(\underline{X} \mid \theta) = \prod_{j=1}^{k} f(X_j \mid \theta)$
R=MI: Each item has a monotone (increasing) regression on γ	→	$E(\underline{X} \mid \theta) = \underline{g}(\theta)$, in which $\underline{g}$ is a vector of functions, and $\frac{d}{d\theta}\underline{g}(\theta) > \underline{0}$

E=EIV: Each item is a fallible indicator of γ	$\rightarrow$	$C(\underline{X} \mid \theta) = \Psi$ is diagonal and positive definite (the X_j are required to have positive "error variances")

This class of QC's is called by Holland and Rosenbaum (1986) the unidimensional monotone latent variable (UMLV) models. The first component of this correspondence relation is the equating of the attribute, γ, with a latent, random variate, θ. The second component speaks to the essential paraphrase of the assertion in the TS that the items jointly measure but a single attribute. This linguistic conception is mapped into the mathematical conception of unidimensionality, which here is defined as the conditional independence of the X_j given a random variate, θ, the latter of which is taken to be a proxy for γ. The third component links the linguistic conception of monotone increasing item/γ regressions to the mathematical requirement of monotone increasing X_j / θ regressions: For all j, the derivative of $E(X_j \mid \theta)$ must be positive. Finally, the fallibility of the items as indicators of γ is modelled as non-zero conditional variances of the X_j, given θ.

An analysis of whether a given test, T, with TS(C,1,MI,EIV) conforms to its TS, is then an analysis of whether $f_{\underline{X}}$ satisfies the requirements implied by this QC. In other words, it involves an assessment of whether there exists a UMLV model that describes $f_{\underline{X}}$. Holland and Rosenbaum (1986) have documented properties that can be used to check whether such a UMLV model exists. For

example, any set of X_j whose $f_{\underline{X}}$ conforms to a UMLV model has the following property: For any two disjoint subsets, $\underline{Y}$ and $\underline{Z}$, of the random variates contained in $\underline{X}$, the covariance of any pair of non-decreasing functions of $\underline{Y}$, conditional on any function of $\underline{Z}$, is non-negative, i.e.,

(5.1) $$C(d(\underline{Y}), g(\underline{Y}) \mid h(\underline{Z})) \geq 0, \ \forall\, d, g \text{ non-decreasing}, \ \underline{X}' = (\underline{Z}', \underline{Y}')$$

(cf. Holland and Rosenbaum, 1986). An $f_{\underline{X}}$ that has this property is called *conditionally associated* (CA) (Holland and Rosenbaum, 1986).

2. A QC for TS(C,1,LI,EIV): Unidimensional, linear common factor models

TS		QC
γ	→	Random, unobservable, latent variate θ
D=1: The items measure in common but one attribute, γ	→	$C(\underline{X} \mid \theta) = \Psi$, a $k \times k$ diagonal matrix
R=LI: Each item has a linear (increasing) regression on γ	→	$E(\underline{X} \mid \theta) = \underline{\Lambda}\theta$, with the elements of $\underline{\Lambda}$ having the same sign
E=EIV: Each item is a fallible indicator of γ	→	Ψ is diagonal and positive definite (the X_j are allowed to have positive "error variances")

This class of QC's is recognizable as the class of unidimensional, linear common factor (ULCF) models. The correspondence relation takes, once again, γ to be represented by a latent variate, θ, in this case a "common factor". The linear factor analytic paraphrase of the notion that "the items measure but one

attribute in common" is the uncorrelatedness of the X_j, conditional on θ (i.e., Ψ is diagonal). The linear item/γ regressions specified in the TS are paraphrased as linear X_j / θ regressions in the QC. That is, for all j, the mean of X_j conditional on θ is a linear function of θ. The fallibility of the items as indicators of γ is modelled according to the factor analytic paraphrase as Ψ being positive definite, i.e., the "unique" (or error) variances are non-zero.

Jointly, the components of this QC imply that Σ, the $k \times k$ population covariance matrix of the X_j, has the representation: $\Sigma = \underline{\Lambda}\underline{\Lambda}' + \Psi$, in which Ψ is diagonal and positive definite. This consequence is then a requirement that must be satisfied by $f_{\underline{X}}$ in order that a test with TS(C,1,LI,EIV) be judged as conforming to its theoretical structure.

3. A QC of TS(DI,1,S,EIV): 2-parameter item response models

TS		QC
γ	→	Random, unobservable, latent variate θ
D=1: The items measure one attribute, γ	→	$P(\underline{X} = \underline{x} \mid \theta) = \prod_{j=1}^{k} P(X_j = x_j \mid \theta)$
R=MI: Each item has an S-shaped, monotone increasing regression on γ	→	$E(X_j \mid \theta) = P(X_j = 1 \mid \theta) = \frac{e^{a_j(\theta - b_j)}}{1 + e^{a_j(\theta - b_j)}}$ or $= \Phi(a_j(\theta - b_j))$
E=EIV: Each item is a fallible indicator of γ	→	$V(X_j \mid \theta) = P(X_j = 1 \mid \theta)[1 - P(X_j = 1 \mid \theta)]$

This class of QC's is recognizable as the class of 2-parameter item response models. Unidimensionality is, once again, the quantitative translation of the notion that the items measure in common but a single attribute. Here it is defined in terms of the strong local independence of the X_j. That is, conditional on the latent variate, θ, the joint distribution of the X_j is product Bernoulli. The item/γ S-shaped regressions are quantitatively characterized by one of two choices: normal ogive or logistic X_j / θ regressions. The components of this QC jointly place restrictions on $f_{\underline{X}}$: If a test with the theoretical structure TS(DI,1,S,EIV) is to be judged as in keeping with its TS, it must be possible to express the $P(\underline{X} = \underline{x})$ as

$$(5.2) \qquad P(\underline{X} = \underline{x}) = \int_{-\infty}^{\infty} \prod_{j=1}^{k} \left(\frac{e^{(a_j(\theta - b_j))}}{1 + e^{(a_j(\theta - b_j))}} \right)^{x_j} \left(\frac{1}{1 + e^{(a_j(\theta - b_j))}} \right)^{1 - x_j} f(\theta) d\theta ,$$

for some choice of values of the a_j and the b_j, $j = 1, 2, ..., k$, and density function of θ, $f(\theta)$. If the items can be so described, then the test may be judged as conforming to its TS.

The mismatch of TS and QC, very often a consequence of the failure to even consider the TS's of tests, and thereby bypass the first stage in a construct validation analysis, leads to the production of irrelevant results. For example, linear factor analytic results are relevant to the adjudication of a test's performance if the test's theoretical structure is TS(C,1,LI,EIV). Otherwise, they

are irrelevant. For example, a linear factor analysis could have no bearing on judgments as to the adequacy of the performance of a test with TS(C,1,U,EIV). In the words of van Schuur and Kiers, "Factor analysis is an inappropriate translation of the analyst's assumptions about the structure of a data set that conforms to the unidimensional unfolding model" (1994, p.99). In fact, a reasonable paraphrase of TS(C,1,U,EIV) is the unidimensional, quadratic factor model (i.e., the metric, unidimensional unfolding model), and this QC produces exactly the same covariance structure as the two-dimensional, linear factor model (McDonald, 1967). The point is that there is nothing intrinsic to a linear factor analytic result, nor any other result, that makes it relevant to the aims of a given test analysis.

Now, one might believe that these comments indicate a misplaced, or perhaps overstated, critique of general test analytic practice. However, it requires little diligence to unearth real-world scenarios to which they apply, and for which they have very tangible consequences. Consider, for example, the case of the Self-Monitoring Scale (SMS; Snyder, 1974). For a period of time, this test was frequently "factored" (cf. Briggs & Cheek, 1988; Briggs, Cheek & Buss, 1980; Hoyle & Lennox, 1991; Tobey & Tunnell, 1981), and these linear factor analytic results eventually resulted in the test's indictment. The twenty-five items of which the SMS is comprised were designed to measure but one thing, the tendency to self-monitor. The linear factor analytic results, on the other hand,

suggested multidimensionality. Analysts debated over the exact "dimensionality" of the test,[49] and opinion on the matter ranged from "two" to "six". Belief in the relevance of these results to judgments as to the test's quality was strong enough to motivate a number of major revisions of the test (e.g., Gangestad & Snyder, 1985; Lennox & Wolfe, 1984). However, a careful consideration of the test's TS, as described by its creator, suggested TS(DI,1,MI,EIV). Failure of $f_{\underline{X}}$ to conform to a *linear* factor analytic QC, however, is not grounds for indicting a test with TS(DI,1,MI,EIV).

In fact, Fleisher and Baize (1982) had argued convincingly that the theoretical structure of the self-monitoring scale was, in fact, TS(DI,1,U,EIV). A QC that is isomorphic to this TS is the unidimensional, quadratic factor model. Once again, if this is the TS of the self-monitoring scale, then the mountain of factor analyses conducted on the test is irrelevant to judgments about its performance. As is well known (e.g., McDonald, 1967), the fact that a test is multidimensional in a linear factor analytic sense does not, in any way, imply its lack of conformity to either of the above-stated QC's.

3. Test of the Conformity of Data to Model

To engage in a test analysis is, minimally, to assess whether a test conforms to its theoretical structure. But, because the theoretical structure of a

[49] Inadvertently infecting the concept of *dimensionality* with a generic quality that rendered it meaningless.

test is a pre-mathematical specification of how a set of test items must behave, TS/T conformity is only possible given that the TS has been translated isomorphically into quantitative terms, the translation called the QC. The test analyst tests whether $f_{\underline{X}}$ satisfies the requirements specified by the QC, based on a random sample of respondents from focal population **P** of interest. It must be emphasized that, just as the drawing of invalid implications of theory in standard construct validation investigations results in the carrying out of irrelevant tests (Cronbach & Meehl, 1955), so too does the construction of a poor TS/QC match.

A general approach for assessing the conformity of $f_{\underline{X}}$ to a QC is as follows: Let $\underline{M}$ be the vector of parameters of $f_{\underline{X}}$ about which the QC makes claims; let $\underline{M}_{\mathbf{P}}$ be the value of $\underline{M}$ in population **P**; let $\tilde{\underline{M}}_{\mathbf{P}}$ be a value of $\underline{M}_{\mathbf{P}}$ that is generated by the QC (i.e., results from a numerical instantiation of the QC's parameters); let $F(\tilde{\underline{M}}_{\mathbf{P}}, \underline{M}_{\mathbf{P}})$ denote some fit function which quantifies the "distance", in some particular sense, between $\underline{M}_{\mathbf{P}}$ and any possible $\tilde{\underline{M}}_{\mathbf{P}}$; and let $\tilde{\underline{M}}^{\bullet}_{\mathbf{P}}$ be that value of $\tilde{\underline{M}}_{\mathbf{P}}$ chosen so that $F(\tilde{\underline{M}}^{\bullet}_{\mathbf{P}}, \underline{M}_{\mathbf{P}})$ is a minimum over all possible $\tilde{\underline{M}}_{\mathbf{P}}$. Test T can justifiably (but provisionally) be said to conform to its TS in population **P** if i) The chosen QC is a "good" paraphrase of TS; and ii) $F(\tilde{\underline{M}}^{*}_{\mathbf{P}}, \underline{M}_{\mathbf{P}})$ is "small". For example, if the TS of test T is TS(C,1,LI,EIV), then an isomorphic QC is the unidimensional, linear common factor model; hence, a test

of T's conformity to its TS in population **P**, is a test of whether there exists a model-implied covariance matrix $\tilde{\Sigma}_{\mathbf{P}}$ that is "close" to $\Sigma_{\mathbf{P}}$, in a sense of closeness defined by some fit function $F(.,.)$.

In practice, a decision regarding the size of $F(\tilde{\underline{M}}^{*}_{\mathbf{P}}, \underline{M}_{\mathbf{P}})$ is made on the basis of a sample of respondents drawn from population **P**, this adding a further complication due to the now inferential nature of the problem. Regardless of which particular fit function is employed, the point is that in order to justifiably claim that a test is (or is not) in keeping with its TS, there must be some *formal* assessment of whether $f_{\underline{X}}$ conforms to an appropriately chosen QC. In the absence of sound justification for judging a test's conformity to its TS, any further steps taken in an evaluative test analysis are inherently ambiguous, as the analyst has not established that the test items are indicators of a single attribute, nor that they relate to this attribute in a construct valid manner (i.e., in the manner described by TS).

4. Derivation of an Optimal, Model-Implied Composite

To composite a test is to produce a scalar function of $\underline{X}$. The objective in employing a test comprised of k items, with these items designed to be indicators of an attribute, γ, is to generate a composite whose realizations can justifiably be seen as error-laden measurements (estimates) of γ. If such a composite can be produced, then it assigns to each individual in focal population **P** a single real

number, thereby scaling this population of individuals with respect to attribute γ. Very often the tests employed in the social and behavioural sciences come with "off-the-shelf" compositing rules, and the undisputed champion of such rules is the unweighted sum. However, in true, evaluative test analysis, the very issue of a test's compositability in some population **P** is open to question, and claims of compositability must be justified. If a test can be shown to be justifiably compositable, then the issue becomes *which* composite to use. The logic is as follows:

i. The TS of given test, T, claims that the *k* items of T are indicators of, or measure in common, a single unobservable attribute, γ.

ii. The TS is mapped into an appropriate QC.

iii. If $f_{\underline{X}}$ satisfies the requirements imposed by the chosen QC, then the performance of test T is in keeping with its TS. In particular, for any of the standard undimensional QC's that dominate applied test analytic practice, satisfaction by $f_{\underline{X}}$ of the requirements imposed by QC means that the X_j, $j = 1, 2, ..., k$, are unidimensional in some particular sense. As follows from the TS/QC relationship, this, in turn, is taken as meaning that the items measure but one thing in common, arguably γ.

iv. However, the TS/QC relationship paraphrases γ as a synthetic random variate of some sort. In the case of the usual latent variable model paraphrases, this synthetic variate is a random, latent variate.

v. Thus, the task of scaling individuals in **P** with respect to the unobservable attribute γ is paraphrased as the task of deriving an *optimal* predictor of the random variate proxy to γ, i.e., the latent variate.

vi. When $f_{\underline{X}}$ satisfies QC, and QC is a unidimensional (latent variable) model, then one is justified in predicting (estimating) the *single* latent variate. That is, this is the condition under which it makes *sense* to develop a single optimal predictor of a single unobservable latent variate that is related to the X_j in a manner described by the QC (i.e., in a manner that is in keeping with TS). This predictor will be, of course, a composite of the X_j.

vii. The particular form of the optimal compositing rule will be determined jointly by characteristics of the QC, commitment to a particular definition of *optimal*, and pragmatic considerations.

As an example, for a test with TS(C,1,LI,EIV), an appropriate QC is the unidimensional, linear common factor model. If $f_{\underline{X}}$ satisfies the requirements imposed by this QC, then the items are unidimensional in the linear factor analytic sense,[50] and the common factor is taken as a proxy for the attribute the items were designed to measure. Under this condition, one is justified in deriving a predictor of the common factor, an unobservable random variate.

[50] That is, have a representation as unidimensional in the linear factor analytic sense.

And, according to the TS/QC correspondence, prediction (estimation) of the common factor is the operational counterpart of scaling individuals with respect γ.

There have been derived many different brands of prediction (estimation) of the latent variate in a unidimensional, linear factor model, each answering to a different sense of optimality (see, e.g., McDonald & Burr, 1967). Moreover, there exist several sub-classes of the unidimensional, linear common factor model, with the particular sub-class of models depending on what, if any, restrictions have been imposed on the model parameters. Employing the principle of (conditional) maximum likelihood estimation,[51] Thissen et al. (1983) show that, if the X_j have equal loadings and equal residual variances, then the unweighted sum of the X_j (or any statistic proportional to it) is an optimal compositing rule. If, on the other hand, both the loadings and unique variances are free to vary over items, the compositing rule assumes the decidedly more complex form

[51] The latent variable models considered herein, and throughout psychometrics, are random latent variable models (i.e., those in which the latent variable has a distribution). Such models are to be contrasted with those in which each person has a "person parameter" to be estimated. However, as Holland (1990) points out, there is no true sense to the notion of maximum likelihood estimation (or any other type of estimation) of θ in the random models, as θ is not a set of person parameters, but, rather, a random variate. Hence, in random latent variable models, θ may be "predicted" but not estimated. However, maximum likelihood terminology will be used here in order to maintain consistency with standard treatments. An additional complication arises in cases in which the chosen QC is an indeterminate latent variable model. In such models, there exist an infinity of random variates, U_i, each of which satisfies the requirements for latent variable-hood (cf. Guttman, 1955 for a discussion of the determinacy of factor score matrices), hence leaving completely ambiguous the question of what exactly is being predicted. Although these issues will undoubtedly ultimately need to be resolved by psychometricians, they do not bear on the logical coherence of the framework proposed here.

$$\left\{ \frac{\sum_{j=1}^{k} \left(\frac{\hat{\lambda}_j}{\hat{\sigma}_j^2} \right) (x_{ij} - \hat{\mu}_j)}{\sum_{j=1}^{k} \left(\frac{\hat{\lambda}_j^2}{\hat{\sigma}_j^2} \right)} \right\}. \tag{5.3}$$

However, as Wainer (1976) has indicated, it is frequently the case that the properties of unweighted and weighted composites of the same set of X_j are virtually identical. Thus, it is often reasonable to prefer an unweighted composite over a weighted counterpart on the grounds of ease of calculability. On the other hand, the possibility of this virtual exchangeability should not be taken as justification for the lazy practice of choosing an unweighted composite as the default choice. The preference of a particular composite over all others should be the result of a careful consideration of optimality and practicality tradeoffs.

A final point should be underscored: There is no globally correct compositing rule. Any legitimate scoring rule is tied to the union of a particular QC (i.e., test theory model), statistical principle, and pragmatic considerations, and there exists latitude in regard to the choice of each. That is to say, the choice of compositing rule may only be justified on situation-specific grounds, i.e., justified conditional on the pairing of the chosen QC and reasonable choice of statistical principle, along with pragmatic considerations.

5. Estimation of the Reliability of the Composite

There is no such thing as "test reliability". A test is comprised of k items, a set of response options, and a scoring rule that converts the responses of individuals to the items, as encoded by the response options, into a set of real numbers. Reliability is defined as the ratio of true score variance to observed score variance, and, hence, is clearly a property of a random variate in some focal population **P**. The random variates whose reliabilities are of interest in test analysis are those of composites of $\underline{X}$, which contains the k random variates that represent, in **P**, the distributions of the scored responses to the k items. analyst refers to "the reliability of test T in population **P**", what she can only be referring to is the reliability, in population **P**, of some particular composite of the X_j. But if test T is comprised of items that were designed to measure an unobservable attribute, γ, then the aim is obviously to estimate the precision of some particular composite of the X_j, this composite taken, provisionally, as yielding measurements of γ. It is here that the sequential nature of true, evaluative test analysis is perhaps most evident. For one cannot estimate the reliability of a composite whose values can justifiably be interpreted as error-laden scale values of γ unless one can produce such a composite. And one can produce such a composite only if there can be derived an optimal, model-implied compositing rule that predicts (estimates) a random variate that is a proxy for γ. And such an

optimal, model-implied compositing rule may be justifiably derived only when $f_{\underline{X}}$ satisfies the requirements of a QC that is a sound paraphrase of the TS(.,1,.,.) of test T.

In an analysis of the performance of test T in population **P**, if a sound QC of the TS is shown to describe $f_{\underline{X}}$, one is justified in compositing the X_j, with the optimal, model-implied composite symbolized as $\phi = f(\underline{X})$. The precision of the composite will, in general, be a function of θ, and can be given quantitative expression as the "information" function of ϕ:

$$I(\theta,\phi) = \frac{[E(\phi \mid \theta)']^2}{V(\phi \mid \theta)}, \tag{5.4}$$

The numerator is the squared derivative of the expectation of ϕ conditional on θ, and the denominator is the variance of ϕ conditional on θ. As is implied by (5.4), the precision of the composite will be greatest at points on θ at which the slope of the "test response function" (i.e., $E(\phi \mid \theta)$) is steep relative to the variance of ϕ at θ. However, for homoscedastic models, in which the IRF's are linear, precision will not vary across different values of θ, i.e., $I(\phi \mid \theta)$ will be a constant. Once again, a general expression for a lower bound to the reliability of ϕ (a lower bound because θ, defined under the principle of local independence, is not the same thing as a true score) is

(5.5) $$\rho^2_{\phi\theta} = 1 - \frac{E[V(\phi \mid \theta)]}{V(\phi)}.$$

Distinctive forms of both (5.4) and (5.5) can sometimes be derived for particular QC's, but it is a misunderstanding to believe that there exist multiple "types" of reliability (e.g., test-retest, alternate forms, etc.), or, correspondingly, a single thing called "test reliability".

6. Entering the Composite into Construct Validation Studies

The traditional categories of reliability and validity enter the sequential framework at very particular points. The TS/QC pairing, and the subsequent testing of the conformity of $f_{\underline{X}}$ to QC, represents the internal facet of a construct validation program involving a test T. It is the first part of the overall validity case that arises when the test user wishes to claim that the items of a test are indicators of an unobservable attribute, γ. Satisfaction of this first validity requirement is the justification for the test user producing a composite of the X_j, and entertaining the possibility that the scores on this composite are (error-laden) measurements of γ. If, furthermore, this composite has been shown to possess adequate reliability in population **P**, then it can justifiably be used in further investigations. To this point, there has been amassed support that the *k* items of T measure in common but one thing, and there has been generated no evidence that this one thing is *not* γ. On the other hand, although a composite has been produced to predict (estimate) the attribute that the items measure, it has also not

been settled that this one thing *is* γ. In fact, no definitive case can ever be made about the identity of the attribute that the composite measures (Cronbach & Meehl, 1955). Certainly, however, a great deal more evidence can be accumulated that has direct bearing on the provisional claim that the scores on the composite are error-laden measurements of γ. This evidence is accumulated in an ongoing program of (external) construct validation.

The logic of such a program of investigation can be outlined as follows:

i. Let test T be comprised of k items $\{t_1, t_2,...,t_k\}$, these designed to be (observable) indicators of an unobservable attribute γ, $\{\eta_1, \eta_2, ...\}$ denote additional putative observed indicators of γ, and $\{\pi_1, \pi_2, ...\}$ denote putative observed indicators of additional unobservable attributes (properties, mechanisms, etc.) $\{\omega_1, \omega_2, ...\}$. Also, let OT_{ϕ}, $OT_{\eta i}$, and $OT_{\pi i}$ stand for the "observation terms" that designate the observables η_i and π_i, and $TT\gamma$ and $TT_{\omega i}$ stand for the "theoretical terms" that designate unobservables γ and the ω_i.

ii. Variation in the responding of individuals in focal population **P** to the items of test T, this variation encoded in the distribution of the random variates $X_j, j = 1, 2, ..., k$, is caused by a complex web of action involving unobservables γ and the ω_i.

iii. Test T is, in the sense of Cronbach and Meehl (1955), a *construct valid* measure of γ if responding to the items of T is causally determined by γ.

In practice, variation in ϕ will not be due strictly to γ, but also to additional unobservable sources (e.g., situational or method factors); hence, the proportion of variance in ϕ due to γ will be less than unity. A construct valid test of γ is, thus, a test for which γ is largely responsible for the responding of individuals to the items of T (i.e., one in which the test items are, to a high degree, pure indicators of γ).

iv. Because γ is unobservable, a direct assessment of its causal action on the responding of individuals to the test items is not possible. Instead, theory that postulates the relationships between γ and other constituents of its *nomological network* must be developed, and testable consequences of this theory, deduced and tested.

v. A nomological network, TH, consists in an interlocking system of hypotheses and laws relating a) observable entities to each other, b) theoretical entities to observables, and c) theoretical entities to other theoretical entities (Cronbach and Meehl, 1955). In terms of the current notation, TH then consists in a network of relations between the sets $\{\gamma\}$, $\{t_1, t_2,...,t_k\}$, $\{\eta_1, \eta_2, \ldots\}$, $\{\pi_1, \pi_2, \ldots\}$, and $\{\omega_1, \omega_2,\ldots\}$. TH is divisible into theory describing the relationships between $\{\gamma\}$ and $\{t_1, t_2,...,t_k\}$ (this theory called, herein, the TS) and theory (called, herein, TE) relating the test items and γ to "external" observables and unobservables.

vi. If steps (1-5) of the test analytic framework have been satisfied, then the optimal composite, ϕ^*, is a predictor of latent variate, θ, that is taken to be a proxy for unobservable attribute, γ. Thus, ϕ^* stands in for the items of test T in all external construct validation analyses. In particular, evidence of test T's construct validity accrues from ϕ^* behaving in a manner that is in keeping with testable consequences of TE.

vii. Because there are, in principle, an infinity of testable consequences of TE, and, at any stage in an ongoing program of construct validation only a small subset of these can be derived and tested, a given test, T, is, in principle, only ever deemed to be *provisionally* construct valid.

A Case Study

The Attributional Complexity Scale (ACS) (Fletcher, Danilovics, Fernandez, Peterson, & Reeder, 1986) is a twenty-eight item test, each item having associated with it a seven-point Likert response scale (strongly disagree to strongly agree). The test was designed to measure the "complexity of attributional schemata for human behavior" (Fletcher et al., 1986), and the twenty-eight items are organized into sets of four, with each set designed to measure one of seven facets of attributional complexity: 1) level of interest or motivation; 2) preference for complex rather than simple explanations; 3) presence of metacognition concerning explanations; 4) awareness of the extent to

which people's behavior is a function of interaction with others; 5) tendency to infer abstract or causally complex internal attributions; 6) tendency to infer abstract, contemporary, external causal attributions; and 7) tendency to infer external causes operating from the past.

In the original study, the scale was given passing marks with respect to its psychometric characteristics: "The results of these studies provide encouraging support for the internal and external validity of the Attribution Complexity Scale" (Fletcher et al., p.682); "The positive significant item-total correlations, the positive correlations between the seven attributional constructs, and the factor analysis results all support our contention that the scale measures one construct—attributional complexity" (Fletcher et al., p.682).

1. The Theoretical Structure

What is the theoretical structure of the ACS? The discussion of Fletcher et al. is somewhat obscure. The items are seven-point Likert and so may be treated as "pseudo-continuous", i.e., I=C. However, the twenty-eight items fall into seven sets (four items per set), each set corresponding to one of seven facets of attributional complexity. How many attributes are the items designed to measure? On the one hand, the reader is told that "...the scale measures one construct - attributional complexity" (Fletcher et al., 1986, p.878). On the other hand, the items are partitioned into seven sets, and each is treated as if it is designed to measure a distinct attribute: "The central hypothesis underlying the

development of this scale is that the attributional constructs just described are all related in a consistent fashion..." (Fletcher et al., 1986, p.877). Furthermore, "...these seven dimensions may be related to attributional complexity..." (Fletcher et al., 1986, p.877). It would appear then that, although there may be higher-order structural requirements at the facet scale level (i.e., in which the seven facet attributes are related to the attribute of attributional complexity itself), each facet scale should be treated as comprised of items designed to be indicators of one of the seven facet attributes. Hence, it may be concluded that, for each facet scale D=1.

Little guidance as to the form of the item/attribute regressions is provided by the original test analysis of Fletcher et al. However, the general tone of the article suggests that these regressions should be taken as monotone increasing (MI). Finally, in modern test construction, it is practically a default that the items are viewed as fallible, errors-in-variables (EIV), indicators. Hence, the best guess at a theoretical structure seems to be, for each set of four items, TS(C,1,MI,EIV). Further theoretical considerations would be required to specify the TS of the seven facet scales as putative indicators of attributional complexity. And, of course, to even enter into such higher-order analyses, each of the seven facet scales would have to be shown to be compositable, and if so, the resulting composites shown to possess adequate precision (i.e., to satisfy steps one to five of the proposed framework).

2. The Quantitative Characterization

An appropriate QC for this TS is a UMLV model. As described earlier in this chapter, a UMLV model is a paraphrase of TS(C,1,MI,EIV) in the following sense:

TS		QC
γ	→	Random, unobservable, latent variate θ
D=1: The items measure one attribute, γ	→	$f(\underline{X} \mid \theta) = \prod_{j=1}^{k} f(X_j \mid \theta)$
R=MI: Each item has a monotone (increasing) regression on γ	→	$E(\underline{X} \mid \theta) = \underline{g}(\theta)$, in which $\underline{g}$ is a vector of functions, and $\frac{d}{d\theta}\underline{g}(\theta) > \underline{0}$
E=EIV: Each item is a fallible indicator of γ	→	$C(\underline{X} \mid \theta) = \Psi$ is diagonal and positive definite (the X_j are allowed to have positive "error variances")

Now, although the test analyst could test conformity of the ACS to its TS by employing one or more of the UMLV consequences documented by Holland and Rosenbaum (1986), it is also the case that linear increasing regressions are a sub-class of monotone increasing regressions. Hence, the unidimensional, linear factor model is a sub-class of the UMLV models, and the test analyst might begin by testing the hypothesis $\Sigma = \underline{\Lambda}_r \underline{\Lambda}_r' + \Psi$, for $r = 1$. The point to note, here, is that a

lack of conformity of $f_{\underline{X}}$ for a given facet scale to this QC would not constitute grounds for indicting the ACS. Conversely, conformity of $f_{\underline{X}}$ to the unidimensional, linear factor analytic QC would constitute evidence that the items of a given facet scale were in keeping with TS(C,1,MI,EIV).

A Summary of Key Test Analytic Rules

A test theory and a test analytic framework are distinct entities. Whereas a test theory is a collection of mathematical tools developed to model the relationship presumed to exist between responding to a set of test items and the attribute the items were designed to measure, a test analytic framework consists in a set of rules that stipulates how such a theory should be employed to pass judgement on the performance of a test. Such a framework must settle what are meant by the key evaluative notions that will be employed to express the evaluative decisions that are the end product of the application of the framework. As has been described, the test analytic framework presented herein is intrinsically sequential in nature. In the following, the rules of which the framework is comprised are reiterated in a stipulative, axiomatic fashion.

Rule 1:

What it means to correctly, but provisionally, claim that the performance of a test is "satisfactory" in a focal population **P** is the following:

1A) The test conforms to its TS, which, empirically, means that $f_{\underline{X}}$ satisfies the requirements imposed by a QC that is isomorphic to TS.

1B) Conditional on (1A) having been satisfied, the test possesses adequate precision in the sense described in step five of the framework.

1C) Conditional on (1B) having been satisfied, the items of the test, or, more usually, an optimal composite of these items, has a nomothetic span that is in keeping with testable deductions from the nomological network of the attribute that the items were designed to measure.

Rule 2:

What it means to have *appropriately* analyzed a test is to have done at least one of the following:

2A) judged fairly the conformity of $f_{\underline{X}}$ to an *appropriately* chosen QC, i.e., a test theory model that is isomorphic to TS;

2B) conditional on (1A) having been satisfied, estimated a lower bound to the reliability, or generated the information function, of an optimal, model-based composite of the items;

2C) conditional on all requirements specified in steps 1-5 having been satisfied, judged fairly the agreement of elements of the test's nomothetic span with testable deductions from the nomological network of the attribute that the items were designed to measure.[52]

Rule 3:

In a given test analysis, a fact is *relevant* to the passing of judgment on the performance of a test if it:

3A) bears on the judgment of the conformity of the test to its TS;

3B) conditional on (1A) having been satisfied, i) is an estimate of a lower bound to the reliability of an optimal composite of the items; ii) is the information function of an optimal composite of the items;

3C) conditional on (1B) having been satisfied, bears on judgments regarding the conformity of elements of the tests nomothetic span to testable deductions from the nomological network of the attribute that the items were designed to measure.

[52] Stipulations (2A) through (2C) imply that one could, conceivably, carry out but a subset of the steps outlined above, and still be conducting a test analysis. Although it is difficult to imagine scenarios in which one would be interested in, for example, establishing only that a test conforms to its theoretical structure, or that a composite, whose formation has been justified on the basis of T/TS conformity having been demonstrated, has adequate reliability, doing so does not represent a logical fallacy of any sort. In other words, even though researchers will often be interested in investigating the "external" validity of some composite of the items of a test, and will, hence, need to satisfy the requirements in all six of the steps in the framework, there is justification for necessarily doing so.

Rule 4:

A test is compositable if it conforms to its TS, and its TS is isomorphic to a unidimensional QC. The exact form of the compositing rule is determined jointly by the particular features of the QC, a chosen statistical principle, and pragmatic considerations.

Rule 5:

The "measurement precision of a test" *means* "the measurement precision of a composite of a test's items". Hence, the measurement precision of a composite of a test's items is coherently estimable only if the test is, in fact, compositable (see rule 4).

Rule 6:

To "enter a test into (external) construct validation investigations" is to enter an optimal composite of the test's items into such investigations. Hence, such investigations can coherently be carried out only if the requirements of steps one to five of the framework, the internal test analytic components, have been satisfied.

American Psychological Association (APA) Standards for Educational and Psychological Testing and Other Potential Competing Test Analytic Frameworks

APA Standards for Educational and Psychological Testing

A moment's consideration will make it clear that the framework that has, herein, been explicated, does not exhaust the rules that bear on coherent test analytic practice. The opening statement of the introduction to the APA *Standards for Educational and Psychological Testing* (1999) reads:

> Educational and psychological testing and assessment are among the most important contributions of behavioral science to our society...The proper use of tests can result in wiser decisions about individuals and programs...and also can provide a route to broader and more equitable access to education and employment. The improper use of tests, however, can cause considerable harm to test takers and other parties affected by test-based decisions. The intent of *Standards* is to promote the sound and ethical use of tests and to provide a basis for evaluating quality of testing practices. (p. 1)

And then later, "The purpose of publishing the *Standards* is to provide criteria for the evaluation of tests, testing practices, and the effects of test use...*Standards* provides a frame of reference to assure that relevant issues are addressed" (p. 2).

Standards defines reliability as "the consistency of...measurements when the testing procedure is repeated on a population of individuals or groups" (p. 25), and states that "Information about measurement error [i.e., "*un*reliability"] is essential to the proper evaluation and use of an instrument";..."The ideal approach to the study of reliability entails independent replication of the entire measurement process" (p. 27). *Standards* claims that, in addition to the usual

classical reliability coefficients (i.e., alternate forms, test-retest, and internal consistency),[53] reliability information may also be reported in terms of variances or standard deviations of measurement errors (as in the Generalizability Theory approach) or IRT-based test information functions. Thus,

> **Standard 2.1**: For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant reliabilities and standard errors of measurement or test information functions should be reported (p. 31).
>
> **Standard 2.4**: Each method of quantifying the precision of consistency of scores should be described clearly and expressed in terms of statistics appropriate to the method...(p. 32).
>
> **Standard 2.7**: When subsets of items within a test are dictated by the test specifications and can be presumed to measure partially independent traits or abilities, reliability estimation procedures should recognize the multifactor character of the instrument (p. 33).

The conclusion is that the APA requires that estimates of reliability (consistency, or precision), and associated standard errors, be reported for each composite, whether it is based on a subscale or the total test, that the nature of such estimates should be described clearly, that the decision to use particular estimates be justified, and that the dimensionality of the responses to test items be reported.

According to *Standards*, legitimate sources of validity evidence include evidence based on: test content, response processes, internal structure of the test,

[53] Which, once again, it should be clarified, are not distinct types of reliability, but are different means of producing parallel items.

relations with other variables (in terms of convergent or discriminant evidence or test-criterion relationships), and consequences of test taking. A sound validity argument is, according to *Standards*, one that "integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses" and "may indicate the need for refining the definition of the construct, may suggest revision in the test or other aspects of the testing process, and may indicate areas needing further study" (p. 17). Twenty-four standards with regard to validity are specified, these concerning issues that range from test interpretation and use to justification regarding the selection of additional variables whose relations to the test may be sited as validity evidence.

> Of particular interest here are the following two validity standards:
>
> **Standard 1.11**: If the rationale for a test use or interpretation depends on premises about the relationships among parts of the test, evidence concerning the internal structure of the test should be provided (p. 20).
>
> **Standard 1.12**: When interpretation of subscores, score differences, or profiles is suggested, the rationale and relevant evidence in support of such interpretation should be provided. Where composite scores are developed, the basis and rationale for arriving at the composites should be given (p. 20).

The first states that the test developer (or user) must explicitly report on dimensionality of the test items, and must provide evidence that the (hypothesized) structure is, in fact, in keeping with expectations regarding which and how many attributes are being measured by the test. The second can be read

as the requirement that the test developer (or user) must make explicit the nature of any composite, i.e., must specify the compositing rule employed.

Standards contains many other rules, including those that bear on the use of tests in populations other than that in which the test was originally validated, on the rights and responsibilities of test takers, and on various testing applications. Such rules are, of course, extremely important. However, the tone with which many of the rules in *Standards* are presented suggests that, taken together, they represent a test analytic framework: This is not the case. Jointly, the rules laid down in *Standards* provide a broad, non-specific take on the concerns that test developers, analysts, and users must address. The "criteria for the evaluation of tests" that it does provide are non-technical, loosely-defined, and non-methodical. It does not give formal definitions of any of the technical test analytic concepts to which it refers, and is completely non-specific with regard to the conditions under which particular techniques (e.g., the various reliability/validity coefficients, item response models, etc.) may be justifiably employed. Furthermore, it confuses certain fundamental distinctions, for example, between coefficients of precision of measurement and coefficients of stability of measurement, and between internal and external components of construct validation. Finally, it does not ever specify in unequivocal terms *how* the test analyst is to employ the tools provided by test theory such that he may pass sound judgment on test *performance*. When it does stray into such

specification, what it offers is often confused, as in its implied reference to test analysts being able to choose freely among distinct reliability coefficients.

For example, although it is true that the veracity of the results of a given test analysis requires that the estimates of reliability, or precision, reported are the "relevant" ones, that each method employed in order to quantify precision is "expressed in terms of statistics appropriate to the method", that "reliability estimation procedures should recognize the multifactor character" of certain tests, that it must be demonstrated that the "internal structure" of the test conforms to pre-specified "premises about the relationships among parts of the test", and that a rationale be given for employing particular composite scores, it is never specified *how* the test analyst is to justify that a given estimate of reliability is *relevant* or *appropriate,* that test behaviour is in keeping with expectations about how the test *ought to perform,* and so on. In other words, *Standards* does not provide a *logic,* specified in technical terms, for pronouncing on the quality of tests. Rather, it consists in a set of *ethical criteria* for test constructors and analysts alike for how tests may be *used,* and for which issues must be addressed if the results from a test analysis are to be taken seriously.

The framework elucidated in the current work, by contrast, provides a detailed, technical, sequential framework to which test constructors and test analysts may refer in order to determine what is, for a given test analysis, the "relevant", or "appropriate", coefficient of precision, of external validity, and so

forth, to be used. It specifies the conditions under which the forming of particular composites, and the subsequent estimation of the precision with which they measure, can be legitimately *called for*. It specifies the conditions under which a composite may be legitimately entered into (external) construct validation studies. More generally, it lays down *rules* that fix how the test analyst should coherently proceed with the two most fundamental (and commonly carried out) steps of a test analysis: 1) The adjudication of whether the items of a test can justifiably be said to measure reliably, but one attribute in common and; 2) the adjudication of whether this single attribute is the attribute that the test items were designed to measure. It will answer to questions about whether a given compositing rule may be rationalized in a particular situation, about the *relevancy*, or *appropriateness* of a given coefficient for estimating reliability, about the conformity of the internal structure of a test to certain theoretically derived premises, and so on.

Other Competing Frameworks?

Messick

Samuel Messick has written extensively on the issue of validity, addressing issues pertaining to the ethics of assessment (Messick, 1980), the evidential basis of test interpretation (Messick, 1989), assessing the meaning and consequences of measurement (Messick, 1988), and the components of a

construct validity approach to validation (Messick, 1995). He defines validity as "an inductive summary of both the existing evidence for and the potential consequences of test interpretation and use" (1988, p. 43).

Messick has long highlighted the distinction between what he considers to be the two fundamental aspects of validity, viz., *construct* and *consequential* validities, the former of which bears on the adequacy of the test as a measure of some attribute it is interpreted to measure, and the latter of which refers to the appropriateness of the employment of the measure in specific applications (cf. Messick, 1980). He has described construct validity as "the evidential basis of test interpretation" (1980, p. 1019), involving both convergent and discriminant evidence pertaining to theoretically relevant empirical relationships 1) between the test and different methods for measuring the same construct and 2) between measures of some construct of interest and measures of different constructs predicted to be related in particular ways to the primary construct under study. Consequential validity, conversely, he has described as involving an evaluation of the impact of potential consequences of both test use and test interpretation, "especially those unintended side effects that are distal to the manifest testing aims" (1980, p. 1020).

Test validity, Messick has argued, is "an overall evaluative judgment of the adequacy and appropriateness of inferences from test scores" (1980, p. 1023), this evaluation resting on four bases: 1) an inductive summary of convergent and

discriminant evidence bearing on the interpretability of test scores in reference to a particular construct; 2) an appraisal of the value implications of such interpretations; 3) a rationale and evidence for the relevance of the construct and utility of test scores for particular applications; and 4) an appraisal of potential social consequences of the proposed use and actual consequences of use.

Mislevy, Steinberg, and Almond

In 2003, Mislevy, Steinberg, and Almond presented a complex and provocative framework, which they call the "evidence-centered" assessment design (ECD), that "makes explicit the interrelations among substantive arguments, assessment designs, and operational processes", the motivation behind the creation of which was "the need to develop assessments that incorporate purposes, technologies, and psychological perspectives that are not well served by familiar forms of assessments" (p. 3). These researchers described the relationships among 1) the motivation behind the assessment, i.e., the claims that one desires to make about students; 2) the principles upon which this reasoning is based; and 3) the "pieces of machinery", i.e., the tasks, responses, rubrics, statistical routines, score reports, and so on, that one assembles in order to gather evidence in support of claims about students.

Four stages of assessment design are identified: 1) Domain Analysis—collecting substantive information about the assessment domain, 2) Domain Modelling—organizing the information collected in (1) in terms of design

"paradigms, viz., proficiency paradigms, evidence paradigms, and task paradigms; 3) Conceptual Assessment Framework (CAF)—specifying a model which consists in a "blueprint" for the "operational elements" of an assessment; and 4) Operational Assessment—involving four principle processes that occur in assessment delivery, viz., presentation, evidence identification, evidence accumulation, and activity selection.

Kane

Michael Kane has proposed an "argument-based approach" to validity, in which an interpretative argument is adopted as a framework for collecting and presenting validity evidence (cf. Kane, 1992). Interpretative arguments, according to Kane, may be evaluated, by three general criteria: clarity of argument, coherence of argument, and plausibility of assumptions. Kane claims that, like all practical arguments, interpretative arguments "may have some inferences and assumptions that can be evaluated unambiguously", but that "Confidence in other inferences and assumptions depends on the accumulation of various kinds of evidence, none of which is completely decisive" and, furthermore, that the "plausibility of the argument as a whole is limited by its weakest assumptions and inferences" (p. 528).

Kane specifies six categories of inferences that standardly appear in interpretative arguments, each of which rests on assumptions that provide justification for the inference. Theses are: 1) *observation*—the acceptance that the

methods used to assign a particular numerical value to a given examinee's response (i.e., to produce a "score" for a given item) were consistent with the definition of the measurement procedure; 2) *generalization*—drawing conclusions about a universe of possible observations on the basis of a limited sample of actual observations (i.e., sample of item composites). The veracity of inferences of this type rests on the assumption that "the results are largely invariant with respect to changes in the conditions of observations" (p. 529), i.e., that the composite scores are sufficiently "reliable"; 3) *extrapolation*—making inferences about non-test behaviours on the basis of test behaviour; 4) *theory-based inferences*—explicitly or implicitly explaining test scores in terms of the theory or theories about the construct thought to be measured by the test; 5) *decisions*—interpreting test scores in light of a decision or set of decisions which motivate test use; 6) *technical inferences*—taking into account any assumptions which may be attached to the various technical apparatuses that are employed in examining validity, e.g., those associated with the employment of particular score-equating procedures.

Although the above-described treatments address, each in its own way, the complexity of the concept of validity, and a potentially diverse set of issues subsumed under the banner of "validation", it would, I believe, be a mischaracterization to refer to any of these works a *test analytic* framework, i.e., a logically coherent set of interrelated steps for pronouncing on the performances

of tests. Each consists in a commentary on components of validity and on methods of validation, or assessment more generally. However, although each of these treatments presumably *assumes* the existence of a coherent framework for analyzing test data such that validity evidence may be accumulated, not one includes a clearly specified prescription for how the test analyst is to proceed, i.e., what he or she must do, and in what order, such that any test-based "evidence" may coherently be considered to bear on the broader issue of validity/assessment. That is, each *presupposes*, and requires, that a test analytic framework such as the one proposed here could be "inserted" in the relevant place into the broader framework described. Hence, they do not constitute competitors to the current framework, but, rather, descriptions of particular orientations to validation in which such a framework could reasonably be employed.

Future Directions

Although the framework proposed herein consists, I believe, in a sound and practically useful set of test analytic rules, there is no question that it does not constitute an exhaustive treatment of the many contours of the test analytic game, and could be improved upon in a number of important ways. Three of the perhaps more pertinent areas requiring further work are outlined briefly below.

The Specification of Sufficiency Conditions for the QC

Maraun, Slaney, and Goddyn (2003) describe the logic underlying the justification that a given manifest property, C, is a criterion for a particular latent structure, LS. They specify two distinct senses of criteria of latent structures: 1) A sense 1 criterion for a latent structure states that C is a criterion for LS if LS⇒C and, equivalently, ~C⇒~LS, i.e., C is a *necessary condition* for LS; 2) a sense 2 criterion for a latent structure states that C is a criterion for LS if C⇒LS and, equivalently, ~LS⇒~C, i.e., C is a *sufficient condition* for LS.

In the context of the present work, as it stands, the proposed framework specifies only a sense 1 criterion for a given QC. In Maraun et al.'s (2003) terms this can be represented as follows: Let LS= $f_{\underline{X}}$ described by a specific QC, and C=particular empirical requirements of the QC for $f_{\underline{X}}$. As above, if it is true that LS⇒C, then C is a necessary condition for LS, and, equivalently that ~C⇒~LS. For example, a necessary condition for $f_{\underline{X}}$ described by a ULCF model (i.e., that there exists a random variate, θ, with $E(\theta) = 0$ and $V(\theta) = 1$, and k×1 random vector, $\underline{\delta}$, with $C(\underline{\delta}) = \Psi$, Ψ diagonal) is that $\Sigma = \underline{\Lambda}\,\underline{\Lambda}' + \Psi$. From this one can claim that if *not* $\Sigma = \underline{\Lambda}\,\underline{\Lambda}' + \Psi$ then $f_{\underline{X}}$ is not described by a ULCF model (and, ultimately, that the test does not conform to its TS).

However, in order to answer to the needs of truly evaluative test analysis, such that one can make legitimate claims that a test conforms to its TS, requires the specification of sense 2 criteria for particular QC, i.e., the specification of

sufficient conditions (and optimally, necessary *and* sufficient conditions) for the QC. Specifically, this means that when particular empirical requirements for $f_{\underline{X}}$ hold, then $f_{\underline{X}}$ is describable by a given QC. For example, suppose the theoretical structure of a test is TS(C,2,LI,EIV), with the chosen QC being a two-dimensional linear common factor model(2-d LCFM). A necessary condition for $f_{\underline{X}}$ to be described by a two-dimensional linear common factor model is that $\Sigma = \Lambda_2\Lambda_2' + \Psi$; however, it is also the case that unidimensional quadratic factor structures imply the covariance structure described by $\Sigma = \Lambda_2\Lambda_2' + \Psi$ (McDonald, 1967). Hence, clearly $\Sigma = \Lambda_2\Lambda_2' + \Psi$ is *not* a sufficient condition for a 2-d LCFM latent structure. The implication is that although the conditions for *nonconformity* of $f_{\underline{X}}$ to the QC (and, hence, a test to its TS) can be established, necessary and sufficient conditions for the QC (and, hence, for TS/T conformity) have not yet been determined. Instead, according to the framework in its current form, as long as sense 1 criteria for the chosen QC met, then the test analyst *acts as if* sufficient conditions (i.e., sense 2 criteria) have been met, but without a logical justification for doing so. Hence, further work establishing sufficiency conditions for the chosen QC is required if a true, evaluative test analytic framework is to be fully worked out.

Integrating a Content and/or Face Validity Component into the Framework

In the current work, a discussion of the importance of content validity (and also "face" validity) has been sidestepped. The proposed test analytic framework in its present form does not include a component bearing on a logic according to which a particular set of items may, on the basis of an analysis of item content, be justifiably claimed to be "indicators" of the attribute of interest. In other words, the framework lacks a coherent means of justifying that the items of a test constitute a representative, relevant, or otherwise appropriate set of indicators of the attribute which the test has been designed to measure. As it stands, the only criterion that the items of a test jointly measure an attribute of interest is that the joint distribution of the set of items, *whatever their content*, is describable in terms of a unidimensional test theory model which represents an appropriate paraphrase of the TS of T. Taken to its logical limit, this means that if a set of observed measures of shoe size, IQ, annual income, number of children, and preferred flavour of ice cream for a sample of respondents from a focal population is describable by the chosen QC, according to the framework, the items could be justifiably composited into a metric for some attribute, γ. However unlikely this scenario is, the point is that the current framework merely *assumes*, but provides no means of justifying, that the content of the items represents relevant features of the attribute which the test is designed to measure.

Further Development of Test Theory Models

Although the current framework does not bear directly on advances and developments in psychometric theory, it pragmatic utility is in part dependent on such advances and developments. For example, more and more complex theoretical structures can be accommodated by the framework only to the extent that there exist mathematical models into which the components of those complex theoretical structures may be mapped. In addition, the soundness of the framework is also reliant on the soundness of the particular statistical modelling procedures employed therein. Since the framework relies on inferential techniques, if a particular procedure is performing poorly, then clearly the claims born out of the test analysis may be compromised, even if the framework is strictly adhered to. Furthermore, the reliance of the framework on latent variable models, many of which of which are indeterminate, and, hence, present certain conceptual hurdles regarding scaling individuals with respect to the attribute measured by the test, may potentially weaken its pragmatic value. Perhaps psychometricians need to focus their energies on developing determinate models which will constitute better paraphrases of E=EIV than is currently provided by latent variable models.

6. AN ANALYSIS AND CRITIQUE OF CURRENT TEST ANALYTIC PRACTICES

Test Analytic Rule Violations

On the basis of the results presented in Chapter Four, I believe that it is not being overly fastidious to conclude that current test analytic practices may in large part be characterized as consisting in an unrationalized mix of test theoretic concepts and techniques, whose use is generally haphazard and ill-guided.[54] It should be noted that *unsound* practice is only identifiable in light of a clear sense of *sound* practice. Chapter Five, I maintain, defines a sound test analytic practice, and, in light of this definition, the current chapter considers the calibre of current test analytic output. Unsound practices can be usually categorized into two primary classes: fundamental logical differences and casual misapplications of test theoretic concepts and tools. Below, points A through C address the former and points D and E the latter.

[54] It must be recognized, however, that had the examined sample of studies included the work of technically skilled test developers and/or the applied test analyses of psychometricians, the situation would likely have been considerably less bleak. Hence, the critique given herein is meant to be directed primarily at applied test analysts with little or no knowledge of psychometric theory and/or statistical expertise.

A. No Antecedent Specification of TS

In Chapter Four it was noted that a fair number of the test analyses reviewed in the current work appeared to be guided by a blend of exploratory and confirmatory aims. Frequently the stated goal was to *confirm* ("assess", "identify") that a test measures a given attribute (or set of attributes) in a manner that is in keeping with expectations (e.g., based on specific findings of previous research, or, more generally, on received theory about the attribute which is purportedly measured by the test), whereas the test analysis actually carried out was exploratory in its orientation, its apparent aim being to find out what and how the test *really* measures. In terms of the test analytic framework proposed in Chapter Five, this is tantamount to trying to make *evaluative claims* about test performance, in the absence of an unambiguous, antecedently specified standard of correctness for what it means for a test to perform *satisfactorily*. Specifically, researchers would attempt to pronounce on the relevancy of results pertaining to the fit of particular models, or with regard to the "reliability", or the "validity" of T in **P**, without first 1) specifying the TS of T, 2) choosing a QC that is isomorphic to its TS, and, then, 3) demonstrating that $f_{\underline{x}}$ conforms to the chosen QC, and, hence, that T conforms to its TS in **P**.

The non-specification of TS was manifest in three common practices. First, many researchers were concerned with *determining* what is the "structure" of a test, T, in particular, with respect to the number of attributes that were being

measured by T in **P**. After examining T's structure, they would judge certain other aspects of the performance of T, such as the extent to which it demonstrated an acceptable "reliability" for the structure thus determined. Procedures such as exploratory factor analysis (EFA) or principal components analysis (PCA) were employed with the aim of "extracting" the number of factors/components "underlying" the test in order to decide how many attributes, or "constructs" are being measured by T in **P**. For example, Francis and Dugas (2004) claimed that an exploratory "Factor analysis shows that the SIBAW has a four-factor structure" (p. 405). On the basis of these results, the researchers grouped the items into four new subscales, provided reliability estimates for both the total score and each of the "derived" subscales, as well correlations between the SIBAW total score and other measures of the *positive beliefs about worry* "construct". They concluded "The SIBAW shows *good* internal consistency and test-retest reliability, as well as concurrent validity" (p. 412; emphasis added).

In order for such evaluative claims about a test's reliability, validity, or any other aspect of test performance, to be *relevant*, one must demonstrate the conformity of $f_{\underline{X}}$ to a QC which is isomorphic to the TS. Since the TS fixes what is meant by *satisfactory*, but also *unsatisfactory* performance, its specification is a necessary first step in an evaluative test analysis. Without it, a given statistical model cannot be considered a QC (sound or otherwise) of the TS, TS/T

conformity cannot be fairly judged, an optimal (model-implied) compositing rule cannot be identified, without which a composite cannot be justifiably formed, and the relevancy of its properties fairly judged. For studies in which no TS is specified (or, the TS cannot easily be deduced from the researcher's description of the test), one might ask to what, exactly, do such descriptors as "high Cronbach's α ", "good internal consistency", "good construct validity", etc., refer? To unweighted or weighted *sums* of test items? To unweighted or weighted *means* of items? To some other function of the k items of T, or of subsets of the k items? Recall that coefficient α constitutes an appropriate estimate of the reliability (or lower bound to the reliability) of the unweighted sum of a set of items. However, one is *justified* in employing the compositing rule which produces an unweighted sum only for cases in which the joint density of the set of items, $f_{\underline{X}}$, has been shown to conform to a QC which is isomorphic to TS(C,1,LI, EIV), such as a ULCF model. Likewise, the relevancy of judgements as to the "goodness" of a test's "internal consistency" or "validity" are fully contingent on $f_{\underline{X}}$ having been shown to conform to a unidimensional QC that is isomorphic to the TS. Clearly, then, the absence of an antecedently specified TS leaves the relevancy of such claims totally open to question.

Now, this is not to say that a researcher may not justifiably be engaged in truly exploratory analyses of the statistical properties of $f_{\underline{X}}$. The joint density of any set of random variates will have certain properties, and, hence, will have

some association structure. Moreover, a model can always be found which provides an adequate description of this structure. Hence, to identify $f_{\underline{X}}$ as having a particular "structure" is to state that $f_{\underline{X}}$ can be described by a given statistical model. But, this is merely to catalogue certain of the *empirical* properties of $f_{\underline{X}}$, and does not inform in any way as to what is the *theoretical structure* of the test, the latter of which is theoretically-derived, and must be specified prior to any empirical analysis, the aim of which is to assess whether a test's behaviour may be said to conform to it theoretical structure. In truly exploratory analyses, wherein there is no antecedently specified TS and, hence, no antecedently specified standard of correctness for "good", "adequate", "poor", etc. test performance, a test can be neither vindicated nor indicted, and, thus, the possibility of making non-ambiguous evaluative judgements about the test's behaviour is precluded.

Second, researchers would hypothesize a set of competing latent variable models, each of which makes particular claims about the parameters of $f_{\underline{X}}$. Of chief interest, once again, was finding the model that demonstrated the "best" fit in some sense, thereby determining what and how T measures in **P**. For instance, in his analysis of the French version of the WAIS-III, Grégoire (2004) compared two-, three-, and four-factor models, and judged that the "four-factor solution fitted the data much better than did the two- and three-factor solutions" (p. 463). He concluded that "According to these results, the two-factor model is no longer

the best way to interpret scores on the WAIS-III" and that "the validity of the four index scores based on the four factor model was supported" (pp. 471-472). O'Connor, Colder, and Hawk (2004) claimed that the "goal of [their] study was to confirm a two-factor structure" (p. 987); they tested not only a two-dimensional restricted LCFM, but also three-, four-, five-dimensional restricted LCFM's, and, after a number of modifications, chose as the "final measurement model" a "trimmed" two-factor model. Finally, they claimed that the "alpha reliability coefficient for both scales were acceptable" (p. 994).

Once again, if the aim is strictly to decide on how to best represent the structure of the joint density, $f_{\underline{X}}$, of a set of random variates, then there is nothing necessarily out of line with choosing among a set of candidate models the one that is "best" according to some (usually statistical) criterion or set of criteria. However, if the aim is to *confirm* that the "test" is in keeping with theoretical expectations, i.e., with the TS that is implied by theory, then the analyst *must* be able to demonstrate that $f_{\underline{X}}$ conforms to some QC that is isomorphic to the TS. What could it possibly mean to claim that the validity of a test was *supported*, or that the reliability was *acceptable*, in the absence of an antecendently specified standard for "test was supported", or "reliability was acceptable"? Furthermore, if the goal is truly to *replicate* a previously identified "structure" of a set of test items, or to *confirm* that a given "structure" holds for the test in some focal population, **P**, then what could be the purpose of

hypothesizing a *set* of models, at least some of which are *not* isomorphic to the implied TS of T? It is not clear how the results from the testing of the fit of a 3-factor (4-, or 5-factor, for that matter) model bear on the performance of a test with TS(C,2,LI,EIV). Contrary to a certain conventional wisdom, in an evaluative test analytic context, more is *not* better. Any model-based result for which the hypothesized model is not isomorphic to the TS will have *no relevancy* with regard to evaluative claims about the test's performance, and, in fact, is more likely to confuse than enlighten the reader.

Third, researchers would test the fit of one or more theoretically-derived models, and, in the face of empirical results indicating poor fit, would conclude that the test might measure something other than what it was designed to measure, or what received theory predicted it would measure. Rather than indicting the test for performing unsatisfactorily, researchers would typically either try to determine the *real* "structure" of T via an exploratory factor analysis, or some such procedure, or would test the fit of alternative confirmatory (i.e., restricted) models, until one was found to provide a good fit. On the basis of these latter empirical findings, *reinterpretations* of what and how T measures in **P** would be given. For example, Osman et al. (2004) contended that "Results of the CFAs showed that none of the models tested meet the preestablished criteria for use [of the Beck Depression Inventory-II]...Thus, we conducted EFAs to identify specific BDI-II factor structures for the sample data" (p. 129); ultimately these

researchers "retained" a two-factor oblique solution, computed coefficient alpha of the BDI-II total and "derived factor scales", and concluded that the "Results of the reliability analyses were *good* across the subsamples" (p. 129; emphasis added).

In the context of test evaluation, to interpret lack of model fit as a discovery about what and how T *really* measures in **P** is to engage in a practice that is, at best, misguided, and, at worst, dubious. If it is clearly implied that T was designed to measure a given construct (or set of constructs), and the stated aim is to confirm that T does, in fact, measure in the manner it was designed to measure in some focal population, **P**, then the only appropriate interpretation of the *nonconformity* of $f_{\underline{X}}$ to the chosen QC is that T does not "perform as it should perform" in **P**. To reinterpret what and how the test measures in the face of TS/T nonconformity is analogous to setting the "level of significance" (i.e., α) for a statistical test, *after* previewing the empirical results of the test, to a level that will ensure that the null hypothesis is rejected. This strategy guarantees certain "success" for a given test analysis: If $f_{\underline{X}}$ is shown to conform to the originally chosen QC, then the researcher *confirms* that the test does, indeed, perform *as it should perform*; if, on the other hand, $f_{\underline{X}}$ fails to conform to the original QC, but can be described by another model (which is *always* the case), the researcher concludes that the latter reveals important information about what the test *really*

measures. A test analyzed in this manner cannot be indicted, and, hence, evaluative claims born out of such an analysis are rendered meaningless.

However, there are a number of scenarios in which the testing of multiple models is a legitimate test analytic practice. Given that TS consists in a loose (and, hence, somewhat vague) linguistic specification of what and how the test measures, there will generally exist many sound quantitative translations of a given TS. Indeed, researchers may not have a solid rationale for choosing one model over another as the QC for the TS in question. In such cases, as long as each model under consideration represents a sound translation into quantitative terms of the components of a *single* given TS, researchers would be justified in testing the set, and choosing as the QC of the TS the model for which $F(\underline{\tilde{M}}^{*}_{\mathbf{P}}, \underline{M}_{\mathbf{P}})$ is at a minimum for a focal population, **P**. For example, for a test with theoretical structure TS(C,1,LI,EIV), the researcher may test the fit of each of three undimensional linear common factor models (ULCF models): 1) a ULCF model in which the item/common factor "loadings" and error variances are each constrained to be equal, 2) a ULCF model in which only the loadings are constrained to be equal, and 3) a ULCF model in which no constraints are imposed on either the loadings or the error variances. Then the researcher may select, on the basis of chi-square difference tests, the model which represents the "best" fit as the QC for the TS at hand.

B. TS/QC mismatches

In some studies, despite there being relatively clear indications of what is the TS of the test being analyzed, researchers assessed the fit of one or more statistical models which were not isomorphic to the TS in one or more of its components. Three commonly encountered TS/QC mismatches were:

1. Principal component model as a QC for TS(C,r,LI,EIV)

In a number of the analyses reviewed, principal components analysis was used to confirm that, in particular, the items of T measure *m* attributes: Harvey et al. (2004) employed a PCA with the aim of investigating "whether the six-factor structure of the Frost Multidimensional Perfectionism Scale could be replicated in a community-based sample" (p. 1007); Williams and Paulhus (2004) used a PCA with oblimin rotation to analyze the Self-Report Psychopathy (SRP-II) scale in order to "determine whether or not Hare's two factors could be reproduced" (p. 768); Knyazev et al. (2004) employed a PCA to analyze the Gray-Wilson Personality Questionnaire which, they claimed, "was devised to measure six animal learning paradigms upon which Gray's theory of personality is founded" (p. 1566).

If $f_{\underline{X}}$ can be described by an *r*-dimensional component model, this means that $\mathsf{rank}(\Sigma) = r$, which, in turn, means that $C(\underline{X} \mid \underline{c}) = \varnothing$, in which $\underline{c}$ is an $r \times 1$ vector of components. This feature of component models, then, constitutes a quantitative translation of E=EF, i.e., the items measure $\underline{\gamma}$ without error.

Although *in theory* there may arise tests for which the TS is TS(C,*r*,LI,EF), in practice this is never the case. Furthermore, *failure* of $f_{\underline{X}}$ to be described by an *r*-dimensional component model does not constitute grounds for indicting a test with TS(C,*r*,LI,EIV) as *not* conforming to its TS. In such cases, an *r*-dimensional linear common factor model constitutes an isomorphic QC of TS(C,*r*,LI,EIV). If $f_{\underline{X}}$ conforms to such a model, this means that $\mathrm{rank}(\Sigma - \Psi) = r$ and, therefore, that $C(\underline{X} \mid \underline{\theta}) = \Psi$, Ψ positive definite. Hence, as long as the items of a test are considered to be "imperfect" indicators of an attribute (or set of attributes), a component model would represent an unsound QC for all commonly encountered TS's.

2. Linear common factor model, or linear principal component model for TS(C,.,MI, EIV)

Despite the fact that the "R" component of the TS for most tests is, at best, loosely specified, linear factor models or linear component models have become by far the default for tests for which the item/attribute regressions might more realistically be conceptualized as generally monotone increasing (MI), as opposed to as the particular linear case of MI. For example, Cepeda-Benito and Reig-Ferrer (2004) assessed the fit of a 2-dimensional LCFM to the items of a test for which, they claimed, "high numbers [indicate] greater level of agreement" (p. 403); with regard to the 3-point response scales for the items of the Quest Religion Scale, Shaw and Joseph (2004) stated that "A higher score indicates

greater quest religion" (p. 1427); Müller, Bühner, and Ellgring (2004) asserted that "The total score [of the Toronto Alexithymia Scale] ranges from 20 to 100 points with high scores indicating high alexithymia" (p. 376). As was described in (1), the lack of conformity of $f_{\underline{X}}$ to a model in which the item/synthetic variates are modelled as linear (increasing) does not constitute grounds for indicting a test with TS(C,.,MI,EIV). Hence, unless there exist compelling theoretical and/or empirical grounds for restricting the item/attribute regressions to be linear, researchers should choose as a QC for the TS a model from a larger class of latent variable models in which the item/synthetic variate regressions are monotone increasing (see Holland and Rosenbaum, 1986, for a discussion of the class of unidimensional monotone latent variable (UMLV) models).

3. Unsound QC's employed for TS's in which D>1

In many of the articles reviewed in the present work it was contended that the test to be analyzed could be conceptualized as a measure of more than one attribute, or more than one facet of a higher order attribute: "the latent structure [of the LASSI] is likely multidimensional and complex, resulting in a lack of simple structure" (Stevens and Tallent-Runnels, 2004, pp. 334-335); "Although it is claimed that the Big Five dimensions...represent the highest level in the hierarchical structure of personality, there is consistent evidence that they are not independent and that two higher order factors underlie them" (Blackburn, Renwick, Donnelly, and Logan, 2004, p. 957); "the proposed 5-factor structure

was tested using confirmatory factor analysis" (Müller et al., 2004, p. 373). However, there was little if any consistency in the approaches taken by researchers in evaluating the performance of such "multidimensional" tests. In addition to a common proclivity towards confounding exploratory and confirmatory aims (as outlined in point A above), many researchers employed models which did not constitute isomorphic QC's of the TS's at hand. For example, in a fair number of studies, the conformity of $f_{\underline{X}}$ to a particular *r*-dimensional unrestricted linear factor model was assessed for tests with TS(C,*r*,LI,EIV), i.e., models in which no association structure for the latent variates (proxies to the attributes purported to be measured by T) was specified.

It is argued here that there exist two legitimate strategies from which researchers may choose in order to justify claims about TS/T conformity for tests with TS(.,*r*,.,.): First, treat T as the union of *r* mutually disjoint sets of items, with each set conceptualized as a "subtest" (or "subscale") of a single distinct attribute (or "facet" of a higher order attribute). Here, the TS will not specify a covariance structure among the attributes measured by T. Then, apply the framework described in Chapter Five to each subtest individually, i.e., for the l^{th} subtest, l = 1, 2, …, *r*, with TS(.,1,.,.), an appropriate *unidimensional* test theory model will chosen as the QC for TS(.,1,.,.), the conformity of $f_{\underline{X}_l}$ to the chosen QC will be assessed, and so on. For this approach, the "performance of T" will be judged in reference to the individual performances of each subtest of T.

Second, one could broaden the admittedly narrow scope of the TS as presented in Chapter Five to include components which imply additional specific empirical requirements with regard to $f_{\underline{\theta}}$. For example, one might add to the specification of the TS for a test with TS(C,4,LI,EIV) that the four attributes measured by T are positively correlated. A sound QC of this TS is a 4-dimensional LCFM, in which the off-diagonal elements of $C(\underline{\theta}) = \Phi$ are constrained to be positive.[55] Then, to justify claims that this particular test is performing adequately the researcher must, minimally, be able to demonstrate that $f_{\underline{X}}$ can be described by a 4-dimensional linear factor model with positively correlated factors. If $f_{\underline{X}}$ is so described, then, at least in theory, optimal model-implied compositing rules may be derived, degree of precision estimated for the resulting composites, and, if appropriate, those composites may be entered into investigations of the nomothetic span of the attributes measured by T. It might be noted, however, that this second option, although feasible, is included here mainly for completeness, as, in practice, researchers are seldom explicit with regard to the more elementary components of the TS that were described in Chapter Five. Typically, an examination of the relations between attributes is reserved as part of the *external* component of test analysis (i.e., step 6), and is *contingent* on the successful completion of the internal components of an analysis (i.e., steps 1-5).

[55] Notwithstanding identification issues which might arise.

C. Sequence Violations

Of the studies reviewed in the current work, researchers rarely specified a clear and unambiguous statement of the TS of the test to be analyzed. For some of the studies, the TS could be deduced; however, for many, it remained completely unclear. Although this does not constitute a violation *per se* of the sequence of "operations" outlined in the test analytic framework presented in the current work, it deserves mention given the paramount importance of TS specification as a *necessary first step* in any analysis whose aim is to evaluate test performance. As was indicated above, in the absence of a clearly articulated, unambiguous TS, TS/T conformity cannot be fairly judged, and, hence, all other claims about the performance of T will be meaningless.

Notwithstanding the inevitable failures of test analyses for which no TS has been specified, there did exist a number of other violations of the sequential order presented in Chapter Five, most notably that 1) composites were produced, and their precisions estimated, *prior to* demonstrating conformity of test performance to some sound QC of TS(.,1,.,.), and 2) entering a test, or a composite, into construct validation studies in advance of having demonstrated that the composite possesses an adequate degree of precision. In particular, it was not uncommon for researchers to mechanically produce the unweighted sum of the "items" of T, provide estimates of its reliability, and investigate T's construct validity, without first: i) demonstrating that $f_{\underline{X}}$ could be adequately

described by a sound QC of TS(.,1, .,.), and, hence, that the X_j could justifiably be taken to be measures of a single attribute of interest, γ; and, having done that, ii) justifying that the particular choice of $\phi = \sum_{j=1}^{k} X_j$ represents, in the sense implied by the chosen QC, an "optimal" measure of γ, the attribute purportedly measured by T.

For example, van der Ploeg et al. (2004), on the basis of the "original structure" of the Impact of Event Scale (IES), provided estimates of coefficient alpha for the previously identified subscales and the total score of the Dutch version of the IES for three separate samples, and *then* tested with confirmatory factor analyses the fits of both a single- and two-factor model on the total sample; Williams and Paulhus (2004) estimated the "alpha reliability" of the total score for the Self-Report Psychopathy scale (SRP-II), and *then* "factored" the 60 items in order to "uncover the factor structure" of the SRP-II; Motl, Dishman, Saunders, Dowda, and Pate (2004) used confirmatory factor analysis to evaluate the "factorial and construct validity" of the Social Provisions Scale for physical activity for two distinct populations; without analyzing measurement precision for either sample; they concluded that their results supported the factorial and construct validity of the test.

In addition to representing instances of TS/QC mismatch, examples such as these constitute particular violations of the correct order in which the different

components of test analysis must be carried out if it is to foster meaningful claims about the test analyzed. As was described in detail in Chapter Five, technically, it is meaningless to speak of the reliability possessed by a given test, T, in population **P**. Rather, composites, $\phi = f(X_j)$, and not tests, may be said to be more or less precise. However, it would make little sense to estimate the precision of a given composite, ϕ^*, without first justifying the formation of that composite as an (error-laden) measure of the *single* attribute, γ, which T is purported to measure. And, one can justify the production of a given composite only if there can be derived a model-implied compositing rule that predicts (estimates) a random variate that may be taken to be a proxy for γ. However, such an optimal, model-implied scoring rule is derivable only when it has been demonstrated that $f_{\underline{X}}$ may be described in terms of a QC that has been chosen as a sound paraphrase the TS(.,1,.,.) of T in **P**. Furthermore, investigations into the "validity of T", are predicated on the notion that ϕ^*, the optimal, model-implied, composite, has been shown to consist in an adequately precise measure of γ, the attribute which T has been designed to measure. Hence, any alternate ordering of the steps described in the test analytic framework proposed in the current work, will result in quantities whose relevance to the evaluative aim at hand is questionable at best.

D. Employing Inappropriate Standards of Correctness for Judging Test Performance

1. Judging the performance of a test on the basis of results from past studies

A substantial proportion of the articles reviewed in the current work included some reference to the results of previous test analyses (including those published in test manuals), presumably as an indication of general performance of a test. Of the articles examined, almost 50% cited previous findings pertaining to the reliability and/or the validity of at least one of the tests employed. Some studies even relied solely on previous findings for pronouncing on the reliability and validity of a test which was to be employed in the study. Although *replication* of findings does, in a certain sense, speak to the overall utility of a given test as a measure of a particular attribute, generally, the results from previous analyses have limited or no relevance to a subsequent analysis, as any judgement of "the performance of T" is strictly in reference the responding of individuals *in focal population* ***P*** to the *k* items of T. Since TS/T conformity in **P** is required in order to justify the compositing of items, previous research findings, which have no bearing on the issue of the conformity of T to its TS *in a distinct focal population* ***P*** *at a latter point in time*, are not relevant to the current aim of passing judgement on the performance of the test for population **P**.

2. Post-hoc interpretations of results indicating poor performance

In a number of studies, researchers attempted to mitigate the impact of low reliability estimates by appealing to, for example, the existence of higher

reliability estimates in previous studies (cf. Egan et al, 2004; Zhang, 2004), the small number of items on which the reliability estimates were based (cf. Francis and Jackson, 2004), or low reliability being compensated for by good validity (cf. Vittengl et al., 2004). This is, once again, an attempt at justifying evaluative claims about a test on the basis of criteria that exist outside of the test analytic rules on which the veracity of such claims are dependent. As argued above, that TS/T conformity has been established for a particular population $\mathbf{P}_1$, does not necessarily mean that it will be so for a different population $\mathbf{P}_2$ (or, for that matter, that the TS's for two distinct populations will be the same). Although it is well known that, all else being equal, reliability will increase as k increases, one should not lose sight of the fact that, in a given test analysis, the performance of a test, with *fixed k*, is being judged, and suggestions about what the reliability *would be* for a test with $>k$ items are inconsequential to the analysis at hand. Finally, since the relevancy of validity "evidence" is *contingent on* the chosen composite of the test items having been shown to possess adequate precision, there is no sense in appealing to "good" validity in order to offset poor reliability.

3. Employing incorrect criteria for unidimensionality

Despite the publication of a number of well-known indictments of the practice of employing coefficient alpha as an "index" of unidimensionality (most notably Green, Lissitz, and Mulaik, 1977, and McDonald, 1981), some researchers continue to interpret the magnitudes of estimates of coefficient alpha as

indicators of the undimensionality of tests. For example, Moneta and Yip (2004) produced coefficient alpha estimates for the subscale "scores" (presumably unweighted sums of items) of two tests, concluding that

> On the whole, the reliability estimates in the sample are satisfactory but slightly lower than those estimated using the original English version of the scales...In particular, the score of the last two [subscales]...are less *internally consistent* than those in the first three [subscales of the NFCS]. (p. 537; emphasis added)

Without recapitulating in full the arguments given by critics of this practice, it can be shown that coefficient alpha,

$$\alpha = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum_{j=1}^{k}\sigma_j^2}{\sigma_c^2}\right), \tag{6.1}$$

is equal to

$$k^2\frac{\bar{\sigma}_{ij}}{\sigma_c^2}, \tag{6.2}$$

in which $\bar{\sigma}_{ij} = \frac{1}{k(k-1)}\sum_{i\neq j}\sum\sigma_{ij}$ and σ_c^2 is the variance of the unweighted sum of the k random variates.

For a set of k random variates that are described by an r-dimensional LCFM, the $k\times k$ population covariance matrix of the variates is equal to $\Lambda_r\Lambda_r' + \Psi$, Ψ diagonal and positive definite. Consequently, the σ_{ij}, and so too $\bar{\sigma}_{ij}$, are strictly functions of the item/common factor loadings, λ_{jm}, $j = 1, 2, \ldots, k$, $m = 1, 2,$

..., *r*. Hence, the existence of a large average interitem covariance can come about in a number of different ways, and is obviously independent of the dimensionality of the latent space. For example, large values of $\bar{\sigma}_{ij}$ would be produced when $f_{\underline{X}}$ is described by an *r*-dimensional LCFM with Λ_r constrained to have a simple structure, and in which the λ_{jm} are large and positive. Likewise, coefficient alpha will be small when *k* is small and $f_{\underline{X}}$ conforms to a *unidimensional* LCFM in which all the loadings, λ_j, are relatively small in magnitude. Clearly, then, the dimensionality of the latent space cannot be distinguished by the magnitude of coefficient alpha, and, hence, coefficient alpha cannot reasonably be employed as an indicator of unidimensionality.[56] Rather, one is justified in employing coefficient alpha as an estimate of the lower bound to reliability of a composite of the form $\phi = \sum_{j=1}^{k} X_j$, if and only if, $f_{\underline{X}}$ has been shown to *conform to* a unidimensional LCFM.

E. Other Misuses and/or Misunderstanding of Test Analytic Concepts

The wanton state of current test analytic affairs should come as no surprise given some common confusions that, at least on the surface of things,

[56] For exactly the same reasons, the "percentage of variance explained" by the , "first factor" cannot be reasonably employed as a indicator of unidimensionality.

continue to plague researchers' understanding of certain fundamental test theoretic concepts. A number of common misconceptions are described below.

1. Misunderstanding models

Although the mathematical sophistication of researchers varies considerably, it is clear that many employ statistical models although having only a tenuous grasp of statistical models and their proper place in test analysis. Doubtless, many of the above mentioned violations speak clearly to certain misapplications of statistical models that occur with some frequency in applied test analyses, e.g., employing a statistical model which is not isomorphic to the (implied) components of the theoretical structure of a test. However, matters are apparently worse, as is evidenced by the fact that one encounters claims about the reliability of *factors,* multidimensional *constructs,* factors *assessing,* internally consistent *models,* items loading onto *subscales,* etc. In a test analytic context, because TS's are linguistic specifications, and do not carry with them particular empirical consequences for $f_{\underline{X}}$, the components of a given TS may be mapped into a set of quantitative requirements for $f_{\underline{X}}$. The employment of statistical modelling represents a sophisticated approach to such a mapping. However, if researchers cannot appreciate the basic distinctions between a subscale and a factor, between a model-specific synthetic variate and the attribute for which the former is a proxy, between item composites and the statistical models which play

a role in justifying their formation, then any elegance that might have been lent to test analysis by the use of such sophisticated techniques will be lost.

2. Reliability

In addition to addressing reliability out of sequence, improperly employing coefficient alpha as an indicator of unidimensionality, or applying the concept of reliability to latent variates (entities which cannot sensibly be said to possess *precision of measurement*), evidence of a number of other common misconceptions about reliability persist. First, test-retest coefficients continue to be employed as either reliability *or* stability estimates, a clear indication that researchers do not grasp that test-retest coefficients *conflate* reliability and stability. Second, the practice of employing coefficient alpha as the "default" estimate of reliability persists, despite the fact that it is only appropriate under certain conditions. Third, there exists a common practice of routinely producing estimates of the reliability of both subscale composites and total test composites; however, for the reasons outlined both in Chapter Five and in the current chapter, one may justify the compositing of a set of items only when it has been demonstrated that the items measure but a single attribute. Given this, there is little sense to the production of a total score reliability estimate for a test with implied TS(.,D>1,.,.). This practice is but one example of the "more is better" dictum that appears to motivate many of the unsound practices that characterize much of applied test analysis.

7. CONCLUDING REMARKS

The concept of "practice" has many senses, the most commonly used of which are 1) the action of doing something with aim of proficiency or perfection with regard to some behaviour for which the action is intended, and 2) an action which is performed habitually or customarily in a certain setting or under certain circumstances by an individual or group of individuals. Within all social groups there exist both individual and shared practices.[57] However, not all practices are *rule-guided,* in the sense of containing *standards of correctness* that fix which behaviours are considered acceptable or preferable in a given situation. Rule-guided practices, as opposed to practices in general, are motivated by particular *prescriptions* or *stipulations* for what behaviour in a particular situation *should* or *ought to* look like.

The practices of science are encoded in *scientific methods.* Although the methods of science are legion, diverse, and often specific to the particular branch of study, they are shared in the sense that a given scientist does not have complete and total freedom to study the phenomena of interest to him in any

[57] Although the bifurcation is somewhat artificial, as many practices fall into both categories.

fashion he so desires. Rather, he is bound, at least to some extent, by the conventions set down by his peers (and scientists in general) for how he is to approach everything from the articulation of theoretical propositions, to the methods he employs in making observations and summarizing findings of empirical investigations.

However, many of the practices in which the scientist engages are ill-defined, providing at best only cursory guidelines for "doing" science. Although there may exist within some scientific practices a certain degree of tolerance of contraventions of the "usual" way of going about business, others require relatively inflexible and explicit standards, or rules, for what needs to be involved in order to meet a particular aim.

In the present work, the distinction between test theory and test analytic frameworks has been emphasized, with the former taken to constitute quantitative "pictures" of the relationship between responding to test items and the attribute for which the items are taken to be measures. Conversely, a test analytic framework has been defined here as a set of rules that stipulates how the theoretical and technical components of test theory are to be employed in passing judgement on the performance of a test.

It is in the spirit of the latter definition that the framework outlined in Chapter Five was proposed, and, furthermore, can be contrasted with the state of affairs which characterizes current test analytic practice: Whereas the former

consists in a well-defined, logically coherent set of *rule-guided* practices for pronouncing on test *performance*, the latter amounts merely to a collection of practices pertaining to the assessment of various features of test *behaviour*, practices which, at least in appearance, are not embedded in a logical structure which stipulates *how* test analyses ought to be done.

The point is that it requires more than simply the existence of practices to judge fairly whether a test is performing as it should perform, or according to expectations. Those practices must also reside within a logically coherent framework, wherein particular rules serve as standards for how a test is to be analyzed if the results from particular empirical analyses are to bear on judgements about the quality of the test analyzed. Without rules which fix what a legitimate test analysis "looks like", one can never meaningfully interpret how mere test *behaviour* bears upon the legitimacy of claims regarding test *performance*.

Now, it is not being suggested that the framework proposed here is the *only* possible test analytic framework, i.e., the only set of logically coherent rules for pronouncing on the performance of a test. Indeed, there could be, conceivably, many such frameworks, each of which would consist in a set of rules that stipulates what it means to have "good", "adequate", and, also "poor", or "inadequate" test performance, and to which the test analyst could refer in order to justify her claims regarding the performance of a given test.

However, if the practices of researchers whose test analyses were reviewed in the present work are indicative of the general state of applied test analytic affairs, it would appear that, if such a framework exists, it is not employed with any frequency by researchers conducting test analyses. Furthermore, it has been argued that the APA *Standards for Educational and Psychological Testing* (1999), the only would-be competitor to the framework proposed here, fails (for reasons outlined in Chapter Five) to constitute a test analytic framework proper. The point is that, currently, the practices that characterize applied test analysis are *not* embedded in a logically sound framework for passing judgement on the quality of tests, and, hence, *some* such framework is required if test analysis is to be rescued from its current haphazard state. The framework proposed here will accommodate the requirements for a sound test analytic framework for general test analyses, but also has the added advantage of being adaptable for more complex test analytic questions.

REFERENCES

Ableson, A.R. (1911). The measurement of mental ability of 'baskward' children. *British Journal of Psychology, 4,* 268-314.

Alexopoulos, D.S. & Kalaitzidis, I. (2004). Psychometric properties of Eysenck Personality Questionnaire-Revised (EPQ-R) Short Scale in Greece. *Personality and Individual Differences, 37,* 1205-1220.

American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin, 51(2),* 1-38.

American Psychological Association. (1999). *Standards for educational and psychological testing.* Washington, D.C.: American Educational Research Association.

Barrett, D.W., Wosinska, W., Butner, J., Petrova, P., Gornik-Durose, M, & Cialdini, R. (2004). Individual differences in the motivation to comply across cultures: The impact of social obligation. *Personality and Individual Differences, 37,* 19-31.

Bartholomew, D.J. (1981). Posterior analysis of the factor model. *British Journal of Mathematical and Statistical Psychology, 34,* 93-99.

Bartholomew, D.J., & Knott, M. (1999). *Latent variable models and factor analysis.* London: Arnold.

Beck, J.G., Coffey, S.F., Palyo, S.A., Gudmundsdottir, B., Miller, L.M., & Colder, C.R. (2004). Psychometric properties of the Posttraumatic Cognitions Inventory (PTCI): A replication with motor vehicle accident survivors. *Psychological Assessment, 16(3),* 289-298.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. Lord & M. Novick, *Statistical theories of mental test scores.* (pp. 397-479). Reading: Addison-Wesley.

Bishop, D.I. & Hertenstein, M.J. (2004). A confirmatory factor analysis of the Structure of Temperament Questionnaire. *Educational and Psychological Measurement, 64(6),* 1019-1029.

Blackburn, R., Renwick, S.J.D., Donnelly, J.P., & Logan, C. (2004). Big five or big two? Superordinate factors in the NEO Five Factor Inventory and the Antisocial Personality Questionnaire. *Personality and Individual Differences, 37,* 957-970.

Blair, R.J.R., Mitchell, D.G.V., Peschardt, K.S., Colledge, E., Leonard, R.A., Shine, J.H., Murray, L.K., & Perrett, D.I. (2004). Reduced sensitivity to others' fearful expressions in psychopathic individuals. *Personality and Individual Differences, 37,* 1111-1122.

Blinkhorn, S.F. (1997). Past imperfect, future conditional; Fifty years of test theory. *British Journal of Mathematical and Statistical Psychology, 50 (2),* 175-186.

Blumentritt, T.L. & van Voorhis, C.R. (2004). The Million Adolescent Clinical Inventory: Is it valid and reliable for Mexican American Youth? *Journal of Personality Assessment, 83(1),* 64-74.

Bogels, S.M. & van Melick, M. (2004). The relationship between child-report, parent self-report, and partner report of perceived parental rearing behaviours and anxiety in children and parents. *Personality and Individual Differences, 37,* 1583-1596.

Bolt, D.M., Hare, R.D., Vitale, J.E., & Newman, J.P. (2004). A multigroup item response theory analysis of the Psychopathy Checklist Revised. *Psychological Assessment, 16(2),* 155-168.

Briggs, S.R. & Cheek, J.M. (1988). On the nature of self-monitoring: Problems with assessment, problems with validity. *Journal of Personality and Social Psychology, 54(4),* 663-678.

Briggs, S.R., Cheek, J.M., & Buss, A.H. (1980). An analysis of the Self-Monitoring Scale. *Journal of Personality and Social Psychology, 38(4),* 679-686.

Brogden, H. (1946). Variation in test validity with variation in the distribution of item difficulties. *Psychometrika, 11,* 197-214.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3,* 296-322.

Brown, W. & Thompson, G.H. (1940). *The Essentials of Mental Measurement.* Cambridge: Cambridge University Press.

Campbell, D.T. & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105.

Campbell-Sills, L. Liverant, G.I., & Brown, T.A. (2004). Psychometric evaluation of the Behavioral Inhibition/Behavioral Activation Scales in a large sample of outpatients with anxiety and mood disorders. *Psychological Assessment, 16(3),* 244-254.

Carroll, J.B. (1950). Problems in the factor analysis of tests varying difficulty. *American Psychologist, 5,* 369.

Cepeda-Benito, A. & Reig-Ferrer, A. (2004). Development of a brief Questionnaire of smoking Urges—Spanish. *Psychological Assessment, 16(4),* 402-407.

Cole, J.C., Rabin, A.S., Smith, T.L., & Kaufman, A.S. (2004). Development and validation of a Rasch-derived CES-D Short Form. *Psychological Assessment, 16(4),* 360-372.

Connor, K.R., Zhong, Y, & Duberstein, P.R. (2004). NEO-PI-R Neuroticism scores in substance-dependent outpatients: internal consistency and self-partner agreement. *Journal of Personality Assessment, 83(1),* 75-77.

Cooke, D.J., Hart, S.D., Michie, C. (2004). Cross-national differences in the assessment of psychopathy: Do they reflect variations in raters' perceptions of symptoms? *Psychological Assessment, 16(3),* 335-339.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297-334.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: John Wiley & Sons.*

Cronbach, L.J. & Meehl, P.E. (1955). Construct validity and psychological tests. *Psychological Bulletin, 52,* 281-302.

Cronbach, L.J. & Meehl, P.E. (1967). Construct validity and psychological tests. In D.N. Jackson & S. Messick (Eds.), *Problems in Human Assessment.* (pp. 57-77). New York: McGraw-Hill.

Currie, M.R., Cunningham, E.G., & Findlay, B.M. (2004). The short internalized homonegativity scale: Examination of the factorial structure of a new measure of internalized homophobia. *Educational and Psychological Measurement, 64(6),* 1053-1067.

Currie, S.R., el-Guebaly, N., Coulson, R., Hodgins, D., & Mansley, C. (2004). Factor validation of the Addiction Severity Index scale structure in persons with concurrent disorders. *Psychological Assessment, 16(3),* 326-329.

Davis, M.H., Capobianco, S, & Kraus, L.A. (2004). Measuring conflict-related behaviours: Reliability and validity evidence regarding the Conflict Dynamics Profile. *Educational and Psychological Measurement, 64(4),* 707-731.

del Barrio, V., Aluja, A., & Spielberger, C. (2004). Anger assessment with the STAXI-CA: Psychometric properties of a new instrument for children and adolescents. *Personality and Individual Differences, 37,* 227-244.

DuHamel, K.N., Ostroff, J., Ashman, T. Winkel, G. Mundy, E.A., Keane, T.M., Morasco, B.J., Vickberg, S.M.J., Hurley, K., Burkhalter, J., Chhabra, R., Scigliano, E., Papadopoulos, E., Moskowitz, C., & Redd, W. (2004). Construct validity and the Posttraumatic Stress Disorder Checklist in cancer survivors; Analyses based on two samples. *Psychological Assessment, 16(3),* 255-266.

Egan, E., Kroll, J., Carey, K., Johnson, M., & Erickson, P. (2004). Eysenck Personality Scales and religiousity in a US outpatient sample. *Personality and Individual Differences, 37,* 1023-1031.

Embretson, S.E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93(1),* 179-197.

Embretson, S.E. & Reise, S.P. (2000). *Item response theory for psychologists.* Mahwah, N.J.: Lawrence Erlbaum Associates.

Finley, G.E. & Schwartz, S.J. (2004). The Father Involvement and Nurturant Fathering Scales: Retrospective measures for adolescent and adult children. *Educational and Psychological Measurement, 64(1),* 143-164.

Finney, S.J., Pieper, S.L., & Barron, K.E. (2004). Examining the psychometric properties of the Achievement Goal Questionnaire in a general academic context. *Educational and Psychological Measurement, 64(2),* 365-382.

Fiorentino, L. & Howe, N. (2004). Language competence, narrative ability, and school readiness in low-income preschool children. *Canadian Journal of Behavioural Science, 36(4),* 280-294.

Fleisher, E. & Baize, H.R. (1982). Self monitoring: A theoretical critique. *Paper presented at the annual convention of the American Psychological Association, Washington, D.C.*

Fletcher, G.J.O., Danilovics, P., Fernandez, G, Perterson, D, & Reeder, G.D. (1986). Attributional complexity: An individual differences measure. *Journal of Personality and Social Psychology, 51(4),* 875-884).

Francis, K. & Dugas, M.J. (2004). Assessing positive beliefs about worry: Validation of a structured interview. *Personality and Individual Differences, 37,* 405-415.

Francis, L.J. & Jackson, C.J. (2004). Which version of the Eysenck Personality Profiler is best? 6-, 12- or 20-items per scale. *Personality and Individual Differences, 37,* 1659-1666.

Gangestad, S. & Synder, M. (1985). "To carve nature at its joints": On the existence of discrete classes in personality. *Psychological Review, 92,* 317-340.

Gagne, F.M., Lyndon, J.E., & Bartz, J.A. (2003). Effects of mindset on the predictive validity of relationship constructs. *Canadian Journal of Behavioural Science, 35(4),* 292-304.

Ghiselli, E.E. (1964). *Theory of psychological measurement.* New York: McGraw-Hill.

Goulding, A. (2004). Schizotypy models in relation to subjective health and paranormal beliefs and experiences. *Personality and Individual Differences, 37,* 157-167.

Grano, N., Virtanen, M, Vahtera, J., Elovainio, M. & Kivimaki, M. (2004). Impulsivitiy as a predictor of smoking and alcohol consumption. *Personality and Individual Differences, 37,* 1693-1700.

Green, S.B., Lissitz, R.W., & Mulaik, S.A. (1977). Limitations of coefficient alpha as an index of unidimensionality. *Educational and Psychological Measurement, 37,* 827-838.

Grégoire, J. (2004). Factor structure of the French version of the Wechsler Adult Intelligence Scale-III. *Educational and Psychological Measurement, 64(3),* 463-474.

Guilford, J.P. (1936). *Psychometric methods.* New York: McGraw-Hill.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10,* 255-282.

Guttman, L. (1950). Relation of scalogram analysis to other techniques. In S.A. Stouffer, L. Guttman, E.A Suchman, P.F. Lazarsfeld, S.A. Star, & J.A. Clausen (Eds.), *Measurement and prediction.* (pp. 172-212). Princeton: Princeton University Press.

Guttman, L. (1955). The determinacy of factor score matrices with implications for five other basic problems of common-factor theory. *British Journal of Statistical Psychology, 8,* 17-24.

Hambleton, R.K & Cook, L.L. (1977). Latent trait models and their use in the analysis of educational test data. *Journal of Educational Measurement, 14(2),* 75-96.

Hambleton, R.K., Swaminathan, H., Cook, L.L., Eignor, D.R., & Gifford, J.A. (1978). Developments in latent trait theory: Models, technical issues and applications. *Review of Educational Research, 48,* 467-510.

Harvey, B. Pallant, J., & Harvey, D. (2004). An evaluation of the factor structure of the Frost Multidimensional Perfectionism Scale. *Educational and Psychological Measurement, 64(6),* 1007-1018.

Hattie, J.A. (1981). *Decision criteria for determining unidimensionality.* Unpublished doctoral dissertation, University of Toronto.

Hattie, J.A. (1984). An empirical study of various indices for determining undimensionality. *Multivariate Behavioral Research, 19,* 49-78.

Hattie, J.A. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9(2),* 139-164.

Heise, D.R. & Bohrnstedt, G.W. (1971). Validity, invalidity and reliability. In E.F. Borgatta & G.W. Bohrnstedt (Eds.), *Sociological Methodology.* (pp. 3-27). San Francisco: Jossey-Bass.

Hewitt, A.K, Foxcroft, D.R., MacDonald, J. (2004). Multitrait-multimethod confirmatory factor analysis of the Attribution Style Questionnaire. *Personality and Individual Differences, 37,* 1483-1491.

Hill, C.D., Neumann, C.S. & Rogers, R. (2004). Confirmatory factor analysis of the Psychopathy Checklist: Screening version in offenders with axis I disorders. *Psychological Assessment, 16(1),* 90-95.

Hill, R.W., Huelsman, T.J., Furr, R.M., Kibler, J., Vicente, B.B., & Kennedy, C. (2004). A new measure of perfectionism: The Perfectionism Inventory. *Journal of Personality Assessment, 82(1),* 80-91.

Holland, P.W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika, 55(4),* 577-601.

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

Hoyle, R.H. & Lennox, R.D. (1991). Latent structure self-monitoring. *Multivariate Behavioral Research, 26(3),* 511-540.

Jay, M. & John, O.P. (2004). A Depressive Symptom Scale for the California Psychological Inventory: Construct validation of the CPI-D. *Psychological Assessment, 16(3),* 299-309.

Jeyakumar, S.L.E., Warriner, E.M., Raval, V.V., & Ahmad, S.A. (2004). Balancing the need for reliability and time efficiency: Short forms of the Wechsler Adult Intelligence Scale-III. *Educational and Psychological Measurement, 64(1),* 71-87.

Joreskog, K.G. (1966). Testing a simple structure hypothesis in factor analysis. *Psychometrika, 31(2),* 165-178.

Joreskog, K.G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34(2),* 183-202.

Joreskog, K.G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36(2),* 109-134.

Karademas, E.C. & Kalantzi-Azizi, A. (2004). The stress process, self-efficacy expectations, and psychological health. *Personality and Individual Differences, 37(5),* 1033-1043.

Kane, M.T. (2001). An argument-based approach to validity. *Psychological Bulletin, 112(3),* 527-535.

Kelley, T.L. (1916). A simplified method of using scaled data for purposes of testing. *School and Society, 4,* 71-75.

Kelley, T.L. (1921). The reliability of test scores. *Journal of Educational Research, 3(5),* 370-379.

Kelley, T.L. (1923). *Statistical Method.* New York: Macmillan.

Kelley, T.L. (1924). Note on the reliability of a test: A reply to Dr. Crum's criticism. *The Journal of Educational Psychology, 14(4),* 193-204.

Keogh, E. (2004). Investigating invariance in the factorial structure of the Anxiety Sensitivity Index across adult men and women. *Journal of Personality Assessment, 83(2),* 153-160.

Knyazev, G.G., Slobodskaya, H.R., & Wilson, G.D. (2004). Comparison of the construct validity of the Gray-Wilson Personality Questionnaire and the BIS/BAS scales. *Personality and Individual Differences, 37,* 1565-1582.

Kohn, P.M, O'Brien-Wood, C, Pickering, D.I., & Decicco, T. (2003). The Personal Functioning Inventory: A reliable and valid measure of adaptiveness in coping. *Canadian Journal of Behavioural Science, 35(2),* 111-123.

Krantz, D.H. (1991). From indices to mappings: The representational approach to measurement. In D. Brown & J. Smith (Eds.), *Frontiers of mathematical psychology: Essays in honor of Clyde Coombs.* (pp. 1-52). New York: Springer-Verlag.

Krueger, R.F., Nichol, P.E., Hicks, B.M., Markson, K.E., Patruck, C.J., Iacono, W.G., & McGue, M. (2004). Using latent trait modelling to conceptualize an alcohol problems continuum. *Psychological Assessment, 16(2),* 107-119.

Kuder, G.F. & Richardson, M.W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2(3),* 151-160.

Lawley, D.N. (1943). On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh, 61,* 273-287.

Lawley, D.N. (1944). The factorial analysis of multiple item tests. *Proceedings of the Royal Society of Edinburgh, 62-A,* 74-82.

Lazarsfeld, P.F. (1950). The logical and mathematical foundation of latent structure analysis. In S.A. Stouffer, L. Guttman, E.A Suchman, P.F. Lazarsfeld, S.A. Star, & J.A. Clausen (Eds.), *Measurement and prediction.* (pp. 362-412). Princeton: Princeton University Press.

Leach, M. & Lark, R. (2004). Does spirituality add to personality in the study of trait forgiveness? *Personality and Individual Differences, 37,* 147-156.

Lennox, R. & Wolfe, R. (1984). Revision of the Self-Monitoring Scale. *Journal of Personality Psychology, 46,* 1349-1364.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3,* 635-694.

Loevinger, J. (1967). Objective tests as instruments of psychological theory. In D.N. Jackson & S. Messick (Eds.), *Problems in Human Assessment.* (pp. 78-123). New York: McGraw-Hill.

Lord, F.M. (1952). A theory of test scores. *Psychometric Monographs,* No. 7.

Lord, F.M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13,* 517-549.

Lord, F.M. (1959). An approach to mental test theory. *Psychometrika, 24,* 283-302.

Lord, F.M. & Novick, M.R. (1968). Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley.

Lowe, P.A., Reynolds, C.R. (2004). Psychometric analyses of the Adult Manifest Anxiety Scale-Adult Version among young and middle-aged adults. *Educational and Psychological Measurement, 64(4),* 661-681.

Lumsden, J. (1976). Test theory. *Annual Review of Psychology, 27,* 251-280.

Maller, S.J. & French, B.F. (2004). Universal nonverbal intelligence test factor invariance across deaf and standardized samples. *Educational and Psychological Measurement, 64(4),* 647-660.

Mantler, J., Schellenberg, E.G., & Page, J.S. (2003). Attributions for serious illness: Are controllability, responsibility, and blame different constructs? *Canadian Journal of Behavioural Science, 35(2), 142-152.*

Maraun, M.D., Jackson, J.S.H., Luccock, C.R., Belfer, S.E., & Chrisjohn, R.D. (1998). Ca and SPOD for the analysis of test comprised of binary items. *Educational and Psychological Measurement, 58(6),* 916-928.

Maraun, M.D., Slaney, K., & Goddyn, L. (2003). An analysis of Meehl's MAXCOV-HITMAX procedure for the case of dichotomous indicators. *Multivariate Behavioral Research, 38(1),* 81-112.

Marsh, H.W. Parada, R.H. & Ayotte, V. (2004). A multidimensional perspective of relations between self-concept (Self Description Questionnaire II) and adolescent mental health (Youth Self Report). *Psychological Assessment, 16(1),* 27-41.

McDonald, R.P. (1967). Factor interaction in non-linear factor analysis. *British Journal of Mathematical and Statistical Psychology, 20(2),* 205-215.

McDonald, R.P. (1981). The dimensionality of tests and items. *The British Journal of Mathematical and Statistical Psychology, 34(1),* 100-117.

McDonald, R.P. (1999). *Test theory: A unified treatment.* Mahwah, N.J.: Lawrence Erlbaum Associates.

McDonald, R.P. & Burr, E.J. (1967). A comparison of four methods of constructing factor scores. *Psychometrika, 34(2),* 381-401.

Messick, S. (1980). Test Validity and the ethics of assessment. *American Psychologist, 35,* 1021-1027.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H.I. Braun (Eds.), *Test Validity.* (pp. 33-46). Hillsdale, N.J.: Lawrence Erlbaum.

Messick, S. (1989). Validity. In R.L. Linn (Ed), *Educational* Measurement (3rd Ed.). (pp. 13-103). New York: MacMillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific enquiry into score meaning. *American Psychologist, 50,* 741-749.

Michell, J. (1990). *An introduction to the logic of psychological measurement.* Hillsdale: NJ: Lawrence Erlbaum Associates.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology, 88(3),* 355-383.

Miller, D. Joseph, S. & Tudway, J. (2004). Assessing the component structure of four self-report measures of impulsivity. *Personality and Individual Differences, 37,* 349-358.

Miller, H. & Bichsel, J. (2004). Anxiety, working memory, gender, and math performance. *Personality and Individual Differences, 37,* 591-606.

Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (2003). On the structure of Educational Assessment. *Measurement: Interdisciplinary Research and Perspectives, 1(1),* 3-62.

Moneta, G.B. & Yip, P.P.Y. (2004). Construct validity of the scores of the Chinese version of the Need for Closure Scale. *Educational and Psychological Measurement, 64(3),* 531-548.

Motl, R.W., Dishman, R.K., Saunders, R.P., Dowda, M., & Pate, R.R. (2004). Measuring social provisions for physical activity among adolescent black and white girls. *Educational and Psychological Measurement, 64(4),* 682-706.

Müller,J., Bühner, M, & Ellgring, H. (2004). The assessment of alexithymia: Psychometric properties and validity of the Bermond-Vorst Alexithymia Questionnaire. *Personality and Individual Differences, 37,* 373-391.

Mungas, D., Reed, B.R., & Crane, P.K. (2004). Spanish and English Neuropsychological Assessment Scales (SENAS): Further development and psychometric characteristics. *Psychological Assessment, 16(4),* 347-359.

Muraki, E. & Engelhard, G. (1985). Full-information item factor analysis: Applications of EAP scores. *Applied Psychological Measurement, 9(4),* 417-430.

Muris, P., de Jong, P.J., &, Engelen, S. (2004). Relationships between neuroticism, attentional control, and anxiety disorders symptoms in non-clinical children. *Personality and Individual Differences, 37,* 78-797.

Nagtegaal, M.H. & Rossin, E. (2004). The usefulness of the thought suppression paradigm in explaining impulsivity and aggression. *Personality and Individual Differences, 37,* 1233-1244.

Novick, M.R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3,* 1-18.

Nunnally, J.C. (1967).. *Psychometric Theory* (2nd Ed.). New York: McGraw-Hill.

O'Connor, R.M., Colder, C.R., & Hawk, L.W. (2004). Confirmatory factor analysis of the Sensitivity to Punishment to Reward Questionnaire. *Personality and Individual Differences, 37,* 985-1002.

Oliver, M.N.I. & Simons, J.S. (2004). The Affective Lability Scales: Development of a short-form measure. *Personality and Individual Differences, 37,* 1279-1288.

Osman, A. Kopper, B.A., Barrios, F., Gutierrez, P.M., & Bagge, C.L. (2004). Reliability and validity of the Beck Depression Inventory-II with adolescent psychiatric patients. *Psychological Assessment, 16(2),* 120-132.

Parkinson, A, Mullally, A.A.P., Redmond, J.A. (2004). Test-retest reliability of Riding's Cognitive Styles Analysis test. *Personality and Individual Differences, 37,* 1273-1278.

Peak, H. (1953). Problems of objective observation. In L. Festinger & D. Katz (Eds.), *Research methods in the behavioral sciences.* (pp. 243-299). New York: Holt, Rinehart and Winston.

Piper, W.E., Ogrodniczuk, J.S., & Joyce, A.S. (2004). Quality of object relations as a moderator of the relationship between pattern of alliance and outcome in short-term individual psychotherapy. *Journal of Personality Assessment, 83(3),* 345-356.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Rodebaugh, T.L., Woods, C.M., Thissen, D.M., Heimberg, R.G., Chambless, D.L., & Rapee, R.M. (2004). More information from fewer questions: The factor structure and item properties of the original and brief Fear of Negative Evaluation Scale. *Psychological Assessment, 16(2),* 169-181.

Roesch, S.C., Rowley, A.A., & Vaughn, A.A. (2004). On the dimensionality of the Stress-Related Growth Scale: One, three or seven factors? *Journal of Personality Assessment, 82(3),* 281-290.

Roskam, E.E. & Ellis, J. (1992). 'The irrelevance of factor analysis for the study of group differences': Commentary. *Multivariate Behavioral Research, 27(2),* 205-218.

Samejima, F. (1974). Normal ogive model on the continuous response level in the multidimensional space. *Psychometrika, 39,* 111-121.

Schuller, R.A., Wells, E., Rzepa, S., & Klippenstine, M.A. (2004). Rethinking Battered Women's Syndrome evidence: The impact of alternative forms of expert testimony on mock jurors' decisions. *Canadian Journal of Behavioural Science, 36(2),* 127-136.

Sears, G.J. & Rowe, P.M. (2003). A personality-based similar-to-me effect in the employment interview: Conscientiousness, affect-versus competence-mediated interpretations, and the role of job relevance. *Canadian Journal of Behavioural Science, 35(1),* 13-24.

Shaw, A. & Joseph, S. (2004). Principal components analysis of Maltby and Day's (1998) amended quest religious orientation scale: A replication of the three component structure. *Personality and Individual Differences, 37,* 1425-1430.

Sirois, F.M. (2004). Procrastination and intentions to perform health behaviors: The role of self-efficacy and the consideration of future consequences. *Personality and Individual Differences, 37,* 115-128.

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

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15,* 72-101.

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *American Journal of Psychology, 18,* 160-169.

Spearman, C. (1910). Correlation from faulty data. *British Journal of Psychology, 3,* 271-295.

Spence, G. Oades, L.G., & Caputi, P. (2004). Trait emotional intelligence and goal self-integration: Important predictors of emotional well-being? *Personality and Individual Differences, 37,* 449-461.

Stevens, S.S. (1946). On the theory of scales of measurement. *Science, 103,* 667-680.

Stevens, T. & Tallent-Runnels, M.K. (2004). The Learning and Study Strategies Inventory—high school version: Issues of factorial invariance across gender and ethnicity. *Educational and Psychological Measurement, 64(2),* 332-346.

Stout, W.F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 55(2),* 293-325.

Stout, W.F. (2002). Psychometrics: From practice to theory and back: 15 years of nonparametric multidimensional IRT , DIF/test equity, and skills diagnostic assessment. *Psychometrika, 67(4),* 485-518.

Taub, G.E., McGrew, K.S., & Witta, E.L. (2004). A confirmatory analysis of the factor structure and cross-age invariance of the Wechsler Adult Intelligence Scale-third edition. *Psychological Assessment, 16(1),* 85-89.

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.L. (1931). *The reliability and validity of tests.* Ann Arbor, MI: Edwards Brothers.

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

Tobey, E.L. & Tunnell, G. (1981). Predicting our impression on others: Effects of public self-consciousness and acting, a self-monitoring subscale. *Personality and Social Psychology, 7(4),* 661-669.

Tucker, L.R. (1946). Maximum validity of a test with equivalent items. *Psychometrika, 11(1),* 1-13.

Ullstadius, E., Carlstedt, B, & Gustafsson, J-E. (2004). Multidimensional item analysis of ability factors in spatial test items. *Personality and Individual Differences, 37,* 1003-1012.

Utsey, S.O., Brown, C. & Bolden, M.A. (2004). Testing the structural invariance of the Africultural Coping Systems Inventory across three samples of African descent populations. *Educational and Psychological Measurement, 64(1),* 185-195.

van der Ploeg, E., Mooren, T.T.M., Kleber, R.J., van der Velden, P.G., Brom, D. (2004). Construct validation of the Dutch version of the Impact Event Scale. *Psychological Assessment, 16(1),* 16-26.

Vittengl, J.R., Clark, A.L., & Jarrett, R.B. (2004). Self-directed affiliation and autonomy across acute and continuation phase cognitive therapy for recurrent depression. *Journal of Personality Assessment, 83(3),* 235-247.

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin, 83(2),* 213-217.

Walker, H.M. (1929). *Studies in the history of statistical method.* Baltimore: William & Wilkins Co.

Wallace, J.C. (2004). Confirmatory factor analysis of the Cognitive Failures Questionnaire: Evidence for dimensionality and construct validity. *Personality and Individual Differences, 37,* 307-324.

Weiss, D.J. & Davison, M.L. (1981). Test theory and methods. *Annual Review of Psychology, 32,* 629-659.

Williams, K.M. & Paulhus, D.L. (2004). Factor structure of the Self-Report Psychopathy Scale (SRP-II) in non-forensic samples. *Personality and Individual Differences, 37,* 765-778.

Wolfe, E.W., Ray, L.M., & Harris, D.C. (2004). A Rasch analysis of three measures of teacher perception generated from the School and Staffing Survey. *Educational and Psychological Measurement, 64(5),* 842-860.

Woolley, S.L., Benjamin, W..J., & Woolley, A.W. (2004). Construct validity of a self-report measure of teacher beliefs related to constructivist and traditional approaches to teaching and learning. *Educational and Psychological Measurement, 62(2),* 319-331.

Zhang, L.-F. (2004). Do university students' thinking styles matter in their preferred teaching approaches? *Personality and Individual Differences, 37,* 1551-1564.

Zweig, D. & Webster, J. (2004). Validation of a multidimensional measure of goal -orientation. *Canadian Journal of Behavioural Science, 36(3),* 232-243.

APPENDICES

Appendix 1: List of Reviewed Journals

Journal Title	Volume number	Date
Canadian Journal of Behavioural Science	35(1)	January 2003
	35(2)	April 2003
	35(3)	July 2003
	35(4)	October 2003
	36(1)	January 2004
	36(2)	April 2004
	36(3)	July 2004
	36(4)	October 2004
Educational and Psychological Measurement	64(1)	February 2004
	64(2)	April 2004
	64(3)	June 2004
	64(4)	August 2004
	64(5)	October 2004
	64(6)	December 2004
Journal of Personality Assessment	82(1)	February 2004
	82(2)	April 2004
	82(3)	June 2004
	83(1)	August 2004
	83(2)	October 2004
	83(3)	December 2004
Personality and Individual Differences	37(1)	July 2004
	37(2)	July 2004
	37(3)	August 2004
	37(4)	September 2004
	37(5)	October 2004
	37(6)	October 2004
	37(7)	November 2004
	37(8)	December 2004
Psychological Assessment	16(1)	March 2004
	16(2)	June 2004
	16(3)	September 2004
	16(4)	December 2004

Appendix 2: Review Form

Article:

_____
_____
_____
_____
_____
_____

Article Number:

Date Reviewed:

Test evaluation?	**Yes**	**No**	**NEI**
Reliability analysis?	**Yes**	**No**	**NEI**
Pre-existing measure?	**Yes**	**No**	**NEI**
Conclude that reliability/validity has been established previously?	**Yes**	**No**	**NEI**

Do the authors:

1) identify the aim of the analysis (i.e., exploratory/confirmatory)	**Yes**	**No**	**Somewhat**
a) exploratory?	**Yes**	**No**	**Somewhat**
b) confirmatory?	**Yes**	**No**	**Somewhat**
c) confound both aims?	**Yes**	**No**	**Somewhat**
2) identify a theoretical structure (i.e., make an explicit statement as to how the test should perform)?	**Yes**	**No**	**Somewhat**
3) explicitly address the dimensionality of the test?	**Yes**	**No**	**Somewhat**
4) examine the structure of the test?	**Yes**	**No**	**Somewhat**
a) using an appropriate model?	**Yes**	**No**	**NEI**
i) based on measurement scale(s)?	**Yes**	**No**	**NEI**
ii) based on the form of the item/θ regressions?	**Yes**	**No**	**NEI**
iii) which is unidimensional in some sense?	**Yes**	**No**	**NEI**
iv) if yes to a), which statistical model?			

_____ _____

	Yes	No	NEI
b) employ an index of fit?			

a) what is the scoring rule, if explicit? _____

	Yes	No	NEI
6) employ a classical index of reliability?			
a) internal consistency/coefficient alpha?			
b) test-retest?			
c) item-total correlations?			
d) alternate forms?			
e) KR20 or KR21?			
f) Spearman-Brown			
g) other?			
7) employ more than one index of reliability?			
8) compute subscale reliabilities, and an overall?			
9) compute subscale reliabilities, but no overall?			
10) compute overall reliability, but no subscale reliabilities?			

	Yes	No	Somewhat
11) use reverse logic (i.e., estimate reliability, *then* examine structure?)			
12) conclude that reliability is good?			
13) consider a particular population?			
14) consider stability?			
a) separately from precision?			
15) explicitly address validity?			
16) estimate validity?			

a) with which coefficient? _____

	Yes	No	Somewhat
17) employ a logic of any sort?			

Comments:

Appendix 3: Summary of Key Findings

N = 251

Nature of Test Analysis	n
At least one property* of the test(s) assessed empirically	**213**
Both structure and precision assessed	79 (37%)
Only structure assessed	12 (6%)
Only precision assessed	109 (51%)
Neither structure or precision assessed	9 (4%)
Not enough information to determine which, if any, properties of the test(s) were assessed	4 (1%)
Aim of the analysis explicitly stated	**85**
Confirmatory aim	32 (38%)
Exploratory aim	6 (7%)
Blend of confirmatory and exploratory aims	47 (55%)
Aim of analysis not explicitly stated	**166**
Both structure and precision assessed	14 (8%)
Only structure assessed	3 (2%)
Only precision assessed	104 (63%)
Neither structure or precision assessed	43 (26%)
Not enough information to determine which, if any, properties of the test(s) were assessed	2 (1%)

Structure Assessed	**94**
Expected dimensionality of item responses explicit	26 (28%)
Unidimensional model employed	4 (5%)
Precision also assessed	79 (84%)
Precision Assessed	**189**
Form of compositing rule explicit	46 (25%)
Cited Previous Reliability/Validity Findings	**119**
Previous reliability findings	38 (32%)
Previous validity findings	7 (6%)
Both previous reliability and validity findings	74 (62%)
Classical Estimates of Reliability Employed	**161**
Coefficient alpha employed	151 (94%)
Test-retest correlation employed	37 (23%)
Item-total correlation employed	15 (9%)

* Precision assessed, association structure examined, and/or validity assessed