CENTRE FOR HEALTH
PROGRAM EVALUATION

RESEARCH REPORT 1

What is Validity?
A Prologue to an Evaluation of Selected Health Status Instruments

Kaye Brown
Research Fellow
National Centre for Health Program Evaluation

Colin Burrows
Senior Research Fellow,
National Centre for Health Program Evaluation

Reprint of NCHPE Research Report 1
June 1997

First Published September 1992
ISSN 1038-9555
ISBN 0 642 18419 4

CENTRE PROFILE

The Centre for Health Program Evaluation (previously known as the NHMRC National Centre for Health Program Evaluation) was formed to meet the need for evaluation of health programs and the development of appropriate evaluation methodologies for the Australian context. It seeks to contribute to government policy deliberations and to the national debate on health insurance policy, hospital waiting lists, the cost of new health technologies, the effectiveness of health promotion programs, and many other issues relating to the best use of the nation's health resources.

PUBLICATIONS

The views expressed in Centre publications are those of the author(s) and do not necessarily reflect the views of the Centre or its sponsors. Readers of publications are encouraged to contact the author(s) with comments, criticisms and suggestions.

A list of the Centre's papers is provided inside the back cover. Further information and copies of the papers may be obtained by contacting:

The Co-ordinator
Centre for Health Program Evaluation
PO Box 477
West Heidelberg Vic 3081, Australia
Telephone + 61 3 9496 4433/4434 **Facsimile** + 61 3 9496 4424
E-mail CHPE@BusEco.monash.edu.au

ACKNOWLEDGMENTS

The Health Economics Unit of the CHPE receives core funding from the Public Health Research and Development Committee of the National Health and Medical Research Council, Monash University, and the Victorian Health Promotion Foundation.

The Program Evaluation Unit of the CHPE receives core funding from the Victorian Health Promotion Foundation and The University of Melbourne.

The research described in this paper is made possible through the support of these bodies.

AUTHOR ACKNOWLEDGMENTS

We are grateful to the above funding bodies for their support.

We would also like to acknowledge the assistance of Sanya Pahoki, Celia Pollard and Adam Maclagan who diligently tracked down countless references for us. Fay Bower (Librarian, Biomedical Library, Alfred Hospital Branch) provided invaluable help in processing numerous inter-library loan requests and verifying incomplete reference citations.

'*Quality of life* is some like *intelligence*. Everyone knows it exists and thinks they can identify it in various ways, but we may not be able to evoke universal agreement on what it is. We are probably better off letting people propose indexes, which we can then use or not use, rather than try to get a multi-individual consensus on what ought to be there.

The situation is much easier when we try to create ailment-oriented indexes [i.e., indexes that describe the things doctors observe as direct clinical events in the practice of medicine and in the evaluation of therapy] for clinical work. Because clinicians have good general agreement on the construct called "congestive heart failure," we might fight about how good a particular index is, but we don't have to fight about the construct itself. Because we do not have unanimous agreement about the construct of quality of life, however, the idea has become a kind of umbrella under which are placed many different indexes dealing with whatever the user wants to focus on.'

FEINSTEIN, 1987a

'An instrument, whether it is a test, scale, observation procedure, questionnaire, or interview schedule, only measures what it measures – nothing more, nothing less. One should take a long look at any instrument, once it is in place in an evaluation design, to be clear regarding what it can and cannot register. Beware of the "naming fallacy" – giving a name to a test or other instrument such as ... ["quality of life", "adjustment to illness", etc} and thereby maintaining that is what it measures. … The issue raised is the validity of the instrument. It should be a key consideration in its selection in the first place. Nevertheless, the restrictions inherent in any instrument should be kept to the forefront of attention in designing and evaluating any program'.

ISAAC & MICHAEL, 1981

'Far better than to construct measures ad hoc for particular investigations of change is to select existing measures that have proven themselves with respect to . . . [accepted psychometric] criteria. ... Even in very large-scale programs of research, it takes years to develop standard measures of psychological characteristics that meet these standards. In particular, gathering evidence for construct validity is a matter that takes numerous years, at best. Consequently, it is usually foolhardy for those who are entering a program of evaluation research (which is usually limited both in terms of time and funds) to undertake the development and standardization of most of the measures that will be employed. The far better part of valor and the far better part of commonsense is to seek suitable measures from those that have been ripening over the years.'

NUNALLY, 1975

‘Fitting quantitative variables to abstract constructs is a bit like using a luminous ruler to measure an elephant on a moonless night. We can obtain clear numbers, but we know that the numbers do not perfectly capture the dimensions of the beast. The ruler does not bend where the elephant bends; it slips when the elephant stamps its feet; and as we grope in the dark, it is hard to tell what portions of the elephant we have measured and which parts remain untouched. When we transfer our numbers to paper and try to sketch the elephant from the measured inches, part of our sketch is derived from what we already know about elephants – our intuition and commonsense knowledge about the shape and size of an elephant.’

KIDDER & JUDD, 1986

TABLE OF CONTENTS

Page

FOREWORD v

1 INTRODUCTION 1

2 MEASUREMENT 6

Scaling: A Caveat 7

Random and Non-Random Measurement Error 8

3. VALIDATION OF MEASUREMENT 15

Means-Ends Relationships in Health Status Assessment 17

X Is A Part of Y: Content Validity 18

X As A Sign of Y: Criterion-Related Validity 23

X Is An Approximation of Y: Construct Validity 30

The Multitrait-Multimethod Paradigm 33

Hypothesis Testing and the Significance of Significance 43

4. LONGITUDINAL VALIDITY – AND RESPONSIVENESS 46

Responsiveness Considered Further 49

5. GENERALIZABILITY 53

Cultural Considerations in Generalizability 57

6. CONSEQUENTIAL BASIS OF VALIDITY 62

7. SUMMARY AND CONCLUSION 71

8. REFERENCES 74

FOREWORD

Measurements of health, well-being and quality of life in the social sciences have a respectably long history going back at least to World War II with the development of instruments to measure physical and psychological well-being. However, in the last two decades, development and application of health status and health-related quality of life measures have become a growth industry, crossing, though seldom combining, the disciplines of medicine, sociology, psychology, epidemiology, operations research, statistics and economics.

The result has been a proliferation of measures ranging from special-purpose single-use indexes developed for particular population segments, diseases or purposes to broadly-defined generic quality of life measures designed for multi-purpose use in many (some claim all) populations. Unfortunately, though perhaps inevitably, this considerable endeavour has been largely ad hoc, often atheoretical and indicates, in many if not most cases, a worrying lack of scholarship. Too many measures have not been validated beyond simple test-retest reliability and in very few indeed have there been serious, let alone acceptable, validation studies.

This is, to us, both surprising and disturbing. It is disturbing because applications using unvalidated measures are entitled to be dismissed simply because the measures are not validated. It is disturbing, too, because the results of such studies are intended to be used in health policy formulation or clinical decisions about specific treatments and patient management. The paucity of validation research is surprising because there is an extensive, well-developed literature on health research and program evaluation methodology. While this does not give simple solutions in an admittedly-complex field, it does provide a conceptual and practical framework for the design of validation studies. Importantly, too, it also illuminates those areas where adequate validation is difficult or contentious – and therefore where one should be very careful about claims for research finding or interpretation of evaluation studies.

This monograph had its genesis in the receipt of a modest research grant to collect and review the literature on the validity of existing health status/health-related quality of life measures. The literature collection, as expected, led us into many fields of research and many source disciplines and resulted in the accumulation of several hundred books and papers. It led, too, to the realisation, again not unexpected, that most researchers worked mainly within their own disciplines and referred largely to the literature within their own or closely-related disciplines. Certainly, there appeared to be a lack of awareness of the specialised literature on evaluation methodology and, in particular, on validity in areas, such as this, where one is necessarily dealing with difficult abstract constructs.

The purpose of the monograph is to explore, and we hope illuminate, the concept of validity as it pertains to health status and health-related quality of life constructs. In so doing, we accept and follow the view that construct validity acts as a unifying concept and validation exercises are concerned with inferences drawn from research findings. Questions of content and face validity and scaling, whilst critically important, are subsumed into the operational realisation of a valid construct. For completeness, too, some attention is given to generalizability or external validity, cross-cultural considerations and the consequential basis of validity. However, because the central theme of the paper is the necessity to develop valid operational constructs of health status or health-related quality of life, these matters are treated in a more cursory manner.

There has been no attempt, either, to specify or illustrate detailed validation procedures though references are given to books and papers where such details are available.

It is our intention to use the paper as a basis for examining the validity of several well-known and commonly used measures of health status and health-related quality of life. We hope it will be of assistance to researchers intending to develop such measures and to those who wish to use existing instruments but are rightly concerned about their validity.

Kaye Brown

Colin Burrows

1 INTRODUCTION

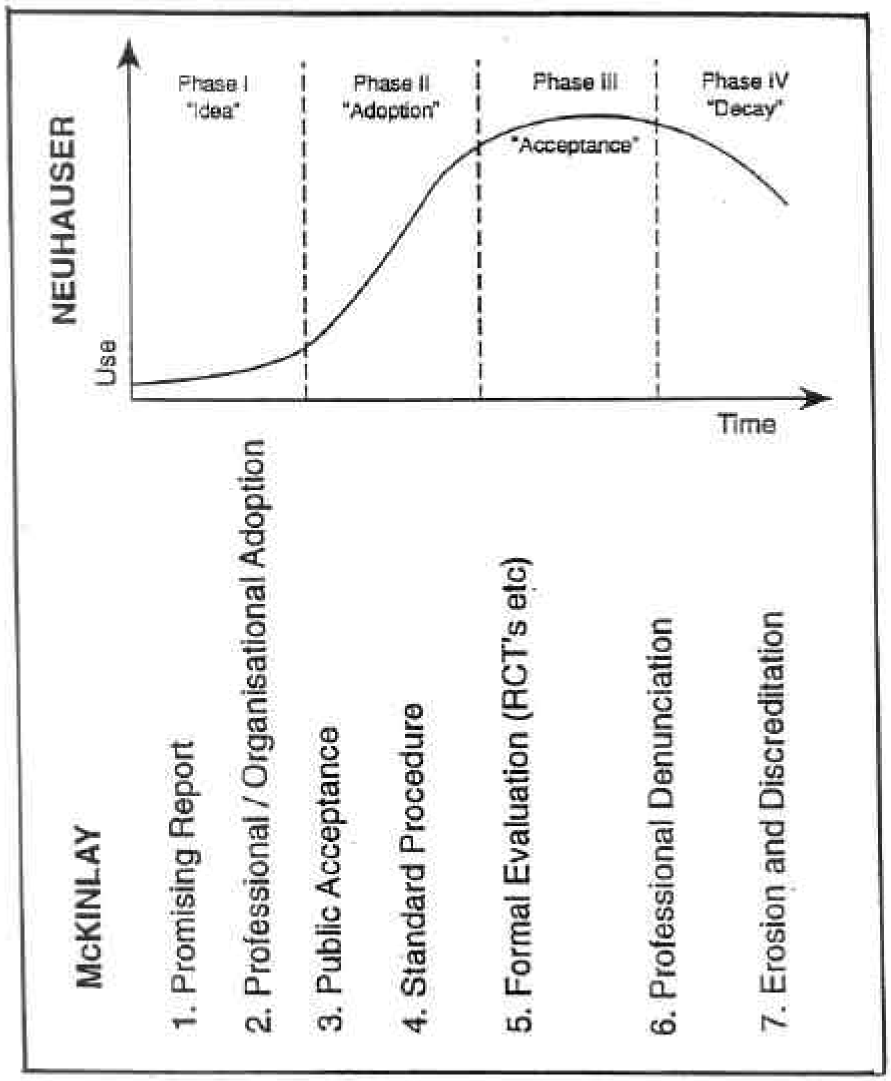
The realization that there is more to life than not dying was a long time coming. Now broader definitions of health are current. The impetus to broaden the focus of health status beyond mortality and other traditional biomedical parameters can be traced to the confluence of a number of factors, including: recognition that outcomes like disease-specific mortality are, at best, crude indicators of health; the trend toward greater consumer participation in health care decisions; increased emphasis on the formal evaluation of health care programs and services; and the necessity to choose among alternatives or to set priorities due to resource constraints. Whatever the catalyst(s), we have now reached the point where there is much interest in, and a burgeoning literature on, the measurement of health status and health-related quality of life.

The immediate problem is that much of this activity has been generating more heat than light. Poor scholarship and a cavalier approach to validation has led to a profusion of health status instruments to the point where Spitzer (1987b) has declared an epidemic – of a kind that may undermine the evidential basis of the whole endeavour. There are too many measures of unknown validity and too few researchers inclined to examine further the nomological validity of extant measures of health status and health-related quality of life. In this regard it seems to us that health program evaluators, health economists and others engaged in research in this area need to demonstrate their bona fides, as they would have others do.

In models of the diffusion of medical technologies it is clinicians who are cast as villains. It is they (we say) who advance the career of medical technologies from the stage of “promising report” to “standard procedure” (McKinlay, 1981) in short order and only belatedly engage in the formal evaluation that ought to antedate the adoption of new technologies. Yet, it seems to us that we are about to become guilty of doing likewise, if we are not already. In terms of the product life cycle curve for medical technology (see Figure 1) we are probably located at or around McKinlay’s Stage 2, and trying hard for Stages 3 and 4, without having paid too much attention to the exigencies of Stage 5. We should be evaluating what we’re doing in the areas of health status assessment and quality of life measurement with the same rigour as that we are supposed to apply to the technologies and programs we investigate. Very often, and with good reason, we are critical of clinicians, in particular, for not assessing efficacy and effectiveness properly. With respect to health status measures the shoe may be on the other foot, as the following quotations from two of the more eminent researchers in the field indicate:

> ‘Very seldom do we see statements submitted in advance in a protocol that specifies at what point a measure will be declared valid. It is true that validity work on a measure is a never-ending process, but there is a point in the continuum where it is possible to say that minimal criteria have (or have not) been met.
>
> How often do you hear somebody get up and say, “we’ve attempted to validate this and we’ve found that it is not valid so we’ve abandoned it”? Most of the time people talk about the superficial exercise they have done with their data, which they almost invariably choose to interpret as positive evidence of validity, and we are asked to accept it. That’s not good enough. We ought to tighten up our rigor in this area’. (Spitzer, 1987a, p 188)

Figure 1: Product life cycle Curve for Medical Technology: The Neuhauser and McKinlay Models.



Sources: Neuhauser D (1979), International workshop on the evaluation of medical technology, Stockhom, 18-19 September, SPRI Report; and McKinlay JB (1981), From promising report to standard procedure: seven stages in the career of a medical innovation, Milbank Memorial Fund Quarterly, 59(3): 347-411

> 'We especially need to emphasize that practitioners must do more to nail down the validity and reliability of their measures: McDowell and Newell [1989] strongly recommend that we focus hard on a few measures and get them well established and not veer off to develop a new scale on every occasion and for every problem. ... [I]n applications we need to reduce, not increase the variety of measures and to establish more firmly the value of the measures that we do use. We simply must have more on the validity and reliability of those measures than we have already seen.' (Mosteller, 1989, pp S282-3).

There is little point in delivering a variety of half-way technologies where the measurement of health status and health-related quality of life is concerned. The existence of too many measures militates against their understanding and acceptance in clinical research and clinical practice, and, importantly, detracts from the comparison of results across different studies addressing similar problems. It dissipates resources that may be allocated to further validation work (Spitzer, 1987a, 1987b). Less-than-well-validated instruments do not win over the sceptics and the promulgation of unvalidated ones is positively counterproductive.

In each of several monographs that follow this one we propose to examine the state and stage of development of the major measures of health status that would most likely represent the set of candidates from which measures, given the criterion of "well-validated and deserving of further development", would come. More than a bare literature review is required here. Rather, what is needed is a close and rather more detailed examination of the nature and extent of validation studies undertaken by the developers of the instrument *and* by other researchers than is customary in the many "shopping guides" to choosing a measure of health status or health-related quality of life. Particular attention will be paid to the incidence of hypothesis testing versus *post hoc ergo propter hoc* reasoning and the coherence of the findings that emerge across studies. Our emphasis will be on the degree to which researchers have engaged in construct validation and especially on the degree of convergent validity among the more widely-cited instruments.

It is not that others have not reviewed the literature that prompts us to take this on. Rather it is a frustration with the precised version of this endeavour that so often reduces everything to a brief description of the instrument, followed by, at best, a few short and not necessarily informative paragraphs about reliability and validity.

As an initial step, this monograph seeks to explore the concept of validity and its nuances. This discussion will provide a basis for reflecting on the validity of the several measures of health status reviewed in detail and will, we hope, make plain to those with limited acquaintance with the concept, why construct validity is the *sine qua non* of usable and useful measures of health status or health-related quality of life.

First, however, we should nominate the measures of health status and health-related quality of life we hope to review in some detail. They are (in the probable order in which they will be tackled and in order of the number of articles we have collected at this stage): the Sickness Impact Profile, the Quality of Well-Being Index, the General Health Rating Index, the Nottingham Health Profile, the Time-Trade-off approach, the Rosser Index and the McMaster Health Index Questionnaire.

Figure 2 shows where these standardized measures of health status fit in a schematic representation of different approaches to the measurement of health status. The first fork in Figure 2 reflects the general recognition of health as a kind of continuum in characterizing health status measures according to whether they focus on well-being/good health or illness/disability (resulting from the impact of disease or treatment outcomes). As indicated by the subsequent branches, health status measures are still based primarily on the negative aspects of health – notwithstanding the fact that the spectrum of health states extends from "perfectly healthy", if that exists, to "near death".

The second fork separates health status measures into general-purpose and disease-specific measures. Disease-specific measures are designed to assess the health of a particular patient population with a given disease or condition, in contradistinction to general-purpose measures which are designed to assess health across the broad spectrum of patients and populations. This fork reflects the trade-off between measures with a limited range of applicability that may be sensitive to small, clinically relevant differences and more robust measures that facilitate a more gross level of comparability across populations or programs. This important distinction seems to be lost on some investigators who have attempted to answer validity-related questions about generic measures using data obtained from small, relatively-homogeneous groups.

The next division focuses on the scale of application. This is not unrelated to the purposes for which health-related quality of life are measured. The compromise is struck between asking "large" numbers of individuals "small" numbers of questions about their health and asking a "small" numbers of individuals a "large" number of questions about their health and asking a "small" number of individuals a "large number of questions. Examples of the former approach include regular surveys (e.g., National Health Interview Survey in the United States, the health component of the General Household Survey in the United Kingdom, and the Australian Health Survey). Typically, most applications of the other health status measures identified in Figure 2 have more circumscribed target populations, though the number of individuals interviewed may still be very large, as in the best known applications of the Rand General health Rating Index or the Nottingham Health Profile. They reflect the goal of case-finding rather than that of measuring the level of a health component or components of health status (Ware, 1984c). The goal of case-finding is to place people into categories: to ascertain whether individuals have a particular illness or health problem. Methods for assessing the level of health status, on the other hand, attempt to locate people precisely along a health status continuum: to measure individual differences or changes in the level of a particular health status concept over time.[1]

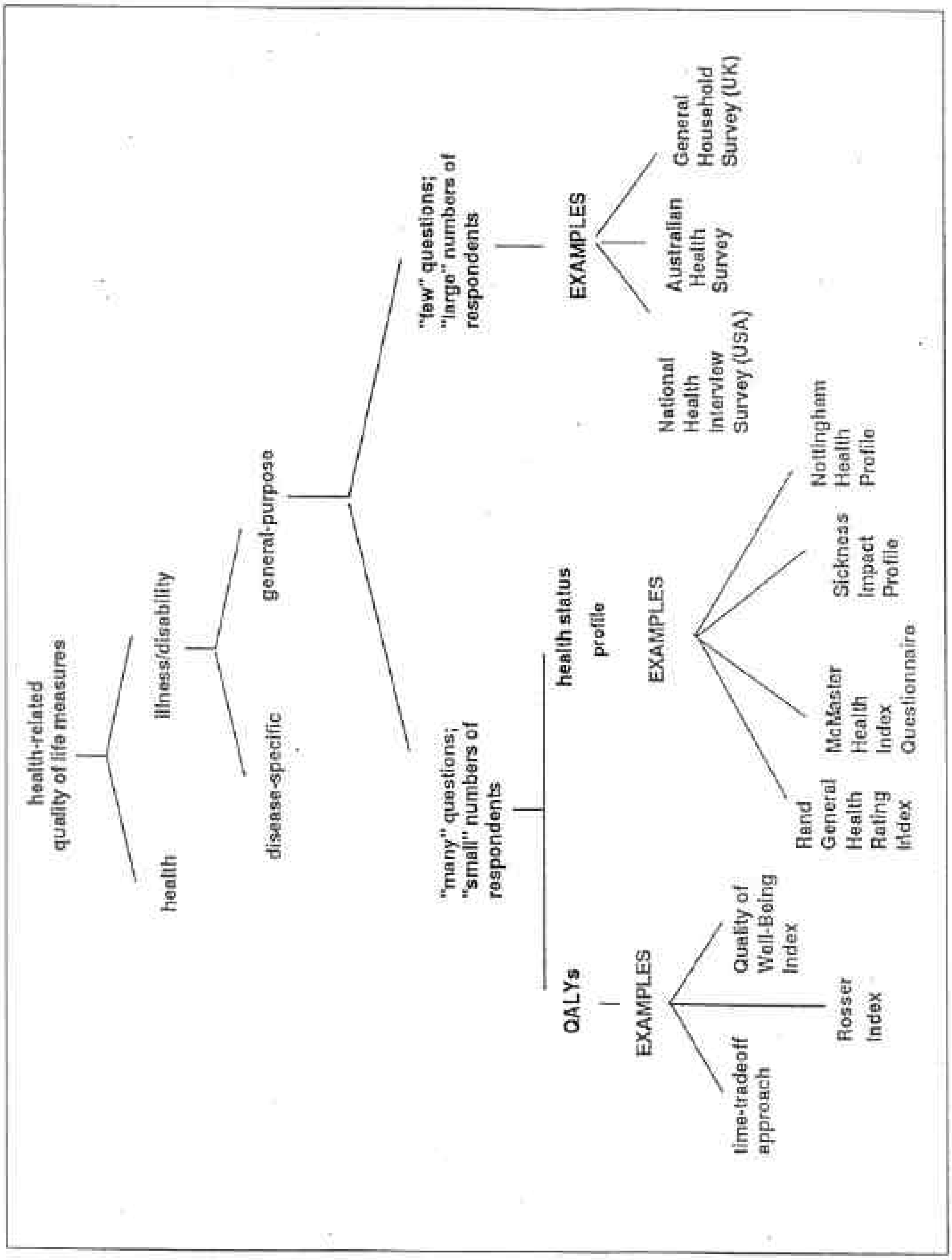
The last fork focuses on measures of health status that require the input of not inconsiderable levels of resources gauged either in terms of the demands placed on respondents and/or the necessity to use interviewers to administer the instruments. Examples of these measures include the Rand General Health Rating Index (Ware, 1984b; Ware, Davies-Avery &

[1] Note that the three forks discussed thus far cut across a distinction that Spitzer (1987b) has drawn between health status and (health-related) quality of life, terms which tend to be used as synonyms in the literature.

Spitzer's view is that the 'measurement of health status should be reserved primarily for assessments of ostensibly health people, usually in the context of aggregates of unselected geographically-defined populations or "catchment area" delineated clinics of a service program'. In contrast, 'the measurement of quality of life should be restricted to the assessment of a series of attributes among those *definitely* sick. The person classified by health-related or health-sensitive quality of life measuring instruments should have clear-cut manifestations of disease according to established explicit or at least implicit criteria of one or more diagnoses' (pp 467-568).

Spitzer's distinction between health status and health-related quality of life is seemingly at odds with Ware's (1984b) distinction between case-finding and measuring the level of a health component or components of health status, as discussed in the text.

Figure 2: Approaches to the Measurement of Health Status


health-related quality of life measures
health
illness/disability
disease-specific
general-purpose
"many" questions; "small" numbers of respondents
"few" questions; "large" numbers of respondents
QALYs
health status profile
EXAMPLES
time-tradeoff approach
Rosser Index
Quality of Well-Being Index
EXAMPLES
Rand General Health Rating Index
McMaster Health Index Questionnaire
Sickness Impact Profile
Nottingham Health Profile
EXAMPLES
National Health Interview Survey (USA)
Australian Health Survey
General Household Survey (UK)

Brook, 1980); the McMaster Health Index Questionnaire (Chambers, 1988); the Sickness Impact Profile (Bergner, 1987; Bergner, Bobbitt, Carter & Gilson, 1981; Bergner, Bobbitt, Kressel, et al, 1976); Nottingham Health Profile (Hunt, McEwen & McKenna, 1986, McEwen 1988); the Quality of Well-Being Index (Kaplan & Anderson, 1988, 1990; Kaplan & Bush, 1982); the Rosser Index (Kind, Rosser & Williams, 1982; Rosser, 1987a, 1990; Rosser & Watts, 1972; Williams, 1985); and the time-tradeoff approach (Feeny, Labelle & Torrance, 1990; Torrance,1976, 1986, 1987; Feeney & Torrance, 1989). This fork then separates these "high-intensity" health status measures according to the level of aggregation of their output. Multidimensional health measures can provide a single aggregated score across all dimensions or scores for all disaggregated dimensions or the option of both if they prescribe a method of aggregation. Holistic measures such as the time-tradeoff approach, health year equivalents (Mehrez & Gafni, 1989a, 1989b) or utility measures (e.g. Boyd, Sutherland, Heasman, et al, 1990; Llewellyn-Thomas, Sutherland, Tibshirani, et al, 1982; Read, Quinn, Berwick, et al, 1984) yield only a single score and, therefore, minimal information. QALYs (quality-adjusted life years) represent, perhaps, the most sought after single index of health status for those involved in economic evaluations.

2 MEASUREMENT

There is nothing inherent in the intention to measure that says the something to be measured must be clearly defined at the outset. Measurement, by definition, is simply the assignment of numbers to events, objects or individuals, according to specified rules. Whether the attribute being measured is physical or psychological, "hard" or "soft", the focus of measurement is necessarily on the "something" that is measured. The something in question may be abstract constructs that are incorporated into a theoretical or conceptual framework; "unobservables" that exist in the minds of researchers, such as health status, social support, and adjustment of illness; or concrete variables like the everyday physical attributes of height, weight, temperature, blood pressure and serum cholesterol. Whatever the something, the question of the validity of measurement arises. As it happens, we have a common understanding of the last class of attributes – there are explicit definitions, generally accepted measurement instruments, and so on. Not so for the constructs of health status, social support or adjustment to illness.

As nonobservables, constructs have no direct measures. To collect data about a construct, a researcher must choose one or more observable things to serves as instances or indicators of the phenomenon. The class of observable things chosen to represent an unobservable construct is known as an operational definition. The operational definition specifies what the researcher will do to ascertain the value of the conceptual variable in a given empirical instance. Not surprisingly, this 'process of linking abstract concepts to empirical indicants' (Zeller & Carmines, 1980, p 2) raises questions of the good ness of the mapping or correspondence between concept and operationalization. This is the theme of validity. The validity question asks, in effect, whether a test measures what it purports to measure. Does the operational definition "truly" measure the corresponding property as conceptually defined? A valid indicator of health status or health-related quality of life allows us to make inferences about the health status of the individuals assessed, and not about some other variable (eg, adjustment to illness).

Scaling: A Caveat

Measurement is a process that involves both theoretical and empirical considerations. From an empirical point of view, the focus is on the observable response – the answer given to the interviewer, the mark on a self-administered questionnaire. Theoretically, interest lies in the underlying unobservable (and directly unmeasurable) concept that the response represents. Another important facet to the usefulness and meaningfulness of health status indexes concerns the procedures for inferring numerical values from the kinds and types of responses elicited. This is the issue of scaling, 'the process by which we record and measure variables' (Ghiselli, Campbell and Zedeck, 1981, p 391). Scaling is an attempt to quantify individuals' responses to stimuli (eg, descriptions of health states, health as it is experienced, etc).

Scaling methods (eg, standard gamble, time-tradeoff, category ratings, magnitude estimation, equivalence and willingness-to-pay) differ in the level of measurement they achieve, and in terms of the level of measurement required of respondents *versus* the level of measurement associated with the resulting scale. For instance, with direct scaling, subjects are instructed to respond with or generate the scale required and the resulting data are treated as reflecting this level of measurement. No analytical techniques are required to transform the response from one type of scale to another; rather the desired scale is obtained directly from the subject. With indirect scaling, on the other hand, subjects are instructed to respond at a certain level of measurement, and the data are subsequently converted to a different scale by the researcher.

It is important to know what level of measurement a particular scaling method yields because the range of algebraic operations applicable to the data is constrained by the level of measurement achieved. An obvious example is the application of quality of life measurement to resource allocation decision. The comparisons across programs inherent in the compilation of "league tables" from cost-utility analyses requires a ratio-scaled denominator. Most health economists have assumed that quality-adjusted life years (QALYs) have ratio scale properties but so far only a very few (eg, Nord, 1991, in press) have addressed this issue. Similarly, other researchers concerned with the measurement of health status have typically *assumed* that subjects are able to generate interval and ratio scales directly. Needless to say, the assumption *should* be tested.

We will take up the issue of scaling as it applies to the measurement of health status in a later paper because it is critically important in broadly-based generic measures and is a difficult area. Here we will concentrate on the relationship between observed responses and unobservable constructs; ie, on the assessment of validity. Of course, to the extent that the assumptions made about the scales associated with health status measures are not upheld, this puts the cart before the horse.[2]

[2] The level of measurement that researchers need to demonstrate depends on the use(s) to which health status scores are likely to be put and the requirements of analytical techniques that are brought to bear on the data. For example, in controlled clinical trials evaluative indexes (as defined later in the text, see p.) are used to determine whether there is a treatment effect and often the data are analysed using analysis of variance techniques. Application of analysis of variance (ANOVA) requires only interval scale data.

Random and Non-Random Measurement Error

According to classical test theory, empirical measurements are affected by only one type of error: *random error*. In this formulation, an observed score, X, is equal to the true score, T, plus a measurement error, E: X = T + E. The random error component is uncorrelated with the true score and subsumes all the chance factors that operate to confound the measurement of any phenomenon. In survey research, the types of errors that may be assumed to be random include those due to coding, ambiguous instruction, interviewer fatigue and the like. The amount of random error is inversely related to the **reliability** of a measuring instrument.

Reliability concerns the extent to which measurements are **repeatable** in independent assessments. The more consistent the results achieved by repeated measurements, the greater the reliability of the measuring procedure: conversely, the less consistent the results, the less reliable the instrument. 'Thus, a highly reliable indicator of a theoretical concept is one that leads to consistent results on repeated measurements because it does not fluctuate greatly due to random error' (Carmines & Zeller, 1979, p 12).

A number of observations about random error and reliability are commonly made which are relevant to the measurement of health status. The first is that indicators always contain some degree of random error. The very process of measurement means that no indicator can be perfectly reliable: error-free measurement is impossible. As Stanley (1971, p 365) has observed, 'The amount of chance error may be large or small, but it is universally present to some extent. Two sets of measurements of the same features of the same individuals will never exactly duplicate each other'. The choice among indicators does not hinge on whether they contain random error but instead on the *extent* to which they contain random error, *ceteris paribus*.

The second point that should be underscored is that the effects of random error are totally **unsystematic** by nature. The analogy is to the archer who sprays arrows about a point on the target such that they are as likely to hit to the right of the target as to its left or as likely to hit above or below the target. Knowing where the last arrow landed does not allow one to predict where the next one will hit.

Thirdly, it should be clear that reliability is basically an empirical issue, rather than a theoretically-oriented one, and therefore less important than validity. Obviously, it is more important to have a set of indicators that corresponds to the concept one wants to represent empirically than to have a set of indicants that are repeatable but not related to the concept in question. On the other hand, to have a valid measure, one must have a reliable one. The idea is captured schematically in Figure 3.

Still the bald statement of the necessary-but-not-sufficient requirement finesses some of the subtlety of the relationship between reliability and validity. Conceptually, the distinction between reliability and validity *is* clear-cut. In practice, however, the methods used to assess these properties – particularly the general tendency to evaluate validity in terms of bivariate rather than multivariate relationships and to assess the relationship between two measurements in terms of correlation coefficients – suggests that these two different characteristics of measurement lie at either end of a continuum, as illustrated in Figure 4. In the case of reliability, the two measurements come from the same instrument. In the case of validity, the two

Figure 3: Reliability Versus Validity

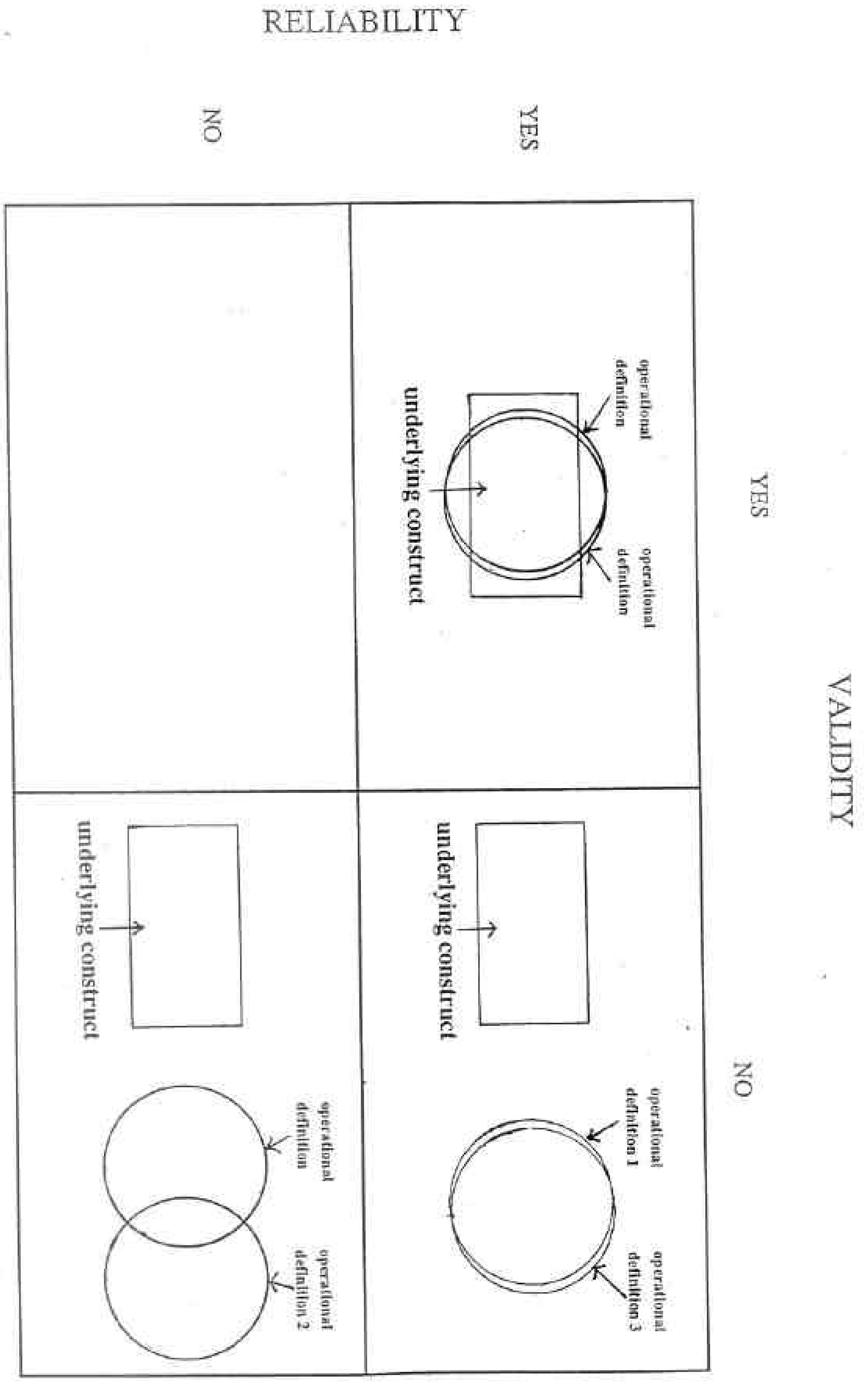
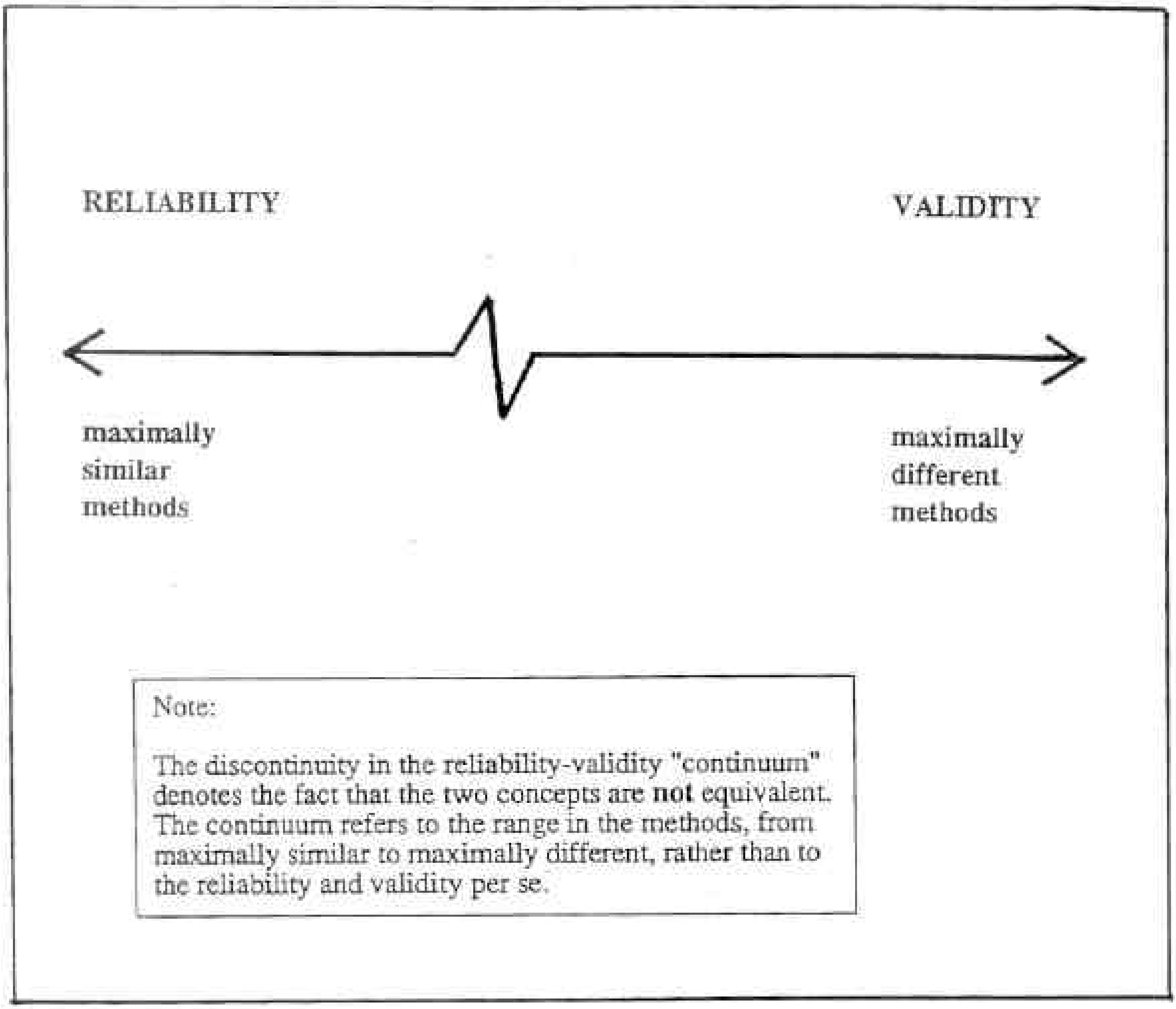

RELIABILITY
YES
NO
VALIDITY
YES
NO
operational definition
operational definition
underlying construct
underlying construct
operational definition 1
operational definition 3
underlying construct
operational definition
operational definition 2

Figure 4: The Reliability-Validity "Continuum"



Note:

The discontinuity in the reliability-validity "continuum" denotes the fact that the two concepts are **not** equivalent. The continuum refers to the range in the methods, from maximally similar to maximally different, rather than to the reliability and validity per se.

measurements derive from different instruments. That being the case, it seems obvious that reliability is different from validity. However, if we recognise that "different" instruments used to assess validity can have varying degrees of "difference", and the "same" instrument can have varying degrees of "sameness", the nature of the continuum becomes plain.

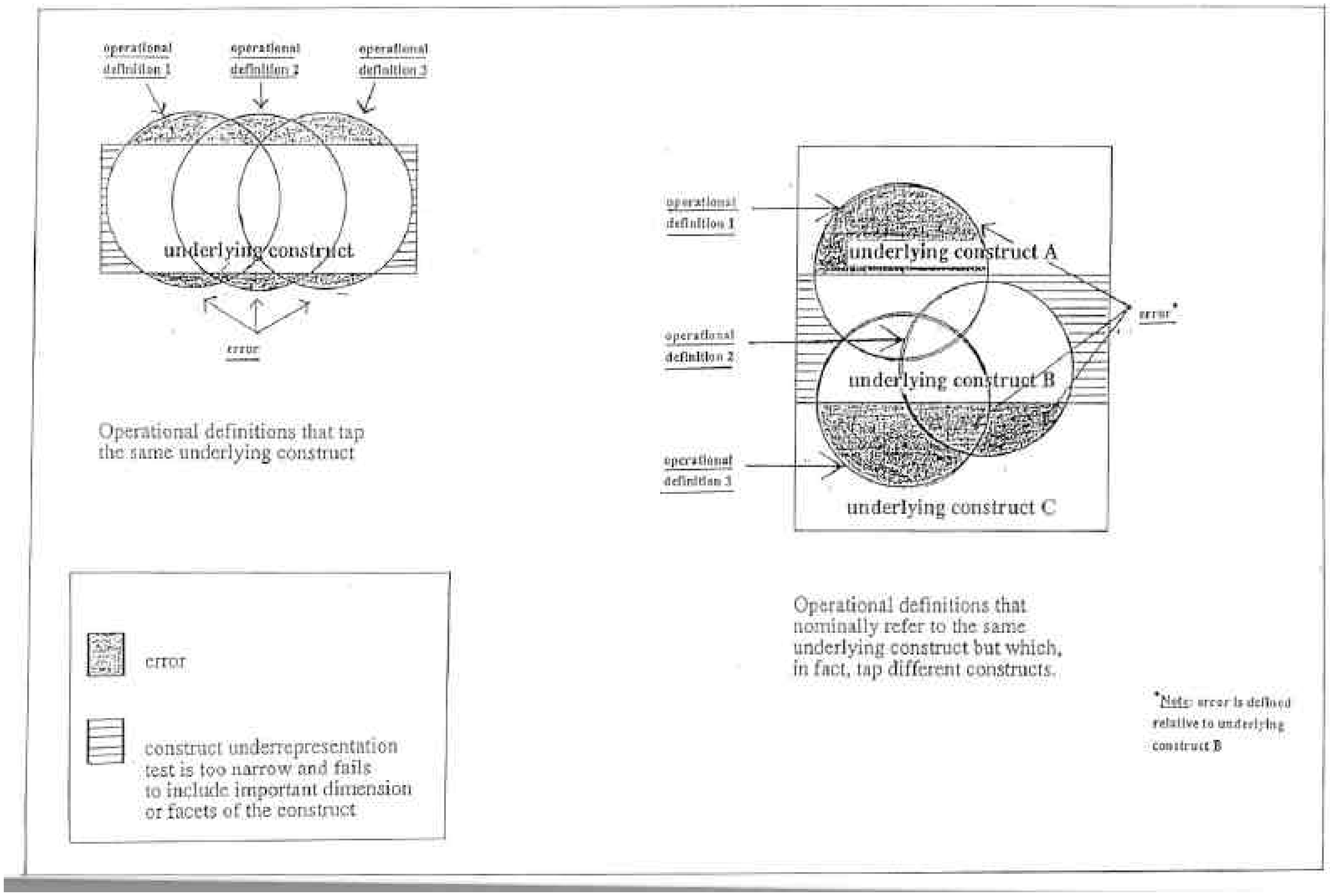
At one end of the continuum are correlations between identical methods of measurement. Test-retest correlations are based on two administrations of the same test. At the opposite end of the continuum are correlations between very different methods of similar methods; validity estimates are correlations between maximally dissimilar methods (Kidder & Judd, 1986). Both ends of the scale have obvious practical limits beyond which the correlations are completely meaningless. The perfect correlation between test and retest measures that would ensue if the second administration of a test occurred immediately after the first and each respondent simply copied their answers from the previous administration would indicate nothing about the reliability of an instrument. It would be a foregone conclusion. Similarly, if two measures were so different that they lacked even face validity as measures of the same variable it would not be surprising to find a low correlation. Figure 5 contrasts the case of three methods which do, in fact, measure the same underlying construct with that of three methods that perhaps nominally refer to the same construct but nonetheless tap three different constructs. The trick is to find methods that are maximally different but which tap the same variable. Then it is easy to defend the proposition that one is measuring validity, not reliability. The problem is that the more methods differ the less likely it is that they are concerned with the same variable. The multitrait-multimethod approach to construct validation, discussed below, takes advantage of hypothesized similarities and differences across measures.

In fact, random error is only one of two types of error. A second type of error that impinges on empirical measurements, which is ignored by classical test theory, is **systematic** or **non-random error**. According to this formulation, the observed score has three components: the true score, T, the systematic error component, S, and the random error component, E: X = T + S + E. A non-random or constant error is one that systematically affects either the characteristic being measured or the process itself, tilting the results in one direction or the other. One kind of systematic error is bias, a consistent tendency for a measure to be higher or lower than it "should be". A bias that is constant across all subjects responding to a survey may seriously distort measures of central tendency but it will not affect the relationships at the bivariate or multivariate level. For example, "yeasaying" (acquiescence) and "naysaying" responses may have a biasing effect on measuring instruments (see for example, Moum, 1988). A second source of systematic error occurs where deviations from "true" scores on one measure *are* related to deviations in another measure being analysed concurrently, ie, measurement errors are correlated. Validity estimates are affected by both systematic error and random error. Reliability is affected only by random error. Table 1 summarises the relationship between random and systematic error and reliability and validity.

Table 1: Categories of Error for Validity and Reliability

	Random Error	Systematic Error
Reliability	Applicable	Not applicable
Validity	Applicable	applicable

Figure 5: Validity: Operational definitions may include some irrelevant components and exclude some relevant portions of the underlying construct



Operational definitions that tap the same underlying construct

Operational definitions that nominally refer to the same underlying construct but which, in fact, tap different constructs.

*Note: error is defined relative to underlying construct B

Method effects are probably the major source of correlated error in survey data. They can occur for any type of survey item if there can be variation among subjects in the interpretation of the introduction, the question put, and/or the response scale employed. More generally, correlated errors will emerge wherever subjects differ, these differences affect the way in which subjects answer two or more times, and these differences are not linked to the concept(s) the items were intended to tap. The following example, after Andrews (1984), illustrates the phenomenon. Suppose a survey item asks subjects to evaluate their own health by selecting one of a number of response categories ranging from "excellent" to "quite poor". Their answers will differ, partly because people's perceptions of their own health do differ (valid variance). However, answers may also differ because people interpret the response categories differently (eg, "quite poor" may be more negative to some subjects than for others). This is measurement error due to method (methods variance). If a second survey item using the same response scale is included in the analysis with the above item, and if each subject interprets the meaning of the response categories in a consistent fashion, the two items would share measurement errors attributable to method.

The overlap in method effects give rise to covariation between the items, and this covariation is added to any covariation that may exist between the concepts which the items tap. This "extra" covariation – which is correlated error – increases the observed correlation if a positive relation exists among concepts, or decreases the observed correlation if a negative relative relation exists among concepts.[3] Correlated errors may be anticipated with health status measures to the extent that they represent patients' self-reports and share similarities in format.

Clearly, validity depends on the extent of random and non-random error present in the measurement process. As Althauser and Herberlein (1970, p 152) have noted, 'matters of validity arise when other factors – more than one underlying construct or methods factors of other unmeasured variables – are seen to affect the measured variable in addition to one underlying concept and random error. Non-random error prevents empirical indicants from representing what they are supposed to represent: the theoretical concept.'

This said, it should be acknowledged, too, that the dichotomy that classical test theory draws between the "true" score and the error score has come to be regarded as rather simplistic – because the many standard approaches to estimating reliability (intra-observer, inter-observer, test-retest and parallel forms) do not exhaust the possible sources of "confounding". For example, in the present context it is probable that the form of the test (interview, telephone or self-administered), the time of day, or the setting in which the patient responds to questions/items about his/her health status (hospital, clinic, home, etc.) may have an impact on the observed scores (Green & Lewis, 1986; Shortell & Richardson, 1978). This harks back to the point made earlier about evaluating test scores versus tests:

> 'The emphasis is on scores and measurements as opposed to tests or instruments because the properties that signify adequate assessment are properties of scores, not tests. Tests do not have reliabilities and validities, only test responses do. This is an important point because test responses are a function not only of the items, tasks, or stimulus conditions but of the *persons* responding and the *context* of measurement. This

[3] The impact of error in the data is not "predictable" except in the case of bivariate random error. Bivariate random error in the data on one or two variables reduces the correlation between them. Bivariate systematic error in the data on one or two variables will either decrease or increase the correlation depending on the ratio of the covariances or error to the covariances of the true values (Rummell, 1970)

> latter context includes factors in the environmental background as well as the assessment setting' (Messick, 1989, p 14, emphasis supplied).

Given that there are multiple "facets" to any measurement situation, and that some of the variables "sampled" in the course of collecting data may contaminate the results, it seems reasonable that we attempt to identify, and then quantify, the sources of measurement error. This is the premise behind *generalizability theory* (Cronbach, Gleser, Nanda & Rajaratnam, 1972). The advantage of conceptualising measurement in this way is that it prompts the researcher to *test hypotheses*: to design studies in which facets of the measurement situation that are possible sources of error (eg, interviewer, setting, mode of administration) are systematically varied, and to calculate the relative contribution of such sources of variation in adding error to a measurement using analysis of variance (ANOVA) techniques. This approach can lead to specific strategies to reduce the major components of error and to improve measurement. It is consistent with the idea that we (should) seek a more *precise estimate*, rather than a "true" score.

The application of generalizability theory to the field of health status assessment has been very limited, notwithstanding its obvious relevance. We are aware of only three studies in this area (viz. Bremer & McCauley, 1986; Chambers, Haight, Norman & McDonald, 1987; Evans, Cayten & Green, 1981).

3. VALIDATION OF MEASUREMENT

Validity is an evolving concept (Angoff, 1988; Anastasi, 1986; Messick, 1989; Cronbach, 1990). The early focus on validity was pragmatic and largely atheoretical, as exemplified by Guildford's statement that 'in a very general sense, a test is valid for anything with which it correlates' (1946, p 429). Then, as now, validity was regarded as pre-eminent among psychometric concepts. It was also recognized that, unlike reliability, validity is not an invariate characteristic of a test, but specific to the particular purpose. Subsequently, validity was thought to be of several types: criterion-related validity (subsuming the initially separate categories of predictive validity and concurrent validity), content validity, and construct validity. This representation of validity persisted until well into the 1970s (and is current in many textbooks), as did the presumption that, as if by corollary, tests could be validated by any of these three general procedures. Thus, the three types of validities were more or less regarded as alternatives -- though construct validity was seen almost 'as a last resort where analysis of content or predictive power could not support a validity claim' (Cronbach, 1990). This prompted Guion (1980, p 386) to cast the three categories of "validity" tests as 'something of a holy trinity representing three different roads to psychometric salvation.'

The contemporary view is that construct validity provides the basis for a unitary conception of validity. Cronbach has put it variously, and unambiguously, as 'all validation is one' (cited in Messick, 1989, p 18); and the '30-year-old idea of three types of validity, separate but maybe equal, is an idea whose time has gone. Most validity theorists have been saying that content and criterion validities are no more than strands within a cable of validity argument' (1988, p 4). The emphasis now is very much on the meaning or interpretation of measurements, as opposed to the test or observation device per se. Validation is driven by theory and what is validated are the inferences derived from test scores.[4] As Messick (1989, p 13, emphasis supplied) puts it: 'Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment'.

Following from this, testing for validity is about supporting or defending *inferences*, not about demonstrating the psychometric properties of a scale. In Landy's words (1986, p 1186), 'researchers are not really interested in the properties of tests. Instead they are interested in the attributes of the people who take those tests. Thus, validation processes are not so much directed at the integrity of tests as they are directed toward the inferences that can be made about the attributes of the people who have produced those test scores'.

Needless to say, different sources and mixes of evidence can be used to support score-based inferences and different sorts of inferences can be drawn from a given set of test scores. Such inferences can be of the ilk, X is a part of Y; X is an approximation of Y; or X is a sign of Y, corresponding to the labels content, construct, and criterion-related, respectively (Landy, 1986). This is no matter of semantics, of swapping one terminology for another – types of validities for types of inferences. What is important is the switch in the focus of validation from the idea of "collecting alternative kinds of stamps" to hypothesis testing. Inferences are hypotheses. To the extent that all the forms of inference previously accorded the status of a type of validity bear on the valid interpretation and use of test scores, they are complements rather than substitutes. By this reckoning, validity is a unitary concept and an evolving property. It is not 'a commodity that can be purchased with techniques' (Brinberg & McGrath, 1985). It is a matter of degree rather than an all-or-none property, and validation is a never-ending process. Indeed, it can be said that 'as a process, construct validation (1) is never "once considered, forever handled"; (2) always involved multiple studies over many situations and populations' (Green and Lewis, 1986, p 109). Most measures should be monitored continually to see if they are behaving appropriately. The nature of the evidence required will depend on the type of validity and the purposes and circumstances relative to which validity is to be assessed.

The nexus between theory and validation is made quite plain by Green and Lewis (1986) in their clear and concise account of the *process* of construct validation. In the beginning, one must specify what is meant by the target construct. This is not a matter of coming up with a new label; it requires a definition and delineation of the construct. Other concepts to which the target construct is hypothetically related must be identified and the nature of the hypothesized relationships among the constructs described. At the outset one simply assumes (hypothesizes) that an instrument measures a specific concept. To get beyond this, however, empirical testing is unavoidable. Empirical evidence must be brought to bear on the hypothesized or predicted relationships among the constructs. These data are very often simple correlation coefficients which are examined in terms of their magnitude and utility, not just their statistical significance. These initial findings provide confirmation or disconfirmation of the hypothesized relationships. Upon reflection, these data may lead to modifications in one or more of the following: specified constructs; accepted indicators of the underlying constructs; hypothesized relationships among the specified concepts; and the method of obtaining measures of the underlying constructs. This step is therefore part empirical evaluation, part theory construction. Many questions must be addressed: "What concepts account for the respondents' observed test scores?" Is the explanation consistent with the hypothesized relationship? Are there alternative explanations that

[4] The term *test score* is generic and is used to mean any consistent behaviour or attribute observed or documented by any means.

can account for the pattern of test scores? Once this round of results is reconciled with the theoretical base, further empirical testing is called for. And so it continues.

Thus, construct validation can be described as 'a *process* by which you test the theoretical relationships among underlying concepts against hypothesized relationships and then revise the theoretical formulations or the measures accordingly' (Green & Lewis, p 108).

By this account, theory informs and is informed by construct validation. This is in marked contrast to much of the research on health status measurement. Patrick and Bergner (1990, p 175) note reproachfully that,

> At present, researchers tend to approach the relationship among endpoints inductively by collecting data and examining the correlation among measures. Little hypothetical or deductive reasoning is involved in either the selection of measures or the analysis of results. A priori hypotheses and head-to-head comparisons of different dimensions will be important for determining the association between specific disease states or disorders and their behavioral, perception, and social consequences.

Of course, where a given researcher enters the process depends very much on the maturity of the construct (the extent of previous research and validation) and the degree of confirmation of the network of constructs in which the target construct is embedded. The testimony of Patrick and Bergner implies that attention to theoretical relationships between health status and other constructs is long overdue.

Means-Ends Relationships in Health Status Assessment

It has long been recognized that, strictly speaking, one does not validate a measuring instrument; rather one evaluates the use(s) to which the instrument is put. The distinction is central to validation, as that term is now understood, since a measuring instrument may be quite valid for one purpose and almost worthless for other purposes. This is not to say that a measuring instrument may not be useful for a number of different purposes. However, the validity of the measuring instrument must be determined empirically for each purpose. In each case, the question is whether the inferences made are appropriate.

In the literature on the measurement of health status, the significance of the purpose or applicability of health indexes has been emphasized by Kirshner and Guyatt (1985), in particular. They identify three broad categories of health indexes: discriminative, predictive and evaluative. Broadly, discriminative and predictive indexes are distinguished according to whether there is a "gold standard" or other external criterion against which the measure can be validated.

A **predictive** index is defined as one that is used to assign individuals to one of a number of predetermined measurement categories, given that a gold standard is available, either concurrently or prospectively, to determine whether individuals have been assigned correctly. A **discriminative** index, on the other hand, is one that is used to distinguish among individuals and groups on an underlying dimension (ie, defining cross-sectional differences) where there is no external criterion against which the measure can be validated. An **evaluative** index is used to measure the magnitude of the change within individuals over time on a dimension of interest.

The vast majority of health status instruments are discriminative *or* evaluative, rather than predictive – since there is no (recognized) gold standard. A number of health status indexes (eg, the Sickness Impact Profile, and the Nottingham Health Profile) are intended to be used as *combined* discriminative and evaluative health status measures ie, 'appropriate and sensitive enough to be applicable to the assessment of the total health status of population groups and specific enough to permit evaluation of a specific health program directed toward a circumscribed group' (Jette, 1980). Whether their use is valid for these different purposes is an empirical question. Certainly, the supporting evidence should be reviewed carefully since *a priori* 'the requirements of maximizing one of the functions of discrimination, prediction, or evaluation may actually impede the others' (Kirshner & Guyatt, 1985).

In subsequent reports, we will examine the available evidence on the validity of a selected set of health status measures for each of these purposes.

At the most general level, there are two types of evidence on which we can base inferences about the degree to which indicants of health status measure the concept of health status per se, rather than systematic sources of variation, and random error:

1. **Internal association:** the pattern of interrelationships among the indicants designed to measure a theoretical concept.
2. **External association:** the pattern of relationships that exists among indicants designed to measure the theoretical concept and other variables.

By rights, what ought to follow from this discussion is a section on the methodology of hypothesis testing. In fact we have opted to stick with the traditional headings, in part because the 'labels … are not completely useless, nor are they interchangeable' (Landy, 1986, p 1185) and, in part because these terms are still very much part of the vocabulary of validity theorists and appear intact in the applied literature. Our compromise is to underscore the nature of the inference(s) that each "type" of validity may help support and to discuss their relationship to the process of construct validation.

X Is A Part of Y: Content Validity

In its classic form, content validity concerns the extent to which a measuring instrument taps a specific domain of content about which inferences are to be drawn or predictions made. Insofar as the items reflect the full domain of content, they are said to be content-valid. In practice, content-related evidence generally takes the form of consensual judgements that the content of the test is representative of and relevant to a particular domain of interest.

It follows that there are two interrelated steps involved in achieving content validity. The first step is specifying the domain of content; it is essentially a requirement of operational definition and is concerned with content relevance. The second step involves specifying procedures for selecting and/or constructing a representative collection of items. It is concerned with content coverage. Neither of these steps is straightforward. For example, it is not enough to use construct theory as a basis for specifying the boundaries and facets of the domain of reference. The items generated need to be judged relevant to the domain with a high degree of consensus. This raises the question of domain clarity, the extent to which a domain is sufficiently well-described that different researchers working independently produce broadly comparable tests.

In fact, often it is not logically possible or feasible to specify the domain of content. Health status is clearly a case in point. Definitions of health abound. The most widely known global definition of health is that offered by the World Health Organization *viz,* 'health is a state of complete physical, mental, and social well being, and not merely the absence of diseases and infirmity' (WHO, 1958). This definition has been criticised for its simplicity and abstractness by Goldsmith (1972, p 213) who regards the difficulty in conceptualising health as 'perhaps the major constraint on the development and usefulness of health status indicators'. Likewise Jette (1980) noted that many definitions of health are ambiguous and abstract to the point that they resist operationalization. More than a decade later, Whitlaw and Liang (1991) have ventured that such operationalization would be advanced by embarking on a series of studies with the aim of tackling 'the specific conceptual and measurement issues which each of the dimensions of health in the WHO definition' (p 333).

Still, researchers have been largely undeterred by the difficulty in achieving a consensus definition of health that can be operationalized. Beginning with the measures rather than the concept, Ware (1984c) observes that 'a review of the content of published health status and [health-related] quality of life survey instruments reveals substantial overlap in the way these concepts have been defined and measured'. The explanation for this informal consensus is apparently straightforward: 'Although there are many different diseases, the problem of conceptualising and measuring the impact of disease on quality of life is not completely overwhelming because there seems to be a manageable number of concepts to be considered' (p 2317). Patrick and Deyo (1989) provide a summary of the major concepts of health-related quality of life contained in six well-known generic measures of health status, presented in Table 2 below (see also Patrick & Erickson, 1988a; 1988b).

As Table 2 indicates, there is a broad consensus about what dimensions should be included, as might be expected given the WHO definition as a usual point of departure, but by no means total agreement about all "ingredients". As Kaplan (1985, p 96) comments, 'with surprising consistency, authors quote the WHO definition and then present their methods measuring each of the three components of health [physical, mental and social health]. … So prevalent is the notion that health measures must include these three components that many reviews now negatively evaluate any measure that does not conform to the WHO definition.' Of course, these remarks are not unpremeditated: the central point of Kaplan's piece is to challenge the orthodoxy of the WHO definition. In particular, he queries whether "social health" is a meaningful and distinct entity.[5] Conversely, Spitzer (1987b, p 468) states:

[5] The nub of Kaplan's position can be summarized in terms of a number of propositions *viz*.

(i) social health consists of at least two dimensions, social activities and social resources;

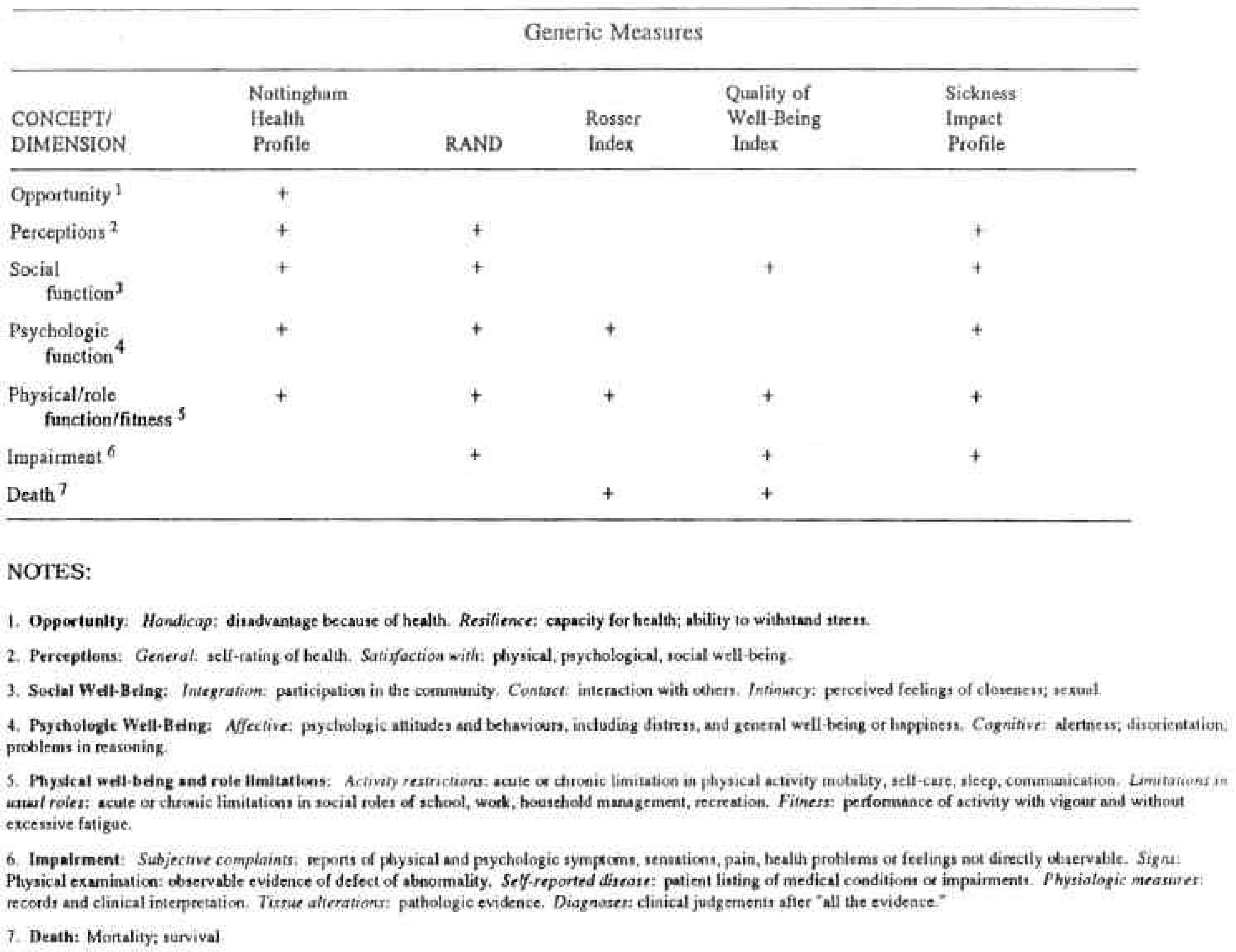
(ii) the social activities portion of social health is included within the measurement of health status; health conditions are important because they disrupt "functioning" or cause premature death;

(iii) since functioning includes social activities it is unnecessary to include social health or social function as a separate component of health status;

(iv) the portion of health missing from the definition of health status is the social resources or social support component of health status;

(v) this being the case, efforts should be directed at identifying the role of social support as a mediator of health status.

Table 2: Major concepts of health-related quality of life contained in selected generic measures



	Generic Measures				
CONCEPT/ DIMENSION	Nottingham Health Profile	RAND	Rosser Index	Quality of Well-Being Index	Sickness Impact Profile
Opportunity[1]	+				
Perceptions[2]	+	+			+
Social function[3]	+	+		+	+
Psychologic function[4]	+	+	+		+
Physical/role function/fitness[5]	+	+	+	+	+
Impairment[6]		+		+	+
Death[7]			+	+	

NOTES:

1. **Opportunity**: *Handicap*: disadvantage because of health. *Resilience*: capacity for health; ability to withstand stress.

2. **Perceptions**: *General*: self-rating of health. *Satisfaction with*: physical, psychological, social well-being.

3. **Social Well-Being**: *Integration*: participation in the community. *Contact*: interaction with others. *Intimacy*: perceived feelings of closeness; sexual.

4. **Psychologic Well-Being**: *Affective*: psychologic attitudes and behaviours, including distress, and general well-being or happiness. *Cognitive*: alertness; disorientation; problems in reasoning.

5. **Physical well-being and role limitations**: *Activity restrictions*: acute or chronic limitation in physical activity mobility, self-care, sleep, communication. *Limitations in usual roles*: acute or chronic limitations in social roles of school, work, household management, recreation. *Fitness*: performance of activity with vigour and without excessive fatigue.

6. **Impairment**: *Subjective complaints*: reports of physical and psychologic symptoms, sensations, pain, health problems or feelings not directly observable. *Signs*: Physical examination: observable evidence of defect of abnormality. *Self-reported disease*: patient listing of medical conditions or impairments. *Physiologic measures*: records and clinical interpretation. *Tissue alterations*: pathologic evidence. *Diagnoses*: clinical judgements after "all the evidence."

7. **Death**: Mortality; survival

> ‘To conclude this section about the nascent consensus concerning what attributes or what constructs we measure with what type of questionnaire, I will point to common features of all the types of generic data gathering instruments I have discussed. In my view, we are not in the domain of either quality of life measurement or health status measurement unless we include physical function, social function, emotional or mental state or mental status, burden of symptoms and perception or sense of well-being. These five groups of attributes seem to be found in most accepted and validated instruments. I would challenge those proposing a scale or index with less [sic] than those dimensions about the content validity of their measure.’

De Groot’s (1986) concerns, on the other hand, are more diffuse and, in a way, even more far reaching. A particular concern for De Groot is the propensity toward an unreflective segue from definition to measurement (and its psychometric sequelae) of health-related quality of life without answering the critical question of what it is that the ensuing instruments are supposed to measure or predict. Thus:

> ‘Curiously, in most research on QL [quality of life] the question of what its assessment is ultimately aimed at is given relatively little attention. It is of course conceded to be a problematic construct on which opinions and theories differ, as do empirical approaches. After a few obligatory statements of this general nature, authors of psychometric studies tend to quickly shift to what they regard as their main business, namely the task of ‘measuring’ QL by means of some preferred operationalization’ (p 67).

The “psychometric shortcut to measurement”, as De Groot refers to this approach, finesses a number of issues which, whilst not central to the present discussion – in part, for the very reason pin-pointed in the above quotation – are very important, nonetheless. One is whether the quality of life value of person P at time T is a momentary feeling of well-being, or whether it is conceived as an attitude, a more or less stable disposition. This is a question that is seldom addressed conceptually and even less so empirically (however, see Moum, 1988, for a discussion of mood-of-the-day effects). Health as well-being is essentially an attitudinal or value concept. We know that “adjustment to illness” is an almost universal phenomenon and the development of scales to measure it is a thriving industry (eg, Arpin, Fitch, Browne & Corey, 1990; Browne, Arpin, Corey, et al, 1990; Derogatis, 1986, Felton & Revenson, 1984; Felton, Revenson & Hinrichsen, 1984; Folkman, Lazarus, Gruen, DeLongis, 1986; McFarlane, Norman, Streiner, et al, 1980; Morrow, Chiarello & Derogatis, 1978; Roberts, Browne, Brown, et al, 1987a; Roberts, Browne, Streiner, et al, 1987b; Viney & Westbrook, 1984). There is evidence, too, that long-term health-related values can differ significantly from current preferences for long-term treatment (Christensen-Szlanski, 1984; Christensen-Szlanski & Northcraft, 1985). At a more general level, it is worth reading the autobiography of the late Alan Marshall, crippled by accident as a boy and writer of exquisite children’s stories for people of all ages. In his late years, he concluded that the richness of his life, consequent upon his reduced physical functioning, was such that, with hindsight, he would again choose the crutches.

Another question worth asking is whether the person P should be regarded simply as a (non-reactive?) “measurement gauge” in a process in which it is his/her task to give ‘relatively simple answers to simple questions’ and it is the researcher’s responsibility to integrate the answers and compute P’s QL-value.

Even within the psychometric tradition, however, there is considerably more that could be done. For example, despite the fact that expert judgement is a key ingredient in attesting to content relevance, systematic attempts to document the consensus of multiple judges are not commonplace in test-construction (Messick, 1989). By and large, the imprimatur of experts seems to be bestowed in an informal manner via a post hoc process. In principle a number of approaches could be used to formalise judgements about content relevance. For example, content experts could rate each item in terms of the degree to which it reflects the dimension of the domain the item is supposed to reflect. Alternatively, content experts could match each item to the domain dimension they think the item best represents. A relevant consideration in applying such tactics to the case of health status measurement is, "who are the relevant experts?" Who should judge the relevance of items purporting to reflect different health status dimensions? consumers, patients, physicians, or care-givers? Any number of researchers (eg, Greenwald, 1987; Hays & Stewart, 1990; Jenkins, Jono, Stanton & Stroup-Benham, 1990; Segovia, Bartlett & Edwards, 1989; Ware, Brook, Davies-Avery, et al, 1980) have used factor analyses and other techniques to explore the "dimensionality" of health status. We are aware of only one published study in the health domain that involves any *analysis* of judges' assessments of whether or not items included in a proposed health status measure reflected the content defined by the corresponding dimensions (Lomas, Pickard & Mohide, 1987).

To address the second facet of content validity, content representativeness, one must know not only the boundaries of the domain but also its logical or psychological subdivisions (Messick, 1975; 1989). As we indicated above, considerably more attention has been paid to exploring the structure of the health-related quality of life than to verifying whether items developed to reflect on a particular dimension do, in fact, do so. With content coverage the concern is with domain sampling, or with whether the items slated for inclusion in the test systematically represent each subdivision (eg, dimension of health-related quality of life). Alternative rules can be specified regarding the actual number of items used to represent each health dimension eg, uniform coverage or inclusion in proportion to judged importance.

The task of *formulating* a collection of items that is broadly representative of a concept is made *ipso facto* more difficult as the number of potential foci increases. The selection of content (eg, dimensions of health and the extent of coverage within each) involves questions about values – decisions about what is relevant and important and should therefore be included. Without an agreed upon domain of content relevant to the phenomenon, there is no prospect of ensuring a random sampling of content, and without that, it is impossible to ensure the representatives of particular items. It is also impossible to specify exactly how many need to be developed to represent any particular domain of content.

The selection of items raises, too, the problem of trading off content, veridicality of responses and measurement. To satisfy requirements for the first two factors, it may be deemed necessary to build in some redundancy; ie, two or more questions that overlap in the particular domain of inquiry. This, of course, yields over-measurement of that content item in response scales and the trade-off between item selection and reduction is not easy to deal with.

An added complication with health status measures is that the purpose for which the measure is being constructed *should* have a significant bearing on item selection and item reduction, too. As Jette (1980, p 568) puts it, 'the number of foci assessed and the extent of coverage within each is determined by … the purpose or applicability of the indicator'. For example, the likelihood that patient status on a particular item will change as a result of an intervention or treatment would be a crucial element of an evaluation instrument. Where the construction of a discriminative instrument is concerned the accent should be on items that are important to patients and are stable over short periods of time (Kirshner & Guyatt, 1985).

These difficulties reveal quite clearly the rather fundamental limitations of the traditional notion of content validity as an arbiter of validity *per se*. In the absence of objective, well-defined criteria, 'inevitably content validity rests mainly on appeals to reason regarding the adequacy with which important content has been sampled and on the adequacy with which the content has been case in the form of test items' (Nunally, 1978, p 93). This problem is associated with another, more fundamental problem: that in content validation, '*acceptance* of the universe of content as defining the variable to be measured is essential' (Cronbach & Meehl, 1955, p 282). For these reasons, it was suggested, even when the Trinitarian approach to the validation of measurement was generally accepted, that content validity is 'basically judgmental and should not be used as the sole criterion of validity' (Messick, 1989, p 40).

The contemporary viewpoint is that 'content validity is not validity at all in the sense shared by the other types, or aspects of construct validity' (Angoff, 1988). As stated by Messick (1975), 'the major problem … is that content validity … is focused upon test *forms* rather than test *scores*, upon *instruments* rather than *measurements*. Inferences … are made from scores, and scores are a function of subject responses' (p 960). Unlike other conceptions of validity, 'content validity gives every appearance of being a fixed property of the *test* … rather than being a property of the test *responses*' (p 959). And, as discussed above, score-based inferences are the foundation of validity. The aim of validity testing is *inferential*.

A measure that includes a more representative sample of the target concept or dimension allows us to make inferences that are broader or more generalizable. If there are important aspects of health outcomes that are omitted on health status measures, we are likely to make some inferences that are wrong. In this case, it is our inferences, not the instruments, that are invalid. For example, knowing a rheumatoid arthritis patient's grip strength does not allow us to make accurate inferences about morning stiffness or joint count, except insofar as these attributes are correlated with grip strength. As Cronbach (1988, p 151) states, content validation stops with a demonstration that a test conforms to a specification; however, the claim that the *specification* is well chosen embodies a CV [construct validity] claim'.

X As A Sign of Y: Criterion-Related Validity

According to Nunally (1978), criterion-related validity 'is at issue when the purpose is to use an instrument to estimate some important form of behaviour that is external to the measuring instrument itself, the latter being referred to as the criterion' (p 87, emphasis supplied). After the criterion has been obtained, determining the validity of the measuring instrument is straightforward. It typically involves correlating scores on the measuring instrument with scores on the criterion variable.

The magnitude of the correlation is viewed as a direct indication of the amount of validity. Indeed, the extent of the correspondence is widely seen as the *only* kind of evidence that is relevant to criterion-related validity. If the correlation is "high", no other standards are necessary. This being the case, criterion-related validity can be described as empirically-oriented and largely atheoretical. It differs in this respect from nomological relatedness, the extent to which different constructs that are related in lawful ways. With criterion validity the emphasis is very much on the usefulness of relationships in applied contexts. As a result there is no criterion or single criterion-related validity coefficient. Rather, there are as many coefficients as there are criteria of interest *vis a vis* a particular measuring instrument – although, in practice, it may well be more difficult to obtain a good criterion than to obtain a measuring instrument. And, it is here that theory tends to be involved, even if only indirectly, since there must be some basis for the selection of the criterion variable(s). This being the case, the practical utility of criterion validation depends as much on the criterion as it does on the quality of the measuring instrument itself. As Cronbach has suggested 'all validation reports carry the warning clause, "insofar as the criterion is truly representative of the outcome we wish to maximise"' (1971, p 488). To the extent that the choice of the predictor test is influenced by hypotheses about the nature of the criterion domain, the criterion-related evidence tends to contribute to the validation of both the criterion and the predictor.

Two types of criterion-related validity are customarily identified, on the basis of when the criterion data become available. If the criterion exists at the same point in time as the measure, then **concurrent validity** is assessed when the correlation between the two sets of scores is obtained. **Predictive validity**, on the other hand, is concerned with a criterion that will become known in the future, which is correlated with the relevant measure. In the first case, it seems legitimate to wonder why it is necessary to introduce a measuring instrument when the criterion or gold standard is readily available. The answer must be that the index offers something the gold standard does not. Perhaps it is quicker, cheaper, less risky or less demanding in terms of the burden it imposes on respondents. With the second-type of criterion-related validity, the advantage of the measuring instrument is more transparent *viz,* timeliness.

As Messick (1988, p 36) indicates, 'the key inference that is sustainable from a statistically significant criterion-related validity study is that there is a dependable relationship in the particular setting between the predictor test or tests and the criterion measure. The inference the practitioner wishes to make, however, is that the test should be used for selection [or classification].' Although the 'proof of the pudding' is often said to be in the correlations where criterion-related validity is concerned, the usefulness of a predictor variable is only partly a function of the strength of the criterion-predictor relationship. In fact, the usefulness of the measuring instrument in a particular applied setting depends on four factors: (i) the strength of the 'true' relationship between the predictor variable and criterion variables; (ii) the selection ratio or location of cut-off point on the predictor variable; (iii) the base rate for success (ie, ratio of successes to failures); and (iv) whether being above the cut-off score on the predictor variable has implications for performance on the criterion variable (Einhorn & Hogarth, 1978; Hogarth, 1980). The significance of each of these factors is examined below. Before we get to this, however, the purposes of testing in applied contexts warrant some amplification.

The reason for testing is very straightforward. Testing is invoked as an aid to making "sorting-type" decisions, involving selection, classification or diagnosis, as a basis for future action. In selection decisions, individuals are either accepted or rejected for a given treatment. In classification decisions, individuals are assigned to one of two or more categories of treatments. Diagnosis-related decisions may involve selecting or classifying individuals on the basis of their current needs (*remedial* diagnosis) and/or predicting which individuals will respond favourably to the available treatment or more favourably to one treatment among a set of alternatives (*readiness* diagnosis). Very often the situation is one where "demand" outstrips "supply" and it is deemed appropriate to make "awards" or allocate resources on the basis of a predictor of "potential to succeed" (according to a criterion). The gist of the classic selection task is represented in Figure 6. Together with Figures 7 and 8 below, it provides the basis for explaining the role played by the above-mentioned determinants of the usefulness of a test.

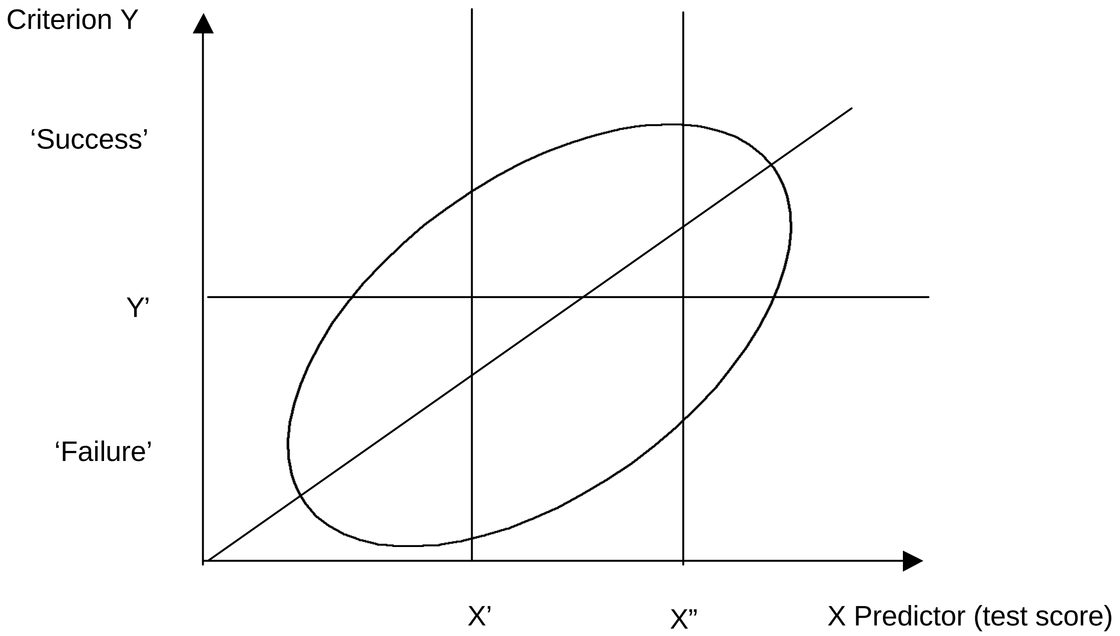
The first of the determinants listed is the 'true' relationship between the test score and the criterion. This relationship, conventionally summarized by the correlation coefficient, is reflected in the shape of the ellipse (covering the full range of scores on the predictor and criterion variables). The narrower the ellipse, the more "predictive" the test – in the sense that higher scores on the measuring instrument are associated with higher scores on the criterion variable. Figure 7 juxtaposes the two extremes: a case where scores on the measuring instrument are perfectly correlated with the criterion (Figure 7a) and a case where the measuring instrument is *not* correlated with the criterion (Figure 7b). In the former case, utilizing scores on the predictor variable as a basis for making decisions is tantamount to random selection, *ceteris paribus*. In the latter case, because the measure is a perfect predictor of the criterion variable, a decision-maker applying a cut-off on the basis of individuals' test scores would identify and select only those who will succeed according to the criterion ie, there would be no "false positives", only "true positives".

Figure 6 depicts the usual case where there is an imperfect positive relationship between the predictor and the criterion. It indicates that, *on average,* candidates with high scores on the measuring instrument should perform better than candidates with low scores. The upshot is that decisions based on test scores are inevitably associated with some true positive outcomes and some false positive outcomes.

Sometimes, the issue is one of *incremental* criterion-related validity ie, what improvement is wrought by using this versus that predictor variable? (Sechrest, 1967). There are applications where the use of a measuring instrument with only a modest correlation with the criterion (eg, correlations of .30 to .40) leads to highly important improvements in the average level of performance.

Of course, the use of correlation coefficients is not without its problems. In most validation exercises, only the coefficients are reported and this is not good enough, simply because the figures *are* averages. A useful point to bear in mind here is that the correlation coefficient is calculated from a linear model. If a set of data points have a distinct non-linear relationship, the pattern will not be well fitted by a straight line (Edwards, 1976; Feinstein & Kramer, 1980). The distribution of data points is important and these are usually neither given nor explained when they are given. The distributions are inherently interesting because low-score and high-score biases can lead to identical coefficients. More importantly, unless the investigators and the reader can be sure that the population is homogeneous with respect to all dimensions that affect responses, the scores may reflect different biases. For example, there is evidence that responses are likely to be affected by age and cultural differences (eg, Aday, Chui & Andersen, 1980; Angel & Cleary, 1984, Angel & Throits, 1987; Baum & Cooke, 1989; Charny, Lewis & Farrow, 1989; Donaldson, Atkinson, Bond & Wright, 1988; Hui & Triandis, 1989; Hunt, McEwen & McKenna, 1986; Hunt & Wiklund, 1987; Lewis & Charny, 1989; Wright, 1986) and low coefficients may be explained by these characteristics. In that sense, correlation exercises can play an investigatory role in developing health status measures.

Figure 6: Continuous test-criterion distribution treated as a dichotomous selection decision



The second determinant of the usefulness of a test is the cut-off score adopted. It also affects the relative number of true positives or "successes" observed. This can be demonstrated simply by observing the effect of applying the cut-off X' versus the cut-off X" in Figure 6. The more stringent cut-off, X", is associated with a lower selection ratio (or "awards" to applicants) and a greater ratio of successes to failures, despite the fact that the strength of the relationship between the predictor and the criterion variables is identical (as measured by the shape of the ellipse). Needless to say the location of the cut-off itself may be affected by the values we attach to different outcomes e.g., true negatives versus false positives.

The third factor is similar to the second. It, too, concerns the location of a cut-off, this time in relation to the criterion variable. The lower the criterion for success the more successes there will be for a given correlation between test and criterion and cut-off score on the test. This factor is known as the base-rate of success.

Figure 7a: Perfect relationship between predictor and criterion variable

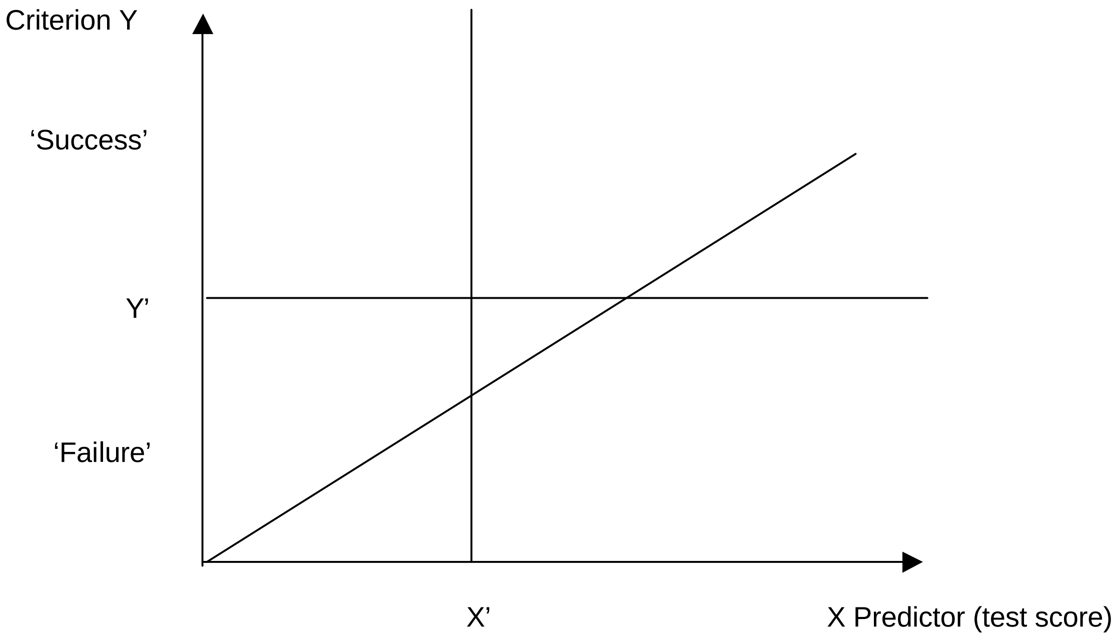
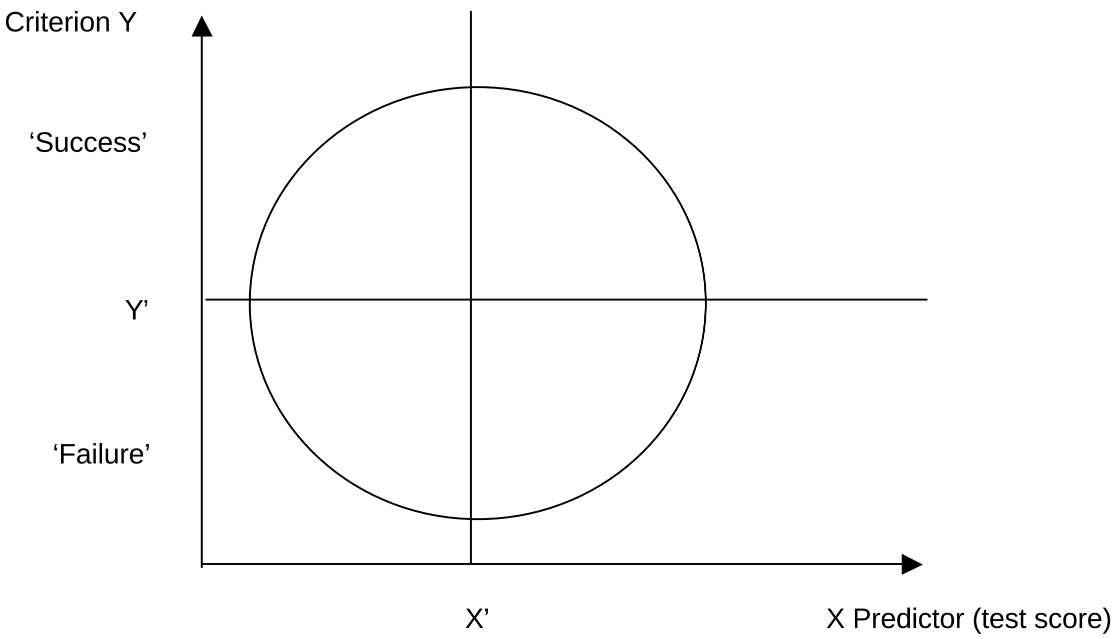
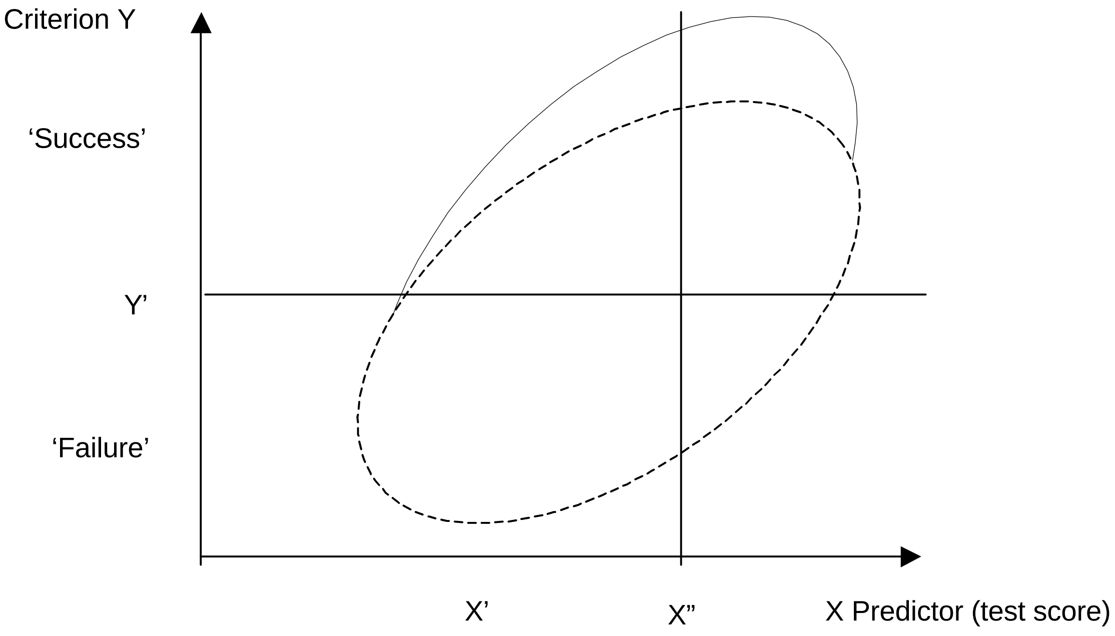


Figure 7b: Lack of relationship between predictor and criterion variable



The fourth factor is more subtle, and pertains to the possible implications of being selected (or not selected). It is often the case that some “treatment” intervenes between the test and the criterion, such that performance on the criterion is above (below) that which would be predicted solely on the basis of the test-criterion relationship. The question is whether there are “real” consequences associated with being accepted versus rejected. Are resources bestowed that cause recipients to perform better than they otherwise would? Figure 8 shows a positive “treatment effect”: there is an upward impetus on criterion outcomes for those individuals whose test scores lie above the cut-off X’.

Figure 8: Selection decision where individuals accepted receive special treatment thus inflating criterion variable scores relative to the “true” relationship between predictor and criterion variables (as reflected by ellipse with broken line)



Source: Adapted from Hogarth, 1980

Thusfar, we have followed the practice of using the correlation coefficient as an index of criterion-related validity. In fact, as Messick (1989) points out, there are sound reasons to rely on regression equations instead – provided, of course, we are dealing with interval-scaled variables. In addition to providing information about the strength of the test-criterion relationship, regression analysis conveys information about test and criterion variance and about mean levels and standard errors of estimate. Further, regression slopes and errors of prediction are more stable across groups. Unlike correlation coefficients, regression coefficients are not subject to attenuation due to criterion unreliability and restriction of range due to selection. Another reason for preferring regression equations over correlation coefficients is the greater statistical power of differential prediction studies (regression systems) vis a vis differential validity studies (correlation coefficients).

Applications of criterion-related validity involving the measurement of health status have generally been as scarce as the proverbial hens' teeth. We discuss some of the reasons why below. It is, however, worth nothing one exception, namely a study by Kaplan (1987) in which the causal direction of much health-related quality of life research was reversed. Kaplan enquired whether patient reports of health status are predictors of physiological health in chronic disease, rather than vice-versa. This study seems to us to invoke the concept of criterion-related validity. Briefly, her results indicate that, although the physiological measures at baseline were the best predictors of the physiological measure at follow-up, explaining 36% of the variance in blood sugar and 24% of variance in blood pressure at follow-up, the inclusion of the survey measures of functional limitations and perceived poor health resulted in a 34% improvement in the prediction of follow-up blood sugar and a 42% improvement in the prediction of follow-up blood pressure, respectively. It therefore focuses on the increment in criterion-related validity that is achieved by adding predictor variables to the regression equation. The study is exploratory, and not without its methodological problems (as detailed in Patrick's 1987 commentary). However, it is likely to be among the first of a genre of studies to explore the interactions between aspects of perceived health status and physiological health outcomes with a view to identifying patients at risk for poor health outcomes in the future and improving the short-term clinical management of patients with discrepant physiological measures and perceived health status (see also Ganz, Lee & Siau, 1991; McClellan, Anson, Birkeli & Tuttle, 1991).

Kaplan's research is also important because it, too, raises the question of whether what is measured in studies of health-related quality of life is quite independent of disease, or characteristic(s) of the individual. The question is whether 'perceived health status, like happiness, is both a state and a trait' (Patrick, 1987, p 39S; see also De Groot, 1986; Donabedian et al, 1987; Miettinen, 1987). Needless to say, the issue and its resolution are very much in the province of construct validation.

Traditionally, the most important limitation of criterion validation procedures is the non-availability of relevant criterion variables. As a general rule, it can be said that the more abstract the concept the less likely it is that one will find an appropriate criterion. And so it is with the concept of health status (Kaplan's study notwithstanding). Indeed, not only is it usually said that there is no criterion available for the validation of measures of health status, but very often in health care selection tasks (such as choosing which patients in the intensive care unit to afford priority or treat aggressively, given limited resources), a measure of health-related quality of life is much sought after *as* the criterion. Certainly, physicians working in intensive care reportedly must often make decisions without knowing anything of the relationship between their *judgements about a patient's likely future quality of life* and the actual quality of life a patient will experience (ie, the criterion) or without having the (ethical) latitude to "experiment" and obtain feedback by varying their thresholds for treating further versus not treating further. In such instances the criterion and the measure could be, alternatively, the patient's expected quality of life as judged by the physician and as assessed by some health instrument. The possibility of switching back and forth between measure and criterion makes the need for a construct-validated criterion quite salient. Cronbach (1990, p 151) puts it this way: 'In criterion-related validation we generally should inspect the criterion for contaminants and missing ingredients. That is, CV [construct validity] of the criterion is wanted.'

An important issue in setting the research agenda for the future is whether it is a satisfactory state of affairs for researchers in the field of health status assessment to shy away from the notion of criterion validation. As we suggested above, Kaplan's study may represent something of a precedent in this respect. Further, Spitzer (1987a: see Donabedian et al, 1987) has recently advocated that we retire the hackneyed observation that we have no gold standard and adopt instead a gold alloy standard. He questions whether after ‘about two decades after we really got going in this area, [it is tenable that] we are *still* saying there is no gold standard’ (p 188). He suggests that either someone should be given the remit (and resources) to set up a reference laboratory to develop a gold standard that can be used by the research community for the purpose of criterion validation, or we should adopt a gold alloy. He argues that, without this step forward, we are doomed to engage in construct validation exercises over and over again.

X Is An Approximation of Y: Construct Validity

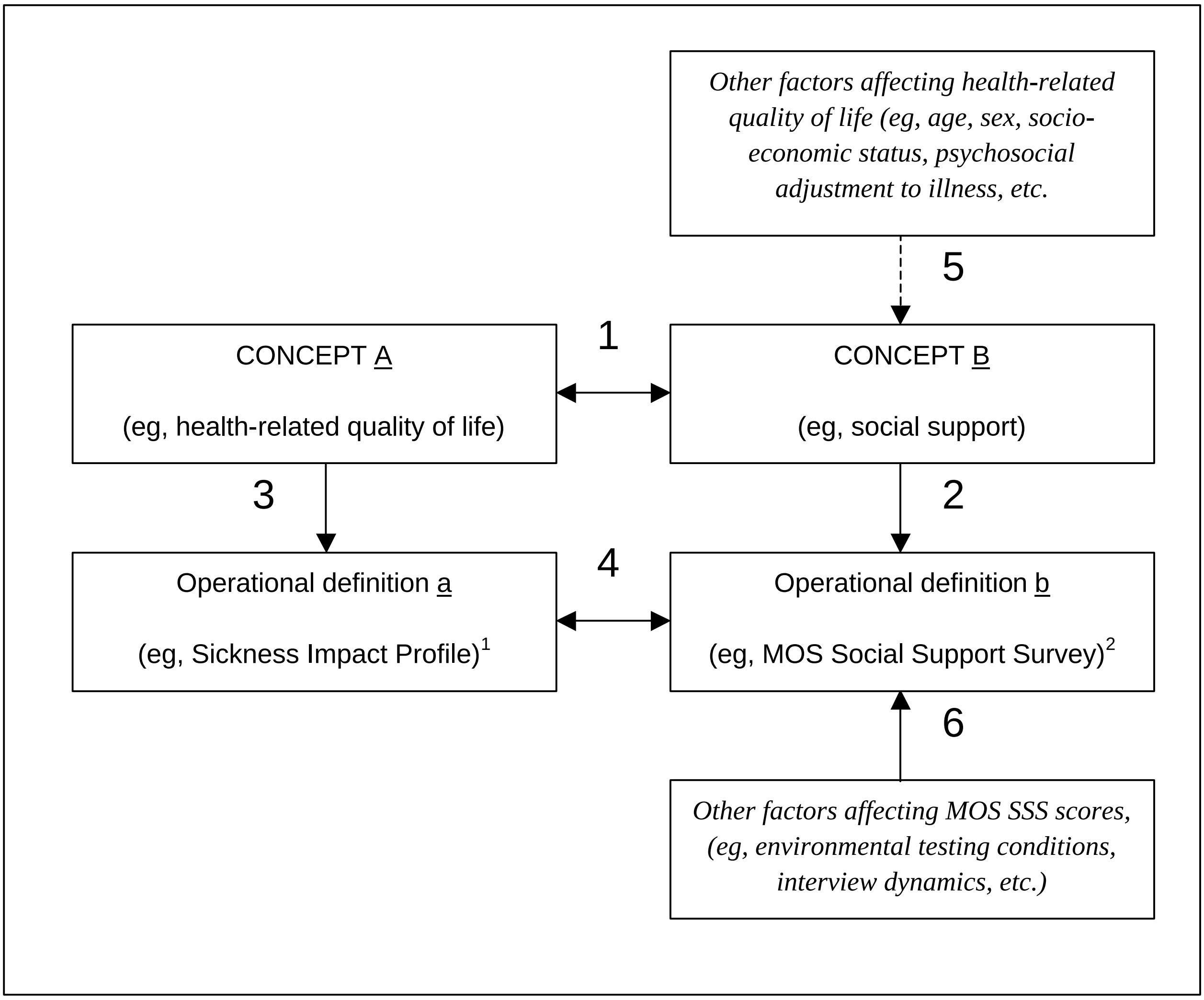
It is an accepted rule-of-thumb that the degree to which it is necessary and difficult to validate measures of theoretical concepts is proportional to the degree to which they are concrete or abstract. In the preceding discussion, we have suggested that the traditional concepts of criterion-related validity and content validity are of limited usefulness *per se* where abstract concepts like health status are concerned. The burden of proof that health status measures measure health status rests with construct validation. As early as 1955 Cronbach and Meehl noted, ‘construct validity must be investigated whenever no criterion or universe of content is accepted as entirely adequate to define the quality to be measured’ (p 282). Today’s Unitarian approach to the validation of measurement merely underscores the point.

Cronbach and Meehl also specify the pre-conditions as follows: ‘Construct validation takes place when an investigator believes 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’ (1955, p 290, emphasis added). Thus, construct validity is evaluated within a given *theoretical* context. More particularly, construct validity is concerned with the extent to which the relationship between a particular measure and other measures are consistent with theoretically-derived hypotheses about the concepts (or constructs) being measured.

According to the most prevalent point of view, construct validation involves three identifiable steps: (1) specifying the domain of observables related to the construct; (2) determining the extent to which the observables tend to measure the same thing or different things; and (3) subsequently doing studies of individual differences and/or controlled experiments to determine whether *presumptive* measures of concepts yield the kinds of results that are predictable on the basis of acceptable theoretical hypotheses concerning the construct. Aspect 3 has to do with determining whether expected correlations between the *presumptive* measure of the construct and measures of other constructs are obtained and/or whether the measure in question is affected in expected ways by experimental treatments.

Evidence in respect of aspect 3 *accrues* from many studies. As Zeller and Carmines (1980) explain, ‘construct validity is not established by confirming a single prediction on different occasions or confirming many predictions in a single study. Instead, construct validation ideally requires a pattern of consistent findings involving different researchers across a significant portion of time and with regard to a variety of diverse but theoretically relevant variables. Only if

Figure 9: The construct validity dilemma


Other factors affecting health-related quality of life (eg, age, sex, socio-economic status, psychosocial adjustment to illness, etc.
5
1
CONCEPT A
(eg, health-related quality of life)
CONCEPT B
(eg, social support)
3
2
4
Operational definition a
(eg, Sickness Impact Profile)1
Operational definition b
(eg, MOS Social Support Survey)2
6
Other factors affecting MOS SSS scores, (eg, environmental testing conditions, interview dynamics, etc.)

Source: Adapted from McGrath, 1982.

Notes:

1. The Sickness Impact Profile (SIP) focuses on the sickness/dysfunction end of the health status continuum and encompasses physical and psychosocial dimensions (see, for example, Bergner, 1987).

2. The Medical Outcomes Study (MOS) social Support Survey is a brief, multidimensional, self-administered measure of functional support (see Sherbourne and Stewart, 1991).

and when these conditions are met can one speak with confidence about the construct validity of a particular measure' (p 82).

It may be argued that there is a logical fallacy involved in claiming that evidence of external associations such as that discussed above constitutes "proof" of construct validity. Following the logic of construct validity, it is reasoned that if concept A and concept B should be related from a theoretical standpoint, the measures that are designed to represent these concepts should be related empirically.

There are a number of weaknesses in this conception of validity (see Runkel & McGrath, 1972) that can be readily explained with the aid of Figure 9. First, its manifestation depends on the validity of measure b, the operational definition of concept B. The inferred validity from measure a to concept A can be no better than the validity from measure b to concept B. Second, the determination of the construct validity of measure a is confounded with a test of the validity of the theoretical relation between A and B (for example, health status and social support). Third, even if A and B are equivalent and the mapping from b to B is valid, demonstrating the validity of the link between a and A depends on the absence of confounding factors with respect to concept B and therefore measure b. Confounding factors will obscure the relation between measures a and b, and therefore the *apparent* validity of a as a measure of concept A.

The technique of construct validation is an attempt to confirm link 3 (the relationship between, say, the Sickness Impact Profile, or a, and health-related quality of life, of A) by assessing link 4 (the relationship between, say, the Sickness Impact Profile, or a, and, say, the MOS Social Support Survey, or b). It is affected by link 2 (validity of the operational definition b of the concept B), by link 1 (validity of the theoretical relation of concept A to concept B), and by links 5 and 6 (effects of other factors on concept B and measure b). Thus, the researcher is able to test the hypothesis that a is a measure of concept A only by *assuming* that the A-to-B and b-to-B links are strong. This involves circular reasoning. From the stand-point of inductive logic it is plain that this paradigm for determining construct validity does not hold water. The only thing that can be validly tested is whether link 4 holds up (i.e., whether measure a correlates with measure b). *If* the assumptions made about the links are correct, then (and only then) the actual correlation between a and measure b permits a valid inference regarding the truth of link 3, that measure a measures what it is supposed to measure.

The assumptions underlying the paradigm are placed on a firmer footing to the extent that the domain of content associated with the other construct (e.g., concept B) is both well-defined and highly restricted. This makes it safer to assume that measure b validly represents concept B. In the limiting case where the "other" constructs are particular observable variables, it is possible to translate the hypothesis "b is related to B" into the assumption "b is B." Then if the assumption that A *relates* to B is quite safe, an empirical correlation between a and b will support the inference regarding the construct validity for the measurement of A with a. From this point of view, studies of construct validity should be undertaken only when, (1) the domain of the "other" constructs is will-defined and (2) and assumption of a relationship between the two concepts is irrefutable (Nunally, 1978).

Up to this point we have considered only two concepts, A and B, and their corresponding measures, a and b. Hence, the foregoing discussion applies equally well to the assessment of criterion-related validity. If instead we conceive of validity as the degree to which a measure ties into a *network of related concepts* then we are concerned with construct validity. Construct validation is of necessity a multivariate approach to validation.

A more “airtight” approach to assessing construct validity, advocated by Nunally (1978), involves determining the internal structures and cross structures in respect of sets of measures concerning observables. Thus set A comprises measures of particular observables a_1, a_2, a_3, etc and set B comprises measures b_1, b_2, b_3, etc. Construct validation, then, consists of the following steps. First, a network of probability statements is formed among the different measures in set A, and likewise for set B, on the basis of a series of empirical studies. For example, if individual differences on the different measures within a set are correlated with one another, one can make probability statements concerning scores on measures a_1, a_2, and a_3. Given the correlations among individual observables, it is then possible to deduce correlations between different combinations of measures in the set (eg, the correlation between any particular measure in the set and the sum of all measures included in the set). The information that is learned about correlations among the measures of observables in a particular set from the accumulation of empirical evidence is referred to as the *internal structure* of the elements in the set. It may provide support for retaining the set as originally defined (where all measures tend to measure much the same thing) or for subdividing the original set A into, say, two subsets (where two things are being measured by members of the set). Alternatively, if all the correlations among the members of a set are very low then it is meaningless to regard the measures as a set, in which case the researcher’s only recourse is to focus on other sets of variables.

The *internal consistency* rationale is a direct extension of the idea that validity can be assessed in terms of the concordance or convergence of the results of *different* operations used to measure the same thing. It simply takes each item or dimension to be an alternative operational definition or measure of the concept in question and asks about the degree to which they yield concordant measurements. This parallels the use of internal consistency as a method for assessing reliability or repeatability as assessed by the correlation between two different applications of the *same measure*. Here, internal consistency is used as a method for assessing the homogeneity or unidimensionality of the components of the measuring instrument.

When internal consistency has been determined for both set A and set B, the *cross structure* between variables in the two sets should be examined. Assume that a particular variable a_1 in A is correlated with a particular variable b_1 in set B. Depending on the size of the correlation, it is possible to make probability statements about unknown correlations between any other member of A and any other member of B. For example, if a_1 and a_2 are highly correlated and b_1 and b_2 are highly correlated then finding a high correlation between a_1 and b_1 allows us to make a prediction about the correlation between a_2 and b_2. Likewise, if the sum of all variables in set A is known to be highly correlated with the sum of all variables in set B, it is possible to estimate the correlation between any particular variable in A and B or the correlation between any two combinations of variables from sets A and B. To the extent that the cross structure between two variables is also satisfactory, there is *circumstantial evidence* for the usefulness of a new measurement method.

The Multitrait-Multimethod Paradigm

Construct validity emphasises two interrelated sets of relationships in respect of a test: (1) that between the test and different methods of measuring the same construct or trait, and (2) that between measures of the focal construct and exemplars of different constructs which are predicted to be related to it on theoretical grounds. The theoretically relevant internal consistencies in the first set have been called *trait validity,* and those in the second set are called *nomological validity* (Campbell, 1960; Cronbach & Meehl, 1955). Trait validity is concerned with the fit between measurement operations and conceptual definitions of the construct, with the meaning of the measure as a reflection of the construct. It is 'the extent to which a measure relates more highly to different methods for assessing the same construct than it does to measures of different constructs assessed by the same method' (Messick, 1989, p 46). The basic notion is that a construct should be neither redundant with other constructs nor tied to a particular method of measurement. Nomological validity, on the other hand, is concerned with the fit between observed data patterns and theoretical predictions about those patterns, with the meaning of the construct as reflected in its relational properties and implications (Messick, 1980; 1989). The basic idea is that the "theory" of the construct being measured should provide a basis for deriving testable linkages between the test scores and measures of other constructs.

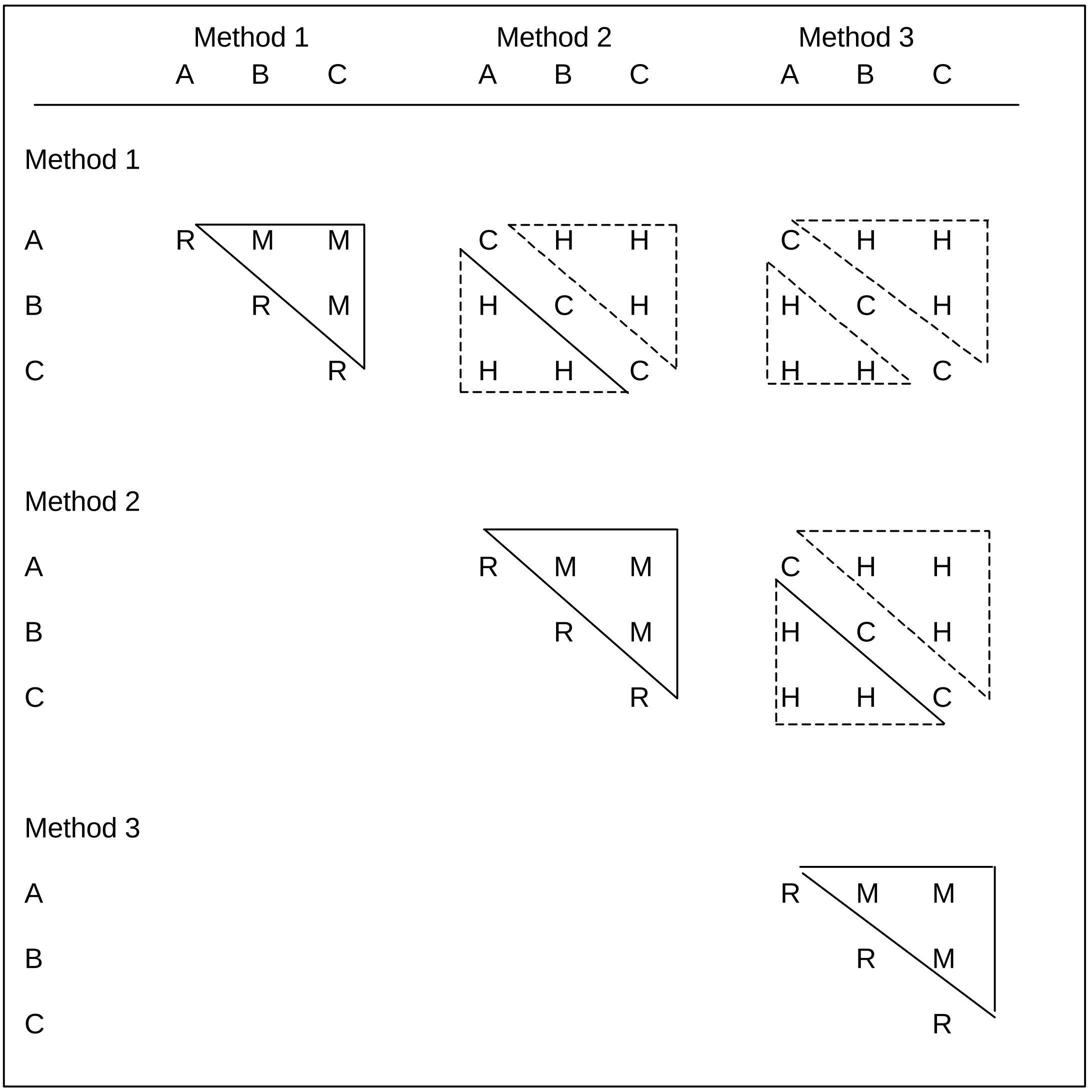
Campbell and Fiske (1959) suggested the multitrait-multimethod matrix (MTMM) as a means of assessing validity. This approach is now regarded as the quintessential construct validation design. It highlights the need for both convergent and divergent evidence in both trait and nomological validity. The basic terms are familiar: *constructs* (for attributes of individuals), called traits in the original exposition of the MTMM approach; and *methods* (for observations of such properties). The MTMM approach recognizes the obvious, as it were: that social science needs to be able to evaluate *both* its constructs and its methods – because it is replete with imprecise or fuzzy constructs that often derive from lay thinking, and very often relies on methods (e.g., observational judgements) that can be affected by extraneous factors (Fiske, 1982). The basic proposition is that different methods of measuring a construct should show some degree of reproducibility and that these methods should discriminate appropriately between pairs of constructs.

Froberg and Kane (1989) urged that the MTMM approach be applied to preference measurement noting that 'so far, no one has attempted to demonstrate discriminant validity' (p 681) of preferences for health states. The imperative for so doing is manifest given the propensity in some quarters to "take the numbers and run". The findings of the first study to use MTMM analysis (Hadorn & Hays, 1991) for this purpose, that 'substantial method variance and little valid trait variance was observed for [health-related quality of life] HRQOL preferences', should add momentum to the normative position. Even more disappointing is the fact that few studies have used the MTMM approach to measure the construct validity of self-reported ratings of health-related quality of life. The number of studies to do so can probably be counted on the fingers of one hand. Two of the more conspicuous and thorough examples are Read, Quinn and Hoefer (1987) and Deniston, Carpentier-Alting, Kniesley, et al (1989).

A MTMM matrix consists of all the intercorrelations resulting when each of several traits or constructs is measured by each of several methods. Each test or task used for measurement purposes constitutes a *trait-method unit*, a combination of trait-related content and a set of measurement procedures that are not specific to that content. The systematic variance among scores may be due to the measurement-related factors as well as to trait content. For example, two methods may evoke the same response set, eg, social desirability.

The MTMM matrix provides a systematic framework for examining two types of validity that fall under the heading of construct validity and for relating them to the concept of reliability. To illustrate the validation process, assume there are three different traits (A, B and C), each measured by three methods (1, 2 and 3), generating a total of nine measured variables. It is convenient to have labels for various regions of the matrix, as per Table 3. The cells on the main diagonal of the matrix (denoted R in Table 3) are correlations between independent measurements of the same construct using the same method (i.e., monotrait-monomethod values). The symbol R denotes reliability or repeatability. Cells below and on the left of the R diagonal are blank because the matrix is symmetrical. Adjacent to each reliability diagonal is the *heterotrait-monomethod* triangle. The correlations in this triangle, labelled M, reflect *method variance* ie, the extent to which there is concordance when the same method is used to measure two different traits. Together with the reliability diagonal these correlations make up the *monomethod block*.

Table 3: MTMM matrix for three traits (A, B, C) and three methods (1, 2, 3)



	Method 1			Method 2			Method 3		
	A	B	C	A	B	C	A	B	C
Method 1									
A	R	M	M	C	H	H	C	H	H
B		R	M	H	C	H	H	C	H
C			R	H	H	C	H	H	C
Method 2									
A				R	M	M	C	H	H
B					R	M	H	C	H
C						R	H	H	C
Method 3									
A							R	M	M
B								R	M
C									R

KEY:

The three validity diagonals are boldface. Each heterotrait-monomethod triangle is enclosed by solid lines and each heterotrait-heteromethod triangle is enclosed by broken lines.

R reliability (same method, same traits);
C trait convergence (same trait, different methods);

M method variance (same method, different traits);
H (different traits, different methods).

There are three heteromethod blocks. Each one is made up of a validity diagonal (which are also known as heteromethod-monotrait values) and two heterotrait-heteromethod triangles located on either side. Note that the two heterotrait-heteromethod triangles are not identical.

Entries on the validity diagonal are labelled C to indicate what Campbell and Fiske call *convergent validity*. They are correlations between two *different methods* of measuring the same trait and reflect the degree of concordance between two operational definitions of the same trait. The heterotrait-heteromethod correlations, labelled H, indicate the extent of the relationship between one measure of one trait and different measure of another trait.

Campbell and Fiske urge us to interpret the entire pattern of correlations in this matrix – the relative magnitudes of R, M, C and H. To begin, we note that the R correlations – same trait, same measure – set the upper limit for other correlations in the matrix since a measure of a trait must correlate at least as highly with itself as with any other measure. Beyond this, they enumerate four aspects of the interrelationships among the correlations that bear on the question of validity (1959, pp 82-83):

1. The entries in the validity diagonal, C, must be 'significantly different from zero and sufficiently large to encourage further examination of validity' (1959, p 101). Given high R correlations, the magnitude of the C correlations (different methods, same trait) indicates the degree of *convergence*. Convergent validity is the extent to which variation in the measure *is* the result of variation in the trait. It involves confirmation of a relationship by independent measurement procedures.

Note that it is necessary for the multiple measures to be of different types, so the weaknesses of any one type of method are countered by coupling it with other methods that have different weaknesses. *Thus, employing multiple operational definitions of a construct, all of which involve the same type of method, is not particularly useful in establishing construct validity. They can establish reliability, but make no contribution to the validation process per se.*

This conclusion follows from our earlier discussion of the concepts of reliability and validity. Although reliability and validity are two different characteristics of measurement, lying at two ends of a continuum, they shade into each other at points in the middle. In our methods of assessing reliability and validity we examine the relationship between two measurements. In the case of reliability, the two measurements come from the same instrument. In the case of validity, the two measurements come from different instruments. This seems very much like a clear-cut distinction which ought to make it obvious that reliability is different from validity. However, to repeat the point made earlier, if we recognise that the "different" instruments used to assess validity may have varying degrees of "difference", and the "same" instrument used to measure reliability may have varying degrees of "sameness", reliability and validity estimates can be viewed as lying on a continuum.

At one end of the continuum are correlations between identical methods of measurement, as indicated by the R entries in the MTMM matrix. At the other end of the continuum are correlations between very different methods of measuring the same variable, as reflected in the C correlations. Finding "maximally different" methods of measuring the same variable is difficult – because the more the methods differ, the less likely it is that they will tap the same variable. However, the consequences of not identifying suitably independent methods is also clear: one is operating towards the reliability end of the spectrum, rather than addressing the question of convergent validity.

2. The C correlations should be higher than the H correlations in the heteromethod block. That is, the validity value for a variable should be greater than the correlation between that

variable and any other variables that have neither the trait nor the method in common, if the traits are independent and the methods are independent.

3. A variable should correlate more highly with an independent attempt to measure the same trait than with measures that use the same method to get at different traits. For a given variable, this involves comparing the C correlations and the M correlations. The M correlations (same method, different traits) indicate the extent to which correlations among measures in the matrix are artefacts of a particular measuring instrument. The *differences* between the C correlations and the M correlations provide evidence of the *divergent validity* of the trait.

The idea here is to define the boundaries of the construct by demonstrating a *lack* of correlation of measurements of the construct with methodologically similar measures of substantively different constructs. That is, the variation in measurements obtained should *not* be the result of variation in other constructs of little or no theoretical interest.

If the M correlations are large and represent a substantial fraction of the C correlations then, even if there is high concordance among alternative operational definitions of a trait (ie, high C correlations), this concordance may not be accepted as evidence of validity for this trait. This is because high M correlations indicate that the methods give more or less the same result regardless of the supposedly different traits to which they are applied. In sum, what is required is high C correlations *and* low M correlations.

4. The same pattern of trait interrelationships must be shown in all the heterotrait triangles of the monomethod and heteromethod blocks. The last three criteria (1-3 above) provide evidence of *discriminant validity*. Discriminant validity must be established when the domain of content is not unidimensional, as, for example, with the construct health status. Thus, we may seek to show that functional status, social functioning and emotional status are different constructs by correlating the measurements with one another and the correlation is lower than correlations between measures of the same construct obtained via different methods (eg, observations by a proxy for the patient versus the patient's responses to a self-administered questionnaire). In the event that the measures of different constructs correlate too highly, the correlations between items should be checked within *and* between clusters. Items that correlate better with another cluster should probably be transferred to that cluster rather than the one to which they were assigned initially.

Simultaneously testing hypotheses about the relations between constructs and considering convergent and discriminant validity presents the researcher with a dilemma. McGrath (1982, pp 97-98) describes it well:

> 'When two measures are very similar in form and substance, we tend to think of them as alternative forms of the same measures and, if they correlate highly, regard that as evidence of the reliability of that construct. If two measures differ in form but are similar in substance, we might well regard them as alternative measures of the same construct and regard their correlation as evidence of convergent validity. But if two measures differ in substance, we are not altogether sure how to regard them. If they fail to correlate, we might regard that lack of correlation as evidence of discriminant validity of one of the constructs. But if they correlate highly, or even moderately, we might regard that correlation as evidence for a relation between two different constructs, as in substantive hypothesis testing. This set of considerations reminds us that 'same' and "different" decisions are made at several levels within the research process, and that they are

> arbitrary. If two measures are too similar, their high correlation is not remarkable, and is regarded as 'merely' a reliability. If two measures are too dissimilar, and they don't correlate, that is regarded as remarkable, and we often take it to be evidence supporting some substantive hypothesis.'

This leads to the paradox that one person's method variance is another's substantive finding.

Although the MTMM strategy is the standard approach to construct validation, its execution has become something of a ritual exercise. As Cronbach (1990) observes, 'a CV [construct validity] study in a journal most frequently consists of cross-trait and cross-method correlations laid out *a la* Campbell and Fiske. Although the MM [multitrait-multimethod] matrix originally rested on subtle reasoning, in most applications the meaning of MM degenerates to "mindless and mechanical". Conclusions are pumped out with no thought to the construction[6] being tested' (p 156). This complaint brings to mind a prevalent oversight: not following one of the main guide-lines for constructing a MTMM matrix *viz.* select at least one trait or attribute that is believed to be independent of the others but which is nonetheless *conceptually related to* the other traits. The rationale is simple: strong validation is preferable to weak validation. Evidence discounting the redundancy of a construct is compelling only if it is pitted against a closely related or rival construct. There are no prizes for distinguishing chalk from cheese!

Though more a reason than an excuse, the explanation for the mechanical application of the MTMM strategy seems to stem, in part, from the fact that Campbell and Fiske did not provide a method for quantifying the degree to which the requirements they specified were met – and judgements based on the visual inspection of zero-order correlations are necessarily qualitative (Jackson, 1969). Such problems of interpretation are compounded by other possible shortcomings of the MTMM approach, including inadequate sampling of individuals, variations in the reliabilities of individual measures and variations in restriction of range across constructs. Another problem is that the Campbell and Fiske criteria for evaluation of the MTMM correlations are incomplete. As has been pointed out (Althauser & Herberlein, 1970; Krause, 1972) they implicitly assume that: (i) there are no correlations between trait and method factors; (ii) all traits are equally influenced by all method factors; and (iii) method factors are uncorrelated. All of this does not gainsay the fact that, in clinical contexts in general and in relation to the validation of health status measures in particular, the principal stumbling block with MTMM approach is likely to be more immediate and practical *viz.* the expense, time and respondent burden involved in MTMM data collection. Almost inevitably, the data collection will be more modest (eg, measuring only one trait with multiple methods or measuring multiple operational realizations of a trait with only one method). However, more concerted efforts to follow the spirit of MTMM studies may well be worthwhile in providing strong argument for discarding many of the extant measures and concentrating on the refinement of those showing most promise – thus following the advice of Spitzer (1987a, p 188).

A second poser applies even if a full MTMM correlation matrix is obtained *viz.* how should it be analysed? A variety of data-analytic procedures have been advanced and more or less discarded, including the analysis of variance (ANOVA) paradigm (eg, Stanley, 1961), a partial correlation method (eg, Schriesheim, 1981) and exploratory factor analysis (eg, Tucker, 1966). Most recently, maximum likelihood confirmatory factor analysis (CFA) has been put forward as a technique that allows less ambiguous interpretation of complete and incomplete MTMM designs

[6] Cronbach uses the term "construction", in preference to the more formal notion of a "theory", to refer to the 'loose assembly of concepts and implications used in typical test interpretations' (p 152).

(eg, Cole, 1987; Schmitt and Stults, 1986). The applicability of covariance structure models, which "extend" the confirmatory factor model by incorporating structural relations among latent (or unobserved) variables, and permit an examination of the relationships among constructs like health status or health-related quality of life and psychosocial adjustment to illness, coping ability and social support, is also gaining recognition (eg, Bentler, 1990; Labuhn, 1984; McSweeny and Labuhn, 1990; Newcomb & Bentler, 1987). The general approach in these path or structural models involves estimating the parameters of a set of structural equations (typically assumed to be linear) that represent hypothesized "cause-effect" relationships. However, as Bergner (1990) points out, the potential of latent variable structural equation models is unlikely to be fulfilled so long as it remains a blackbox that is inaccessible to methodologically-sophisticated but mathematically unsophisticated researchers.

In very general terms, 'factor analysis is a statistical procedure for uncovering a (usually) smaller number of latent variables by studying the covariation among a set of observed variables (Long, 1983). To make the discussion more concrete, suppose we have interviewed a sample of end-stage renal disease patients on different dialysis modalities, using a "pool" of outcome measures pertaining to health-related quality of life and reintegration into daily living. Consistent with the broad definition of health promulgated by the WHO, suppose we include items aimed at assessing end-points in the following areas: physical function, fulfilment of social roles (including paid or unpaid employment), emotional status, interpersonal relationships, cognitive function, social activity and economic circumstances. At one extreme, we could use factor analysis as an expedient way to ascertain the minimum number of hypothetical factors that account for the observed covariation, and as a way of exploring the data for possible data reduction. This form of factor analysis is *exploratory* because the preferred result, the identification of a limited number of coherent and relatively independent factors or dimensions, is data- rather than hypothesis-driven. The researcher does not specify the structure of the relationships among the variables in the model. At this stage, factor analysis is basically "inductive fishing". If the assumptions necessary to estimate the model's parameters are substantively appropriate, the usefulness of factor analysis as an exercise in construct validation depends very much on what steps are taken by way of follow up. As Comrey (1978, p 657) explains,

> 'The usual procedure in interpreting factor results is to inspect the variables that have high loadings on the factor, look for what they have in common, and then name the factor in accordance with the common elements. Up to this point, the activity is little more than factor naming, and if nothing beyond this is done, the value of the analysis may be rather limited. When a name is given to a factor, a hypothesis has been formulated. Untested hypotheses usually have limited value until something is done to test them. Ideally, … the investigator will make plans to carry out additional analyses in which he adds new variables which should have major loadings on certain specified factors and low loadings on other factors if his hypotheses are correct. He will perhaps revise other variables in ways that predict certain outcomes if his factor interpretations are correct. These predictions will be tested by further investigations. Experiments may be carried out, with predictions being made as to the outcome, in which attempts are made to alter scores on one factor but not another. Results of the experiment will confirm or disconfirm the hypothesized factor meanings. In other cases, predictions may be made about how scores for a certain factor will correlate with other variables outside the matrix.'

Comrey's "where to from here" prescription leads rather naturally to consideration of *confirmatory* factor analysis which, in contradistinction to exploratory factor analysis, requires the researcher to specify a factor model before the data are analysed and is theory-based. For example, suppose that the data collected extend to the use of multiple measures of health-related quality of life which each include the dimensions of physical functioning, social functioning and psychological functioning or mental state. Thus, we would anticipate or hypothesize that there are three different underlying dimensions associated with each measure and that certain variables belong to one dimension (eg, physical functioning) while others belong to the second and third dimensions (social functioning and psychological functioning), respectively. If factor analysis is used as a means to test these expectations, then it is used to confirm certain hypotheses rather than to explore underlying dimensions. Hence, it is referred to as *confirmatory* factor analysis.

Specification of the confirmatory factor model requires making formal and *explicit* statement about: (1) the number of unobserved or latent variables (also known as, *common* factors) for which effects are shared in common with more than one of the measured or observed variables; (2) the number of observed variables; (3) the variances and covariances among the common factors; (4) the relationships among observed variables and latent factors; (5) the relationships among errors in the observed or measured variables (or *unique* factors) and observed or measured variables; and (6) the variances and covariances among the unique factors. In contrast with exploratory factor analysis, which has been revered to scornfully as a GIGO (garbage in/garbage out) model because it fails to incorporate substantively meaningful constraints and imposes substantively meaningless constraints (Long, 1983) with respect to these components, CFA allows them to be specified according to the demands of the application.

In CFA analyses of MTMM matrices, the model for each observed variable is comprised of three components: a trait component, a method component and a random error component. A diagram of the general model as it applies to the MTMM matrix (Long, 1983; Schmitt & Stults, 1986) of Table 3 (see p 42) is illustrated in Figure 10. Table 4 summarises the parameters estimated in a confirmatory analysis of the MTMM matrix as shown in Figure 10.

The CFA model assumes that the MTMM matrix can be expressed as a function of common factors as follows:

$$\Sigma = \Lambda \, \Phi \, \Lambda' + \psi \qquad (1)$$

where Σ is the MTMM matrix, Λ is a matrix of factor loadings (as shown in Table 4(A)), Φ is a matrix of correlations among trait and method factors (Table 4(B)), and ψ is a diagonal matrix of unique factor variances or random error components (Table 4(C)). To see how the MTMM approach is formulated as a confirmatory factor model, let the three traits (A, B, and C), correspond to the trait factors ξ_1, ξ_2, and ξ_3 and let the three methods (1, 2, and 3) correspond to the method factors ξ_4, ξ_5, and ξ_6. There are nine observed variables: X_1 to X_3 are measures of traits ξ_1 to ξ_3 by method ξ_4; X_4 to X_6 are measures of traits ξ_1 to ξ_3 by method ξ_5; and X_7 to X_9 are measures of traits ξ_1 to ξ_3 by method ξ_6. The observed variables, X_1, X_2, …, X_9, are represented by squares and latent (or unobserved) variables are represented by circles. A straight arrow

Table 4: Parameters estimated in a confirmatory factor analysis of the MTMM matrix for three traits and three methods [Cross-reference to Table 3 and Figure 10]

A. Trait and Method Factor loadings – Corresponding to $\mathbf{L}$ in Equation (1)

$$\begin{array}{c|cccccc}
 & \multicolumn{3}{c}{\text{Trait Loadings}} & \multicolumn{3}{c}{\text{Method Loadings}} \\
 & A & B & C & 1 & 2 & 3 \\
\hline
X_1 & \lambda_{11} & 0.0 & 0.0 & \lambda_{14} & 0.0 & 0.0 \\
X_2 & 0.0 & \lambda_{22} & 0.0 & \lambda_{24} & 0.0 & 0.0 \\
X_3 & 0.0 & 0.0 & \lambda_{33} & \lambda_{34} & 0.0 & 0.0 \\
X_4 & \lambda_{41} & 0.0 & 0.0 & 0.0 & \lambda_{45} & 0.0 \\
X_5 & 0.0 & \lambda_{52} & 0.0 & 0.0 & \lambda_{55} & 0.0 \\
X_6 & 0.0 & 0.0 & \lambda_{63} & 0.0 & \lambda_{65} & 0.0 \\
X_7 & \lambda_{76} & 0.0 & 0.0 & 0.0 & 0.0 & \lambda_{76} \\
X_8 & 0.0 & \lambda_{82} & 0.0 & 0.0 & 0.0 & \lambda_{86} \\
X_9 & 0.0 & 0.0 & \lambda_{93} & 0.0 & 0.0 & \lambda_{96} \\
\hdashline
 & \xi_1 & \xi_2 & \xi_3 & \xi_4 & \xi_5 & \xi_6
\end{array}$$

B. Intercorrelation of Trait and Method Factors – Corresponding to $\mathbf{F}$ in Equation (1)

$$\begin{array}{c|cccccc}
 & A & B & C & 1 & 2 & 3 \\
\hline
A & 1.0 & \phi_{12} & \phi_{13} & \phi_{14} & \phi_{15} & \phi_{16} \\
B & \phi_{21} & 1.0 & \phi_{13} & \phi_{24} & \phi_{25} & \phi_{26} \\
X & \phi_{31} & \phi_{32} & 1.0 & \phi_{34} & \phi_{35} & \phi_{36} \\
\hline
1 & \phi_{41} & \phi_{42} & \phi_{43} & 1.0 & \phi_{45} & \phi_{46} \\
2 & \phi_{51} & \phi_{52} & \phi_{53} & \phi_{54} & 1.0 & \phi_{56} \\
3 & \phi_{61} & \phi_{62} & \phi_{63} & \phi_{64} & \phi_{65} & 1.0
\end{array}
\qquad
\left[\begin{array}{c|c}
\text{trait/} & \text{trait} \\
\text{trait} & \text{method} \\
\hline
\text{method} & \text{method} \\
\text{trait} & \text{method}
\end{array}\right]$$

C. Random Errors Associated with Each Measured Variable – Corresponding to $\mathbf{U}$ in Equation (1)

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
δ_1	δ_2	δ_3	δ_4	δ_5	δ_6	δ_7	δ_8	δ_9

KEY: X_1 through X_9 indicate the nine measured variables; A, B and C are the three trait factors; 1, 2 and 3 are the three method factors; 1.0 and 0.0 are values fixed by the researcher that represent his/her hypotheses regarding the structure of the MTMM matrix; and the loadings ϕ_{ij}, λ_{ij} and δ_{ij} are parameters that estimated using CFA on the basis of the observed correlation matrix.

pointing from a latent variable (e.g., ξ_1) to an observed variable (e.g., X_1) indicates the causal effect of the latent variable on the observed variable. The unique factors are represented by the unlabelled arrows in Figure 10.

Figure 10 shows the loadings λ_{ij} of the observed variables on the factors. A given method factor is assumed to affect only those observed variables that are measured by that method. For example, since X_1 to X_3 are all measured by method 1, they load on the method factor associated with method 1, ξ_4, but not ξ_5 and ξ_6. Similarly, a given trait factor is assumed to affect only those observed variables that are measures of that trait. For example, X_1, X_4, and X_7 are measures of trait ξ_1 by methods 1, 2, and 3, and load on the trait factor ξ_1k, but not on the trait factors ξ_2 and ξ_3. This information is contained in the trait and method loading matrix Λ of Table 4(A). The X's and ξ_1's are added to this matrix as borders to show which observed variables and common factors are being linked by a particular loading λ_{ij}.

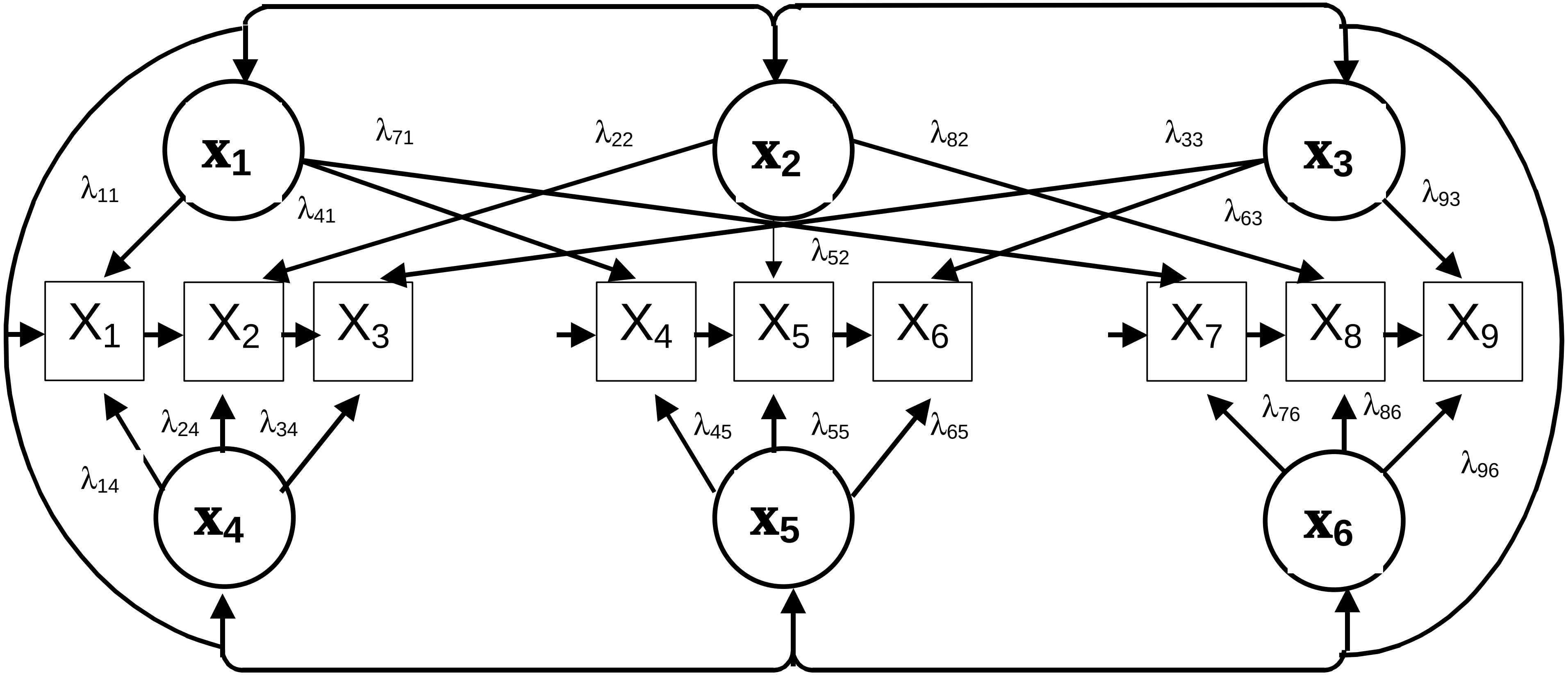
Covariances among latent variables are contained in Table 4(B), a symmetric (6 X 6) matrix that corresponds to Φ in Equation (1). These covariances are indicated by the continuous curve connecting all the latent variables in Figure 10. This matrix contains covariances among traits (trait/trait block), covariances among methods (method/method block), and covariances between trait factors and method factors (method/trait and trait/method blocks).

Table 4(C) contains the vector of random errors or unique factors associated with each measured variable. In the MTMM model the unique factors, δ_1 to δ_9 are usually assumed to be uncorrelated.

The values of 0.0 and 1.0 in the matrices of Table 4 are fixed by the researcher and represent his/her hypotheses about the structure of the MTMM matrix. All together there are 42 (= 18 + 15 + 9) parameters that are free or estimated on the basis of the observed correlation matrix. The confirmatory approach allows for estimation of these parameters, as well as tests of their significance and the decomposition of each bivariate correlation in the MTMM matrix into a trait component and method component.

While the matrix of Table 3 is the smallest MTMM matrix for which a full model can be tested, it is possible to test smaller matrices if any or all of a number of restrictive assumptions can reasonably be made (e.g. lack of correlation among trait and method factors, which involves fixing the set of parameters that comprise the method/trait block in Table 4, $\{\phi_{41}, \phi_{41}, \phi_{42}, \phi_{43}, \phi_{51}, \phi_{52}, \phi_{53}\}$ to 0.0; or a lack of correlation among method factors and among trait and method factors, and an equivalent influence of method factors across traits, which involves fixing parameters comprising the method/trait and method/method blocks in Table 4, $\{\phi_{41}, \phi_{41}, \phi_{42}, \phi_{43}, \phi_{44}, \phi_{51}, \phi_{52}, \phi_{53}, \phi_{54}, \phi_{64}, \phi_{65}\}$ specifying a single loading for each method factor). Thus, the question of discriminant validity can be addressed through the use of a multitrait-monomethod design if, for example, three traits (e.g., adjustment to illness, health-related quality of life, and social support) are measured by three instruments, all of which use the same self-report method. And, in the simpler still monotrait-monomethod design where one trait (say, health-related quality of life) is tapped by, say, three somewhat different self-report measures (e.g., Sickness Impact Profile, Nottingham health Profile, and Quality of Well-Being Index), CFA can estimate the loading of each measure on a common factor. An additional advantage of CFA in this circumstance is that it allows the researcher to test whether the correlation among measures is attributable to shared trait variance and/or shared error variance.

Figure 10: General model of the determinants of nine measured variables each comprising a trait, method and random error component



CFA is a form of structural equation modelling (SEM). The degree of fit between a confirmatory factor analysis model and data can be evaluated with computer programs such as LISREL (Joreskog & Sorbom, 1984) or EQS (Bentler, 1985). The plausibility of the hypothesized CFA model is assessed by comparing the implied data structure with the observed data structure. The model's goodness of fit is usually evaluated using the chi-squared statistic and a number of other measures of practical fit: rho, delta, and the comparative fit index (Bentler, 1990; Bentler & Bonnet, 1980).

Of course, CFA is not a panacea (Marsh, 1989). Not every test-validation data set can satisfy its practical demands (see Cole, 1987) and, as Bergner (1990) reminds us, the parameter estimates in CFA are affected by the underlying model. Results of analyses in respect of a particular model indicate only whether the model fits the data, not whether there are other models that may fit the data equally well or even better. Likewise, a better fit does not necessarily mean that a model is more accurate. What is satisfactory on statistical grounds, may or may not be acceptable with respect to theory.

There are different opinions regarding the usefulness of factor analysis for constructing health indexes. Kaplan, Bush and Berry (1976) reject it on two grounds. First, they consider that its use may result in the dropping of items that are checked rarely or are poorly correlated with other items despite their social significance. Eliminating items that have low "weights" on all factors is, they argue, tantamount to substituting variation in frequency for variation in social importance. This is, of course, a valid criticism. Value weighting schemes derived from relative frequency are fundamentally different from those based on relative importance. As Bush (1984, pp 118-119) writes, 'whether from a preference or a prognostic point of view, one hundred runny noses are not the same as one hundred severe abdominal pains'. However, the problem then becomes how many such items can be retained under the umbrella of a unitary concept and how is social importance to be determined. Clearly, there are judgment calls involved whether or not one chooses to use factor analysis.

Kaplan et al's (1976) second objection relates to health status measurement in evaluation contexts, and arises because the factor structure is a product of the correlations among controllable *and* uncontrollable variables (eg, age, income, etc). A health program that has a significant effect on the controllable set of variables will, they argue, alter the factor structure and invalidate the factor equation carried over from prior analyses, as an outcome index, thereby biasing any estimate of change. Differential rates of change among variables will similarly alter the factor structure and preclude its use as a description of health status change. The main issue here, how should we validate **within-subject changes** in index scores, is taken up below.

Hypothesis Testing and the Significance of Significance

Construct validation often relies heavily on correlational evidence. Regardless of the way in which construct validity is conceptualised, the relationship between measurements is at issue. We have already indicated, in the context of our discussion of criterion-related validity, some reasons why correlation coefficients are problematic. Another problem is that, when recourse is made to correlation coefficients to test hypotheses about the extent of the relationship between two variables (i.e., items included in one or more measures or components measured by different methods), it is often not obvious what we should make of the correlation coefficients obtained. How should they be interpreted?

For example, what correlation should be expected between a new measurement and an existing measurement? What is the null hypothesis? Surely, it is not appropriate to test for a zero correlation between, say, a patient self-ratings provided on, say, a 7-point scale of "overall functioning" and an aggregate score obtained using a multidimensional general health status measure? The hypothesis that there is no correlation between the two sets of measurements reflects either a profound lack of confidence in the new measurement or in the external criterion against which it is to be validated.

This seems to be a common weakness in reports of construct validation. Very few studies articulate hypotheses about relationships among concepts or make predictions concerning the magnitude of relations they expect to obtain if the instrument in question is really measuring health status. To date, we are aware of only one study of health status measures [Guyatt, Deyo, Charlson *et al* (1989)] in which *a priori* hypotheses about the size of correlations are stated. Even then, the authors leave us in some doubt as to what level of correlation they regard as indicative of validity, since the correlations obtained are lower than hypothesized but are nonetheless interpreted as confirmatory evidence.

The problem of the nebulous null hypothesis has attracted the attention of other reviewers, too. McDowell and Newell (1987) note that 'all too often an author reports whatever correlation he obtains and then concludes that the test is thereby shown to be valid?' (p 29). Spitzer (1987b) has voiced a related but more general concern: 'very seldom do we see statements submitted in advance in a protocol that specifies at what point a measure will be declared valid'. Meehl (1986) puts the point most colourfully and offers a ballpark estimate of the level of correlation he would take seriously:

> 'What my colleague Lykken [1968] calls the "ambient noise", or "crud factor", is of unknown average value, but it can hardly be supposed to be less than, say, in correlation terms, Pearson r = .25 in the soft areas of psychology. "Everything is correlated with everything", and .25 is probably not a bad average value. *Randomly chosen individual differences variates do not tend to correlate zero. Of course in real life, the experimenter is usually correlating variates that belong, at least commonsensically, to some restricted domain.* We don't usually do studies correlating social dominance with spool-packing ability or eye color. *So a more realistic guesstimate of the crud factor, the expected correlation between a randomly chosen pair of variates belonging to a substantive domain, would be higher than that, maybe as high as .30* (p 32, emphasis added).

Certainly, not making reasoned declarations of what level(s) of correlation should be construed as adequate from the perspective of demonstrating construct validity becomes less tenable as more and more research on health status measures is done. Often, it seems to us, the correlations obtained are interpreted favourably, regardless of their size *vis a vis* those in earlier studies. As anticipated by Peak (1953, p 288) and cited by Messick (1989, p 49): 'a protest must be entered … against the proliferation of blindly empirical validities which are without the disciplined guidance of theory, for the increment of meaning from the accumulation of miscellaneous correlations may ultimately approach zero.'

This problem of the ready acceptance of whatever correlations are obtained as evidence of construct validity is compounded by a high degree of quite mindless significance testing. Carver (1978) describes the root of the problem as follows:

> ‘Statistical significance is generally interpreted as having some relationship to replication …, and replication is the cornerstone of science. If the results are due to chance, then results will not replicate. The only valid reason for considering statistical significance is to try to determine whether research results are simply a product of chance and will therefore not be replicable. Yet it is not logical to deduce that if the results are statistically significant, they will replicate, or that if the results are not statistically significant, they will not replicate. But if researchers do obtain the same result more than once, it is more reasonable to conclude that the results are not due to chance.
>
> ‘Since one of the primary reasons for being concerned with statistical significance is a threat to replication, replicated results automatically make statistical significance unnecessary (Bauernfeind, 1968). Stevens (1971) stated the relationship between statistical significance and replication this way:
>
>> In the long run scientists tend to believe only those results that they can reproduce. There appears to be no better option than to await the outcome of replications. It is probably fair to say that statistical tests of significance, as they are so often miscalled, have never convinced a scientist of anything. (p 440)’

As we were at pains to emphasize earlier, validation is very much about hypothesis testing, about whether inferences drawn on the basis of test scores are vindicated. Hypothesis testing is very much about whether, and in what contexts (McGuire, 1983; Meehl, 1978), results replicate.

Avoidance of the need to make a scientific judgment about the magnitude of a correlation coefficient by focusing on its statistical significance is commonplace. But it is a tactic that has little to recommend it in that the P value depends strongly on the size of the sample, n, on which the calculation is based (Edwards, 1976; Lykken, 1968; Feinstein & Kramer, 1980). To see this, consider the formula for a t test used to interpret the value of Pearson’s correlation coefficient, r:

$$t = \frac{r \sqrt{(n-2)}}{\sqrt{(1-r^2)}} \qquad (2)$$

As a rough guide, a t value of 2 or higher will usually be statistically significant at P < .05. Setting t = 2 in Equation (1) and solving for the sample sizes that will transform given values of r into statistical significance we obtain the following:

r	.05	.1	.2	.3	.4	.5	.6	.7
n	1,598	398	98	43	23	14	10	7

Clearly, calculating the P value associated with a correlation coefficient is not a useful way to determine whether it is substantively significant. Indeed, it is worth adding that the well-entrenched use of indexes of trend, like Pearson’s correlation coefficient, may itself be inappropriate in such circumstances. Very often we are not so much interested in trend or relatedness (ie, the strength of the tendency for changes in one variable to be reflected by changes in another) but rather the extent to which two variables can serve as a surrogate for one another (Deyo, Diehr & Patrick, 1991; Kramer & Feinstein, 1981). This applies particularly to assessing reliability but may also be applicable to the assessment of validity where the accent is on the degree of agreement between alternative measures (especially if there is an accepted gold standard).[7]

Our lament is that, at present, there is too much inductive and too little deductive reasoning involved in the selection of measures and the analysis of results (Patrick and Bergner, 1990).

4. LONGITUDINAL VALIDITY – AND RESPONSIVENESS

Most discussions of the concepts of reliability and validity are set within a measurement theory framework and are concerned, in the main, with discriminative and predictive indexes. They are concerned essentially with cross-sectional comparisons and predictions, respectively, and, therefore, responsiveness, or sensitivity to change, is not relevant to their purposes (Carver, 1974).

For some purposes, however, most obviously in the case of clinical research where evaluative indexes play a major role, this is a serious error of omission. Where the focus of interest is the measurement of changes in patients' health status, construct validity is concerned not with the representation of health or well-being at some given point in time (do between-subject differences in instrument scores bear the expected relation to differences in other variables measured?) but with the ability of that representation to change in response to real change over time (do within-subject changes in instrument scores bear the expected relation to changes in other variables measured?). The efficacy and effectiveness questions "can it work?" and "does it work?" beg the question of what "working" entails. Presumably, working can be equated with *changing* the patient's health condition for the better and instruments that are more responsive to changes in health status are presumably more sensitive measures of the effects of health care interventions (Normal, 1989). To be useful in the evaluation of clinical interventions, instruments must be able to detect *clinically* significant changes, *even if such changes are small* (Deyo & Centor, 1986; Deyo & Inui, 1984; Guyatt, Deyo, Charlson et al, 1989; Guyatt, Walter & Norman, 1987; Kirschner & Guyatt, 1985), and this is the question addressed in assessing longitudinal validity.

This said, it should be added that teasing apart the concepts of responsiveness and validity is not quite as straightforward as it first appears. It is perhaps tempting to regard responsiveness as an analogue of the concept of validity that applies only in the case of evaluative indexes. This supposition is incorrect.

We know as if by rote that an instrument is valid if it actually measures what it is intended to measure. What is implicit in this definition is that the accoutrements of validity differ according to the purpose for which an instrument is used. This is already well established in the case of

[7] In the case of continuous variables the interclass correlation coefficient is a worthy replacement for Pearson's r. It combines a measure of correlation with a test of the difference between means and adjusts for systematic bias. Like Pearson's r, it varies between –1 and +1, with higher scores reflecting increasing method or observer agreement.

discriminative versus predictive indexes but it applies to discriminative versus evaluative indexes, too. Thus, whereas *cross-sectional* construct validity is central in the case of discriminative indexes, with evaluative indexes we seek *longitudinal* construct validity. The first interpretation focuses on the question, “Do cross-sectional or *between-subject* differences in index measurements taken at a single point in time bear the expected relationship to external measures?” The second interpretation is concerned with the question, “Do longitudinal or *within-subject* changes in index scores associated with an intervention bear the expected relation to changes in other variables measured?” It is not obvious that an instrument that is able to measure satisfactorily between-subject differences in health status should necessarily be able to measure satisfactorily changes in health status within individuals (Churchill, Wallace, Ludwin, et al, 1991; MacKenzie, Charlson, DiGioia, Kelley, 1986a, 1986b; Wiklund & Karlberg, 1991).[8]

Guyatt et al (1989a) provide a number of examples that illustrate this point. In one example, they indicate that an instrument designed specifically to measure aspects of physical, emotional and social functions related to adjuvant chemotherapy in women with breast cancer, the Breast Cancer Chemotherapy Questionnaire (BCQ), was able to detect statistically significant differences in the 24 week minus 12 week change scores of patients in the short (12 weeks) versus long (36 weeks) arms of a randomised controlled trial, whereas the Rand physical function and emotional function instruments, which have previously been shown to be valid discriminative measures (Ware, Brook, Davies-Avery, et al, 1980), did not demonstrate responsiveness. Another example concerns the use of the Sickness Impact Profile (SIP) in the context of a controlled trial of different strategies for managing patients with back pain. Clearly, although items like “I’m eating no food at all, nutrition is taken through tubes and intravenous fluid” or “I do not feed myself at all, and must be fed” may be reproducible and valid in discriminating among individuals in terms of their health status, they will not be responsive in the setting concerned. By contrast, other items such as “I keep rubbing or holding areas of my body that hurt or are uncomfortable” or “I am not doing any of the house cleaning that I would usually do” may prove useful in detecting changes over time in patients with back pain. Needless to say, this example invites consideration of the trade-offs between generic versus disease-specific measures, as discussed below.

Other examples presented by Guyatt et al (1989a) are similarly helpful in amplifying the relationships among the concepts of reproducibility, validity and responsiveness. Collectively, their examples make clear that what counts is whether *changes* can be validated.

An example based on a toxicity questionnaire, the Eastern Co-operative Oncology Group Criteria (ECOG), illustrates that an index may be responsive but not longitudinally valid. ECOG includes items such as white blood cell and platelet counts, laboratory tests of liver function, evidence of allergic reaction and physician assessment of patients’ problems with nausea, vomiting, stomatitis and diarrhoea. In the context of the same randomised controlled trial of adjuvant chemotherapy for breast cancer referred to above, this instrument proved to be responsive despite the fact that it is not valid as a measure of overall health status. It is true that ECOG includes some subjective health components, and it may correlate with the change in subjective health status a patient experiences during the course of chemotherapy. However, it was designed to measure drug toxicity rather than subjective health status. Not surprisingly, it

[8] It should be noted that instruments that measure quality adjusted life years for cost-utility analyses must satisfy both requirements. If a health status instrument is to address satisfactorily questions about the efficacy and/or effectiveness of an intervention, its longitudinal validity should be established. If the same instrument is to be used to examine questions of efficiency – to make comparisons across programs – it also needs to have an acceptable level of cross-sectional validity.

fails the test of clinical sensibility (Feinstein, Josephy & Wells, 1986; Feinstein, 1987a, 1987b): inter alia, it lacks face validity and content validity.

There is always, too, the possibility of a placebo-type effect where patients report an improvement in their subjective health status following an intervention. The change reflected in subjects' responses may indicate satisfaction with the program or a courtesy bias, rather than a change in their subjective health status, in which case the instrument used would be responsive but not valid.

To illustrate the coincidence of longitudinal validity and responsiveness, Guyatt et al (1989) discuss the assessment of disease-specific aspects of physical and emotional function in patients with Crohn's disease or ulcerative colitis, using the Inflammatory Bowel Disease Questionnaire (IBDQ) (Guyatt, Mitchell, Irvine et al, 1989b). Two interviews were conducted, one month apart. The IBDQ's longitudinal validity was gauged in terms of the relationship between change scores obtained from the two interviews and patients' self-ratings of whether, and to what extent, their disease activity had changed for better or for worse. The IBDQ showed only small intrasubject variability over time in patients who reported their health status as stable. Global ratings of change showed moderate to high correlations with changes in IBDQ score. Moreover, the largest changes in IBDQ scores were registered by patients who reported overall improvement or deterioration, suggesting that the instrument is responsive. Coupled with the fact that these changes were greater than the differences in scores for patients whose global rating of their disease activity suggested they were stable, which result provides evidence for the instrument's longitudinal validity.

This last example shows clearly that, in assessing the validity of an evaluative index, we must more or less adopt a gold standard. It is absolutely necessary that we have a way of determining whether or not true change has or has not occurred as a direct result of an intervention. Here, there are two preferred alternatives: use an intervention that we know works or use a health status measure that is regarded as longitudinally valid. A third option is to use a transition index (eg, MacKenzie et al, 1986a, 1986b). Indexes constructed to reflect a single state ("How tired have you been? Very tired, somewhat tired or full of energy?") which are used to collect before- and after-measures of health status are more efficient than instruments that focus on transitions (eg, "How has your tiredness been? Better, the same or worse?") because changes are calculated rather than explicitly assessed. On the other hand, transition indexes avoid the potential problem of floor and/or ceiling effects which may occur with indexes that reflect a single state. In the present context, the ability to document floor effects would seem to be a particularly important performance characteristic; changes for the worse in severely ill patients with low baseline scores should not go unnoticed (see Bindman, Keane & Lurie, 1990, for an example of this instrument bias).

A corollary of the need for a gold standard is the fact that simultaneously trying to validate an instrument *and* assessing the efficacy of a treatment or program involves a problem of circularity. Like many glimpses of the obvious it is an often overlooked trap. Examining scale score changes following an intervention of known efficacy is one means of validating an evaluative instrument. Likewise determining the efficacy of an intervention requires, *inter alia,* an evaluative instrument of known validity. Depending on the nature of the intervention and the nature of the disease condition, accomplishing either task may require research findings that involve measurements made at more than two points in time.

Responsiveness Considered Further

In general term, *responsiveness* is to an evaluative index what *discrimination* is to a discriminative index. As index “dimensions”, they are each concerned with whether “meaningful” differences are, in fact, measured. They lie at the root of the debate about the relative merits of generic versus disease-specific measures.

Discrimination refers to the spread of scores or the number of categories into which individuals can be placed and the degree to which it is present is central to the validation of instruments that aim to define cross-sectional differences among individuals (Bergner & Rothman, 1987; Kirshner & Guyatt, 1985). Discrimination is very much a product of item selection, item reduction and item scaling processes. A crude scale will allow individuals to be placed in only a small number of categories whereas a scale with finer discrimination will allow a greater range of scores. In general, the fineness of the discrimination achievable will depend on the number of items per health dimension and/or the number of categories within a dimension. Crude scales will be adequate to detect differences among groups of individuals when the expected difference is large and this points to an interaction between the desired level of scale discrimination and the intended sample. Noncongruence between the level of health status assessed and the target population will result in a skewed distribution of health status scores ie, a large proportion of individuals will receive the same or similar scores such that the detection of meaningful differences is problematic (Bergner & Rothman, 1987). The reliability coefficient (ie, the ratio of variance between subjects to total variance) is a generally accepted overall measure of the ability of an instrument to discriminate among individuals.

Responsiveness is determined by two properties: reproducibility and changeability (Guyatt, Veldhuyzen van Zarten, Feeney & Patrick, 1989b; Guyatt, Walter & Norman, 1987). The objective is to detect changes above and beyond the variability observed in subjects who do not receive an experimental intervention. A measure that is reproducible yields the same results when repeated in *stable* subjects (ie, subjects whose status has not changed) at two points in time. What is critical is that the magnitude of within-individual variance be small. It represents the “noise” that makes the minimal clinically important difference or “signal” difficult to detect. Changeability is concerned with whether an instrument registers score changes when clinically important improvements or deterioration in quality of life occurs. The responsiveness of an instrument is proportional to the change score that represents the clinically important difference and inversely proportional to the variability in score in stable subjects.

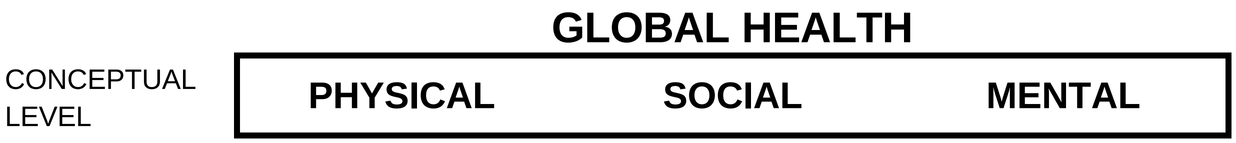
As with discrimination, responsiveness is very much a product of item selection, item reduction and item scaling processes – except that the accent is on measuring within-person change over time rather than on measuring between-person differences. Changeability will depend on the number of response options per item, the number of items per health dimension and the number of categories within a dimension. The criterion governing the selection and reduction of items is the likelihood that an individual’s health status will change as a result of the application of an intervention. Clearly, items that are unresponsive – either unreliable in a test-retest sense with stable subjects or not sensitive to change in individuals whose health status is not stable – should be deleted. Beyond this, deciding how many items to retain or delete is less straightforward than with discriminative indexes. The usual procedures for measuring internal consistency, KR20 and Cronbach’s alpha (Carmines & Zeller, 1979; Cronbach, 1951; Kuder & Richardson, 1937), assume that the precision of the index will increase incrementally with the covariance of the items and the number of items included. It is not clear that these assumptions are appropriate for evaluative indexes. In the first place, items in an evaluative index need not be correlated at a single point in time. Rather, they should be consistent in the way that they measure change in health status *over* two points in time. Secondly, regardless of whether they are correlated, the greater the number of items included in an index, the greater the probability of including items that may prove insensitive to efficacious treatment and, as noted above, any variability in item scores that is not related to the intervention may obscure any treatment effects.

The issue of responsiveness is very much at the root of the debate about whether to use so-called disease-specific instruments rather than generic instruments in clinical research. Generic health status measures are those that are designed to be broadly applicable across types and severities of disease, across different medical interventions or treatments, and across sociodemographic and cultural subgroups. Disease-specific measures belong to a broader class of specific measures and are designed to assess specific diseases, diagnostic conditions or patient populations.[9] Up to a point the relationship between the two approaches is a matter of the level of operationalization, as Spitzer (1987b) has indicated. Figure 11 shows three levels of operationalization of health-related quality of life and functional status, subsumed by a general health status concept, which give rise to a hierarchy of possible data gathering instruments. It is significant that specific symptoms associated with specific diseases are regarded as targets for *hypothesis-determined* measures.

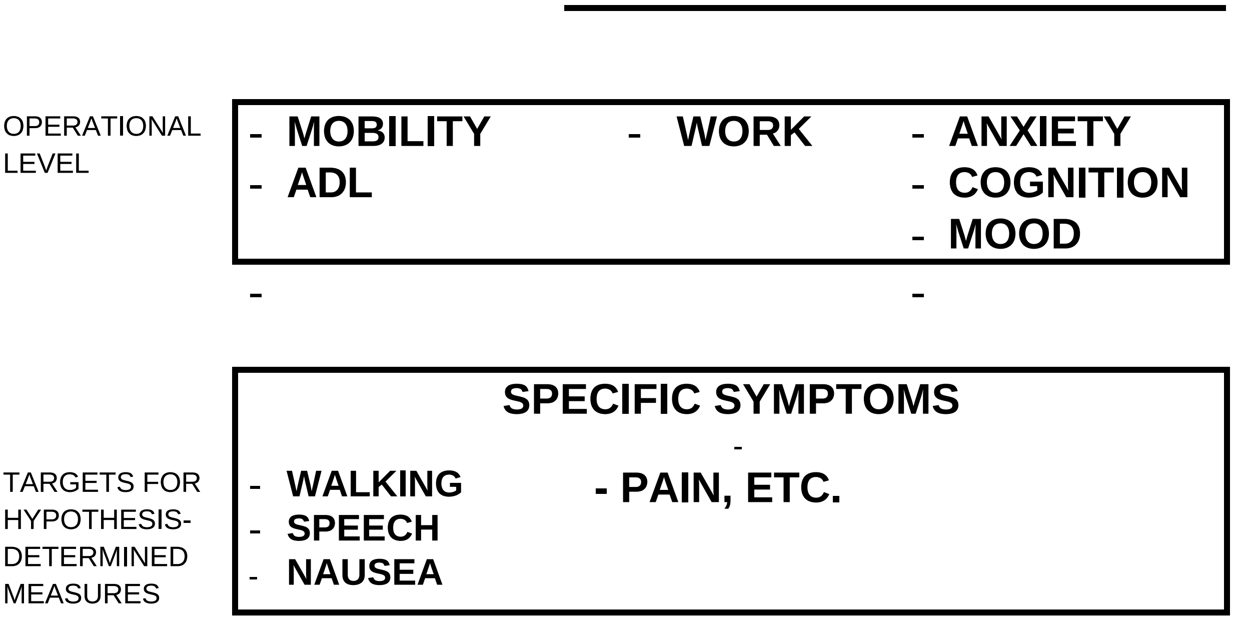
The rationale for disease-specific measures is the increased responsiveness to disease-specific interventions that *may* result from the inclusion of those aspects of quality of life that are of particular concern or relevance to patients being studied. In its simplest form, the idea is that the more focused the instrument, the more responsive it is likely to be. However, this assumption cannot be made on the basis of ipso facto reasoning and the issue has to be resolved empirically. Generic measures may be just as useful and responsive in some settings as disease-specific measures but, to date, there have been very few head-to-head comparisons. This is not altogether surprising when we consider that we do not yet know exactly how to address the issue of relative responsiveness. There is uncertainty about how to calibrate responsiveness and the idea of calculating a "coefficient of responsiveness" is beset by the problem of specifying the smallest clinically important change for alternative health status instruments, which may vary with the application involved. This raises the question of the generalizability or external validity of the findings from head-to-head comparisons. In principle, the relative responsiveness of specific versus generic instruments may have to be resolved seriatum.

Apart from practical barriers (including respondent burden and resource requirements) to the use of generic measures, which are no doubt real, there are other reasons why specific measures seem to be preferred by physicians. These relate to the greater clinical salience of

Figure 11: Levels of operationalization for quality of life and functional status



[9] Although the term 'disease-specific health status measure' is used in contradistinction to the term 'generic health status measure', sometimes it is more appropriate to juxtapose generic versus specific measures per se. This is because not all specific measures are disease related (Deyo & Patrick, 1989; Guyatt & Jaeschke, 1990). An instrument may be specific to a disease (eg, inflammatory bowel disease, rheumatoid arthritis, cancer, diabetes, heart disease, etc); specific to a given condition or problem (eg, back pain, dyspnea, pain); specific to a certain function (eg, sexual or emotional function); or specific to a given population of patients (eg, the frail elderly, developmentally disabled children, etc.).



OPERATIONAL LEVEL

- MOBILITY
- ADL

- WORK

- ANXIETY
- COGNITION
- MOOD

-

-

TARGETS FOR HYPOTHESIS-DETERMINED MEASURES

SPECIFIC SYMPTOMS

-

- WALKING
- SPEECH
- NAUSEA

- PAIN, ETC.

Source: Spitzer WO. State of science 1986: quality of life and functional status as target variables for research. *Journal of Chronic Disease* 1987; 40: 465-471. Reproduced with permission.

specific measures, and more particularly, to the content validity of measures and the meaningfulness or interpretability of change scores.

Specific measures have the advantage that they have high *face validity* and *content validity* for physicians; they appear to be clinically sensible as they relate closely to signs and symptoms and areas of functioning that are routinely explored by physicians. In contrast, generic measures sometimes have low content validity for patients and physicians alike. They may contain items that are of little or no relevance to the study population and which add to respondent burden without contributing to responsiveness (eg, questions about incontinence or eating behaviour may not be of concern to patients with chronic obstructive lung disease). On the other hand, it is not always possible to specify in advance the impact that a treatment or intervention will have on the patient's health status: there may be unanticipated positive or negative side-effects (eg, Rockey & Griep, 1980) and physicians may make incorrect assumptions about the clinical symptoms that have an adverse impact on well-being.[10]

At this point it is worth reiterating that validity is first and foremost concerned with inferences that can be made on the basis of *test scores.* As we noted in an earlier section of this monograph, content validity focuses on instruments rather than measurements, with the inputs to the measurement process rather than the outputs of the measurement process. To the extent that content validity is not concerned with score-based inferences, it should not be allowed to loom large in the calculus when choosing between selecting specific versus generic health status measures.

A related but more telling concern about generic measures relates to the interpretability or meaningfulness of changes measured. This is very much a question of validity: making score-based inferences presupposes that the scores are interpretable.

There is no widely accepted notion of what represents a clinically meaningful change in health status. Once upon a time, the same could have been said in respect of changes in many traditional clinical measures, like blood pressure and forced expiratory volume (FEV_1), that physicians now regard as self-explanatory. The difference is that such measures are now familiar and can be interpreted in terms of well-established or agreed cutoff points that have become linked to concrete performance-oriented outcomes through accumulated experience. Paterson's (1988) observations in respect of the findings of the trial of auranofin therapy for the treatment of rheumatoid arthritis (Bombardier, Ware, Russell et al, 1986) emphasise the contribution of such linkages. Thus, he comments:

> 'to be meaningful a quality-of-life instrument should further a judgment as to the practical importance of the score observed. The traditional measures do not do this, since their units of measure, such as millimetres of mercury or seconds of walk time, have little meaning in the context of daily life. For example, it is probably only those rheumatologists experienced in the use of traditional measures and the literature about them who would know how the grip strength of healthy 16 year-old boys, say, compares with that of 60 year-old women, let alone how grip strength may be expected to change with different

[10] Hunt (1988) provides a nice example, drawn from Priestman (1986), of the discrepancy that can exist between the perspectives of patients and physicians regarding the net effects of treatment. In a comparison of cytotoxic versus endocrine therapy for breast cancer, the prevalence of nausea, vomiting, constipation and total alopecia was found to be greater in women on cytotoxic drugs. Notwithstanding these side effects, their feelings of well-being increased over an eleven week period. Apparently, the symptomatic relief ensuing from tumour shrinkage more than offset the distress caused by these side effects.

> therapies. The global measures' units have no bridge whatever to concrete experience, so that the therapeutic importance of a score change is unknowable by itself. Only through repeated correlation of the results with other concrete results can the global measures take on meaning. Of the simple, scalar type measures perhaps the 10-centimetre Pain line comes closest to having built up a meaningful framework of such correlations. This argues for quality-of-life instruments with component items based on performance, since these have meaning in terms of common experience. Indeed, the score changes on the HAQ [Health Assessment Questionnaire] and QWB [Quality of Well-Being Index] can be expressed in terms of change in performance of a single daily act, and the importance of the ability to perform that act or not can be reasonably judged. While the PUMS [Patient Utility Measurement Set] does not have performance items, its score appears amenable to the same kind of translation into experientially meaningful units – for example chances of death or years of life.
>
> 'Unless concrete equivalents of the QWB or PUMS scores are eked out, the practical importance of a score change requires the same kind of framework of prior correlational experience that any other unfamiliar measure requires. That the two measures each employ a 100-point continuum and are each anchored by death and full health helps orient the lay evaluator but does not communicate the practical importance of, say, a +16 point change – even if expressed as percentage of full health (pp 186-187).

Another interpretability question is whether "a change is a change is a change". As Deyo and Patrick (1989, p S259) observe, without the necessary bridges to concrete experience, 'many will wonder if a given absolute change is equally important at different points of the health status spectrum.' This has an analogue with discriminative indexes of health status in the question whether "a difference is a difference is a difference". For instance, what does it mean to have a total score in the range 0 to 4 versus one in the range 5 to 9 on the Sickness Impact Profile? Changes in disease-specific scores may be easier to interpret because they are more familiar, more specific, more precise, or more closely linked to changes in clinical measures of disease activity (Deyo & Patrick, 1989; Guyatt & Jaeschke, 1990; Patrick & Deyo, 1989). As a result, there may be a greater degree of consensus among physicians about how to interpret and communicate the prognostic implications or practical importance of observed change scores. Determining what it is we can or should conclude about the effects of treatment on the basis of changes in scores on only the most general dimensions of health, such as general health perceptions and social/role functioning, is far from straightforward.

Another wrinkle to the problem of interpreting changes has to do with the level of aggregation or disaggregation of scores. As Feinstein et al (1986) note, an improvement on a single item or dimension that is self-evident and impressive (eg, a 75% change, from a score of 1 to 4) may be obscured in the context of an overall score (eg, a 4% change, from a score of 77 to 80).

5. GENERALIZABILITY

Construct validity asks the question: "Can we generalize from this operational definition (measured variable) or set of operational definitions to an underlying construct?" In juxtaposing the four kinds of validity with which their names have become synonymous, Cook and Campbell

(1979, p 81) point out that 'making generalizations is the essence of both construct and external validity. … Just as one gains more information by knowing that a causal relationship is probably not limited to particular operational representations of a cause and effect, so one gains by knowing that the relationship (1) is not limited to a particular idiosyncratic sample of persons or settings of a given type, and (2) is not limited to a particular population of Xs but also holds with populations of Ys and Zs.' External validity asks the question, "To what subject populations, settings, treatment variables, and measurement variables can findings concerning the effect of the independent variable X (ie, treatment or intervention) be generalized?"[11]

Questions of external validity arise most straightforwardly in contexts where evaluative indexes are used. Evaluative indexes are concerned with measuring change or responsiveness to an intervention and they reflect part of the cause-and-effect relationship to which Cook and Campbell allude. In the context of a randomised controlled trials, for example, the primary goal is to infer treatment effects from group differences by comparing the average treatment effect on health status for the experimental and control groups respectively. Here one is concerned with generalizability *to* the target population that fulfils the study's inclusion criteria. A more differentiated data analysis may seek to discover whether there are individual differences in the amount of change experienced and to identify the factors associated with a good response. Differentiated findings would indicate that treatment effects could not be generalized *across* all subpopulations (Cook & Campbell, 1979; Streiner & Norman, 1989). Certainly, if the clinical context is one where there is reason to suspect a possible trade-off between survival and quality of life, it is probably appropriate to either define an age range or to stratify by age.

The extent to which the issue of generalizability is of concern in the application of health status assessment in clinical trials we cannot say. We do note, however, a degree of disquiet in some quarters about the lack of attention to confounding factors. For instance, Hollenberg, Testa and Williams (1991) have raised a number of external validity-related concerns regarding the use of quality of life as a therapeutic end-point in relation to the therapeutic trials in hypertension. They summarise:

> 'Although some insights have been gained on the relative influence of various therapeutic regimens on the quality of life of treated patients, in many of the studies too little consideration has been given to the use of instruments that have been validated in the patient population to be studied, to the power of the study and its design, to the contribution of confounding variables such as age and gender, and to evidence that short term trials (measured in weeks) can miss important changes that occur over months in a process where treatment is life-long.'

Questions of generalizability, if not external validity per se, may be raised legitimately in respect of discriminative indexes, too. If the above quotation from Cook and Campbell is purged of all references to cause-and-effect relationships, the logic of the resulting statement – 'just as one gains more information by knowing that a relationship is probably not limited to particular operational representations, so one gains by knowing that the relationship (1) is not limited to a

[11] Note that our use of the term generalizability here differs somewhat from that of Deyo and Patrick (1989, pp S257-S258) who discuss the 'uncertain generalizability of *instrument performance*' as a methodologic barrier to the use of health status measures in research, clinical and policy applications. Whereas our emphasis in this section is with the generalizability of health status *scores* for a given sample to *and* across populations, they are concerned also with the "transferability" of the instrument per se (eg, does the instrument's response format need to be modified to optimise performance when dealing with, say, elderly day care subjects?) Practical or "instrumental" aspects of performance, linked to respondent burden, thus seem to be implied. Of course, to the extent that an instrument is demonstrated to be relevant and valid or "equivalent" for different generational clinical or cultural groups, their scores may be compared.

particular idiosyncratic sample of persons or settings of a given type, and (2) is not limited to a particular population of Xs but also holds with populations of Ys and Zs' – is no less appealing than its causal counterpart.[12]

The avowed aim of generic health status measures is comparability. As Patrick and Deyo (1989, p S227) put it, 'comparability across different diseases, conditions, populations, or investigations requires a generic approach to health status assessment. … generic measures provide a common denominator or common unit of outcome by which to judge the relative severity of health outcomes and the relative effectiveness of interventions.' Already some health status measures, most notably the Sickness Impact Profile (SIP), have been applied to a large array of patient groups. The resultant profile of total scores by disease condition or population group is reproduced in Figure 12 as an example of the fruits of this exercise.

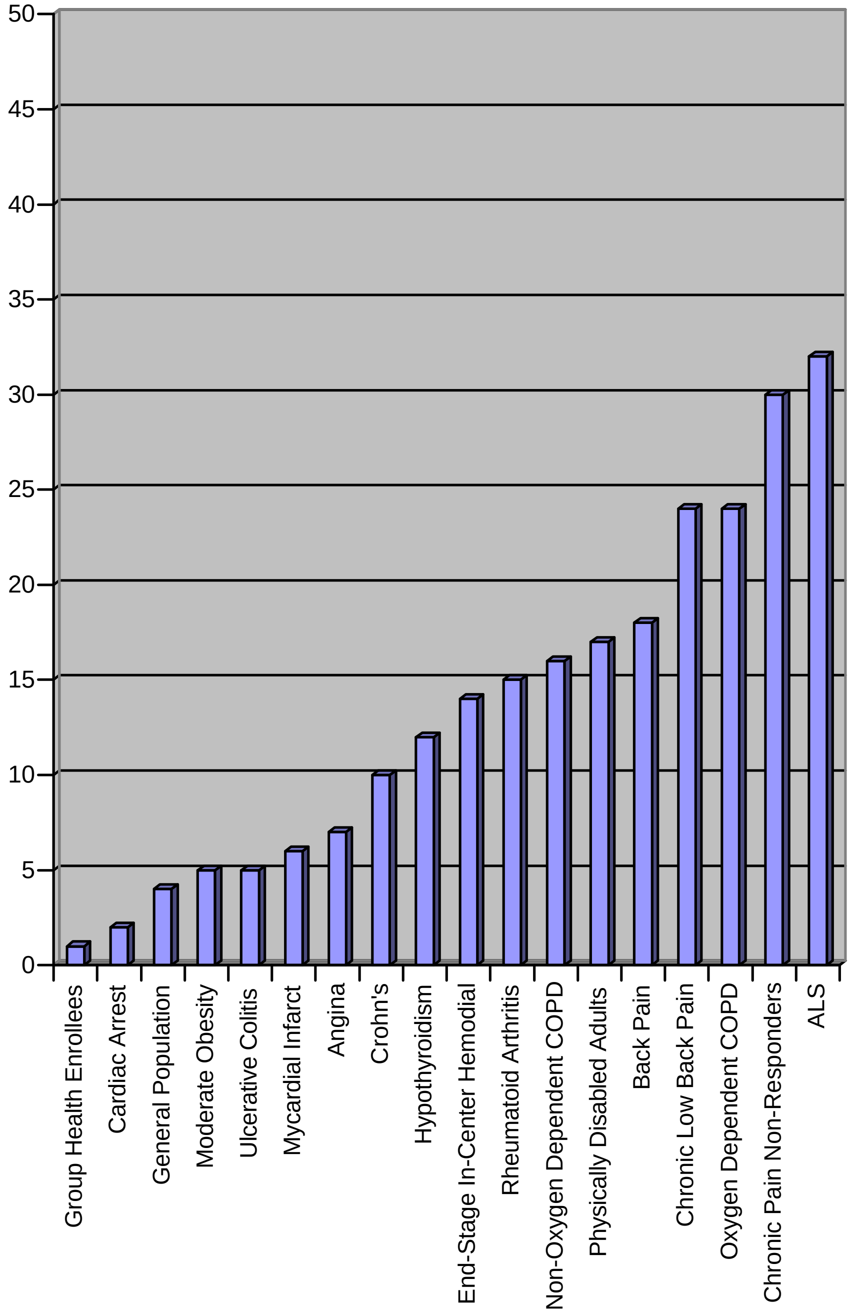
The ability to generalize to and across target populations would seem quite central to this endeavour. It seems to us that we are entitled to wonder whether the placement of specific disease conditions and population groups along the health status continua associated with the scales defined by particular instruments are generalizable. Indeed, not only are we entitled to wonder, as a research community we *should* wonder: afterall, many cross-sectional studies of health status are based on patient samples that are quite small and of unknown representativeness *vis a vis* the relevant patient population. Moreover, in the cases where cost-utility analyses are incorporated in league tables to provide a basis for resource allocation decisions, much may hinge on whether the quality and quantity of life years saved that go into the denominator of the cost-utility ratio are representative of the patient cohort with a particular disease condition and the issue of generalizability across subpopulations has not been much addressed. Indeed, the very manner of reporting of many studies involving health status assessment seems to militate against the very kinds of comparability that, over time, are likely to prove most instructive. The most frequently reported summary statistics with respect to the distribution of health status index scores are the mean (invariably) and standard deviation (usually). In the process, the scores are seldom standardized for age, sex, or other characteristics of potential interest. Here we would argue that it hardly seems tenable to use the predicted relationship between age and health status to gauge the validity of a measure of health status[13] and then to subsequently ignore the fact that age acts as a confounding variable in the presentation of health status scores. Other characteristics of the distribution of patient scores, such as skewness, are seldom mentioned. Yet, even simple frequency distributions could be illuminating when trying to understand what the numbers resulting from the measurement

[12] In this case, the relationship in question is between the nominal/categorical variable having (or, not having, in some applications) a given disease or illness and measured health status.

[13] To some degree the problem stems from the lack of validation work in respect of many health status measures. Hadorn and Hays (1991, p. 830) elaborate as follows:

> 'Also needed, but uniformly lacking to date, is a rigorous assessment of the construct validity of quality-of-life instruments. Construct validity is supported if a measure "behaves" as it is expected or hypothesized to behave. Thus, ratings of self-reported quality-of-life should be correlated with such demographic and clinical factors that can reasonably be expected to influence overall HRQOL [health-related quality of life]. This is an example of a popular means of construct validity testing wherein correlations are assessed between response levels and any of several measurable conditions or factors expected to affect constructs underlying those responses.
>
> 'For example, elderly persons and patients with significant medical illnesses should, other things being equal, report more physical suffering (eg, pain, nausea, shortness of breath), limits in important activities, and other impairments in quality of life compared with younger and illness-free individuals. That is, age and illness are expected to be negatively correlated with physical aspects of quality of life. The construct validation of a questionnaire that failed to detect a difference in reported quality of life across age and illness groups would be suspect.'

Figure 12: Overall Sickness Impact Profile (SEP) scores for different disease conditions or population groups



Source: Patrick & Deyo, 1989. Reproduced with permission.

process mean. For example, how does a SIP score in the range 0-4 compare with one in the range 5-9, 10-14, 15-19, or 45-49? Knowing the proportion of, say, home haemodialysis patients whose total SIP scores lie within these (for now arbitrary) ranges and their medical and other attributes may allow us to articulate better the health status continuum and to specify when a difference in scores translates into concrete differences in what patients can and cannot achieve. This is a discriminative index analogy to the question, “when is a change a change?” raised earlier with respect to evaluative indexes; it is the question when is a difference a difference. Not many cross-sectional studies present the descriptive data behind their summary statistics or give us much insight into the meaning of between-subject variability. Two conspicuous exceptions are provided in the studies by Hart and Evans (1987) and Deyo, Inui, Leininger and Overman (1982). Overall, it seems to us that scant attention has been paid to the question of generalizability in the literature of health status measurement. It is a situation that can and should be redressed.

Cultural Considerations in Generalizability

The issues regarding generalizability of health status scores raised in the preceding section apply across the board. However, when the question of generalizability across different cultural groups is raised, the issue is at once more complicated. The question of whether it makes sense to compare culturally-distinct groups in terms of specific scores on a health status scale is not entirely straightforward because the question of the cross-cultural validity of the health status assessment instrument is a logically prior one. The interpretation of cross-cultural differences in scores on health status measures involves a serious dilemma. Are the results a valid indication of differences in health status between subpopulations, or should they be explained in terms of bias or incomparability of the data?

The relevance of culture to the assessment of health status has not escaped health services researchers entirely (eg, Deyo, 1984; Gilson, Erickson, Chavez, et al, 1980; Hendricson, Russell, Prihoda, et al, 1989; Hunt, 1986; Hunt, McEwen & McKenna, 1986; Hunt & Wiklund, 1987; Patrick, 1981; Patrick, Sittampalam, Somerville, et al, 1985; Wiklund, Romanus & Hunt, 1988), though, to date, it has not been a high priority among developers of standardized measures. Accumulating evidence of the validity of a health status measure in one culture is, in and of itself, time-consuming. Still, knowledge of the cultural relativity of standardized health status items is important for understanding the nature of measured differences in health and illness among different populations and cultures and for designing and evaluating health interventions for different target groups (Patrick et al, 1985).

Health services researchers have long postulated that cultural factors contribute to the illness-disease distinction (eg, Fabrega, 1975; Kleinman, Eisenberg & Good, 1978). Angel and Throits (1987), point to the role that culture may play in explaining differences in self-reported health status. They write that,

> ‘there is only an imperfect correspondence between the clinical fact of disease and the subjective experience of illness and … while the clinical characteristics of a disease are culturally invariant, the phenomenological experience of illness is highly variable. One obvious explanation for this variability is that culture influences the experience of illness’ (Angel & Throits, 1987, p. 466).

Campos and Johnson (1990) are of like mind. They, too, see culture as a moderating variable:

> 'it is clear that the presence or absence of *disease* (measurable biophysiological abnormality) is not as important in predicting quality of life as is *illness* (the subjective psychological and social distress caused by symptoms). …
>
> '… by utilizing *subjective* measures to assess quality of life, individual, social, and cultural factors are automatically included in the analysis, since individuals' subjective experiences result from complex interaction of variables at all three levels' (Campos & Johnson, 1990, p 167).

Certainly, Ware's (1984a) framework for discussing disease and its impact, shown in Figure 13, is compatible with culture's being a moderating variable, as these writers suggest. The framework parallels the dimensions of health summarised in Table 2. It puts disease at the centre of a series of boxes that radiate outwards to indicate the different types and levels of impact as experienced by the individual.

Given that comparability is the *raison d'etre* for discriminative indexes of health status, it is not sufficient simply to note any differences (or the lack thereof) in the health status scores of different ethnic or cultural subpopulations. At least one further step must be taken. It involves establishing that observed differences (or equalities) in the health status scores of different cultural groups *are* real and not simply an artifact of the cultural content and value orientation of measures and items used.[14] This step is fundamental to the interpretation of the scores and to demonstrating the cross-cultural validity of the instruments concerned.

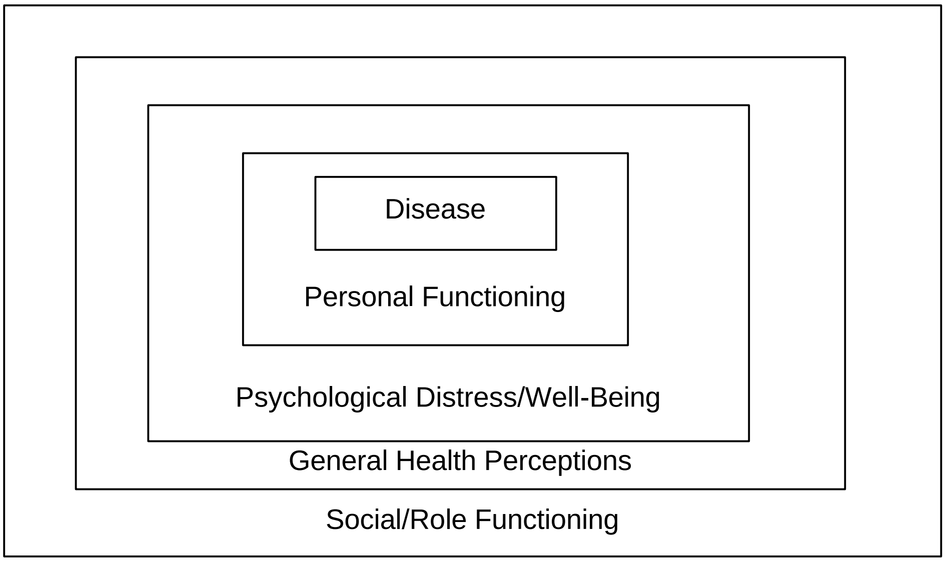
Although a number of researchers have written about the concept of "cross-cultural validity" as a prerequisite for comparisons across cultural and ethnic boundaries, many such discussions occur at cross-purposes. Often their only commonality is a shared point of departure as they refer to different spheres where equivalence across cultures must be demonstrated rather than assumed. In an attempt to summarise the ideas associated with equivalence in cross-cultural measurement, Hui and Triandis (1985) present a four-fold classification of types of equivalence:

Conceptual equivalence: A first and minimal requirement for cross-cultural comparison is that the construct exists in both cultures. A construct that can be discussed meaningfully in the cultures concerned is said to have cross-cultural conceptual equivalence. (If the construct is meaningful and relevant to individuals in both cultures, problems of cultural relevance are confined to items used to operationalize the construct).

This precondition may not be too demanding for constructs of health status where the items in the construct are well-defined and clinically interpretable, at least in populations where survival is not borderline. However, clearly with the broader concept of quality of life, which is largely attitudinal and value-based, it is decidedly problematic.

[14] Similar considerations may apply to evaluative indexes. To the extent that cultural factors contribute to differences between individuals in the amount of change in response to treatment it may be more difficult to detect an overall treatment effect (see Streiner & Norman, 1989).

Figure 13: Framework for discussing disease and its impact



Personal functioning: 'the performance or capacity to perform the kinds of tasks that most people do every day, including self-care, mobility, and physical activities'.

Psychological distress and well-being: or mental health, this level 'gets at feeling states that cannot be known merely from observing or asking about behavior.'

General health perceptions: refer to self-ratings of health that encompass 'an individual's evaluation of personal health along the lines of the three preceding health concepts (disease status, personal functioning, and mental health)'.

Role functioning: 'refers to the performance or capacity to perform activities associated with an individual's usual role, including formal employment, school work, or homemaking'.

Source: Adapted from Ware, 1984a.

Equivalence in construct operationalization: A second requirement concerns the transition from theory to measurement. If a construct is operationalized using the same procedure and is equally meaningful in the cultures being studied, the resulting instrument is said to be equivalent in its construct operationalization across cultures. For example, in operationalizing physical functioning, reference may be made to the performance of, or capacity to perform, certain activities. These activities should be equally meaningful and strenuous in both cultures.

It is possible that a particular item in a scale assessing, say, role functioning or the social health dimension of the construct health status, could be well-translated but ask for the likelihood of behaviours that would not be exhibited under any conditions in the target culture. The use of such items in a scale would generate different results between the source and target languages – because they represent *emic,* or culturally specific manifestations of illness behaviour, in contrast to *etic,* or culturally general responses to ill health (Hulin, 1987).

Item equivalence: Item equivalence is achieved if the two preceding types of equivalence obtain *and* the construct can be measured by the same instrument. The litmus test here is whether each item means the same thing to subjects from Culture A as it does to those in Culture B. This is a matter of the cross-cultural goodness of the translation.[15]

Scalar equivalence: An instrument can be said to have scalar equivalence if the other types of equivalence obtain and if it can be shown that the construct is measured on the same metric. Scalar equivalence requires that a value on a scale reflects the same degree, intensity, or magnitude of the construct regardless of the subpopulation from which the respondent is drawn.

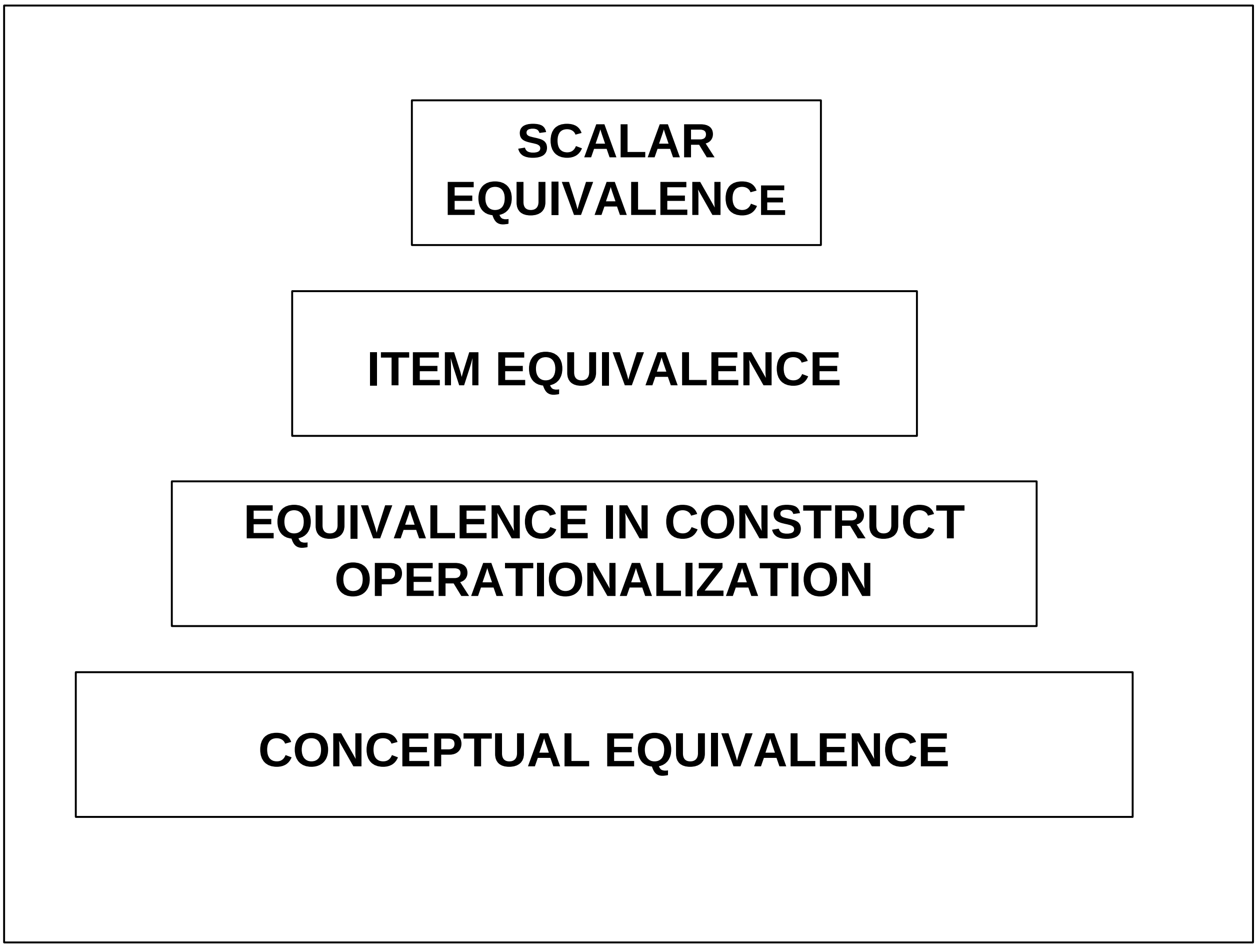
The progression of assumptions between successive types of equivalence is reflected in the hierarchy of Figure 14 and in the following quote from Angel and Throits,

> ‘Psychometric scales are constructed with the assumption that individuals can be arrayed along a common metric in terms of health or distress and that individuals with similar scores are similarly healthy or distressed in some objective sense. Indeed, the entire methodology of scale construction is based on the assumption that there is a phenomenological equivalence which can be identified and compared with the appropriate instruments.' (Angel & Throits, 1987, p 486).

Having only conceptual equivalence and equivalence in construct operationalization, as *may* be the case where the assessment of health status is concerned, does not constitute a psychometric instrument that is useful for cross-cultural comparison. Hui and Triandis indicate that to ‘legitimately attain informative quantitative comparison between two cultures, one needs to check whether the measures have cross-cultural equivalence in all four aspects' (1990, p 135).

[15] According to Hulin (1987, pp 122-123) ‘for translations of psychological instruments, high-fidelity reproduction of the source language input includes, in addition to normal linguistic considerations, similar measurement characteristics of items and scales as evaluated by appropriate procedures for assessing psychometric equivalence across subpopulations. … *psychometrically equivalent items (stimuli) evoke a specified response, from the set of permissible responses, with the same probability among individuals with equivalent amounts of the characteristic assessed by the item or scale comprising the items (stimuli).*' Hulin's concept of psychometric equivalence would seem to combine Hui and Triandis' categories of item and scalar equivalence.

Figure 14: Levels of Equivalence



Source: Adapted from Hui & Triandis, 1987.

In the aggregate, there are many strategies that can be engaged to test whether these types of equivalence hold. For the most part, however, they are not close substitutes. They are differentially predisposed to demonstrate, reject or presuppose, equivalence assumptions. Needless to say, testing a particular equivalence assumption is meaningful only when the underlying assumptions are accepted.

Hui and Triandis present a normative model of the relationship between strategies and equivalence assumptions (see their Figure 1, p 147) that is very instructive. For each strategy, it indicates which equivalence assumptions the strategy presupposes, which ones it can help demonstrate or improve, and which ones it doubts or rejects.

We shall not discuss their model here (but commend it to interested readers). We would, however, like to make one brief observation based on it. It is in the nature of a cautionary tale. Lest we succumb to the temptation, it is well to note that the most prevalent strategy used in cross-cultural comparisons – “crude” translation and direct comparison using t-tests or MANOVAs – *presupposes* all levels of equivalence. Sophisticated translation techniques such as back-translation, bilingual and committee approaches, decentering and pretests (see Chapman & Carter, 1979; Sechrest, Fay & Saidi, 1972) are required to demonstrate or improve equivalence in construct operationalization and item equivalence.[16]

6. CONSEQUENTIAL BASIS OF VALIDITY

So far our discussions of validity issues have been concerned primarily with addressing the question of whether an instrument that purportedly measures health status is a good measure of what it will be interpreted as assessing ie, health status. This is a scientific and technical question; the focus is almost exclusively on psychometric adequacy, especially construct validity. However, it may be argued that this presents a narrow view of test validity. There is also an ethical dimension to testing that calls for evaluation of the consequential basis of test use to which we should pay heed. Cronbach (1988, p 6) recognises this dimension in his observation that ‘tests that impinge on the rights and life chances of individuals are inherently disputable’; with the counsel that ‘validators have an obligation to review whether a practice has appropriate consequences for individuals and institutions, and especially to guard against adverse consequences.’

Messick (1975, 1980, 1981, 1988, 1989), in particular, has consistently argued that test validity should be construed broadly because ‘the process of construct interpretation inevitably places test scores in both a theoretical context of implied relationships to other constructs and a value context of implied relationships to good and bad, to desirable and undesirable attributes and behaviors’; and ‘the process of test use inevitably places test scores in both a theoretical context of implied relevance and utility and a value context of implied means to ends’ (1988, p

[16] As a matter of convention, translations are usually classified according to the linguistic purposes of the translation using a schema formulated by Casagrande (1954) and elaborated by Brislin (1976) viz. pragmatic, aesthetic-poetic, ethnographic or linguistic. Hulin (1987) argues convincingly that translations of psychometric scales and items from a source language to a target language have different goals from those associated with standard classifications of translations, and that the methods of assessing the degree to which these goals are achieved should therefore be different.

41). Thus, *test validity* is defined as 'an evaluative judgment of the adequacy and appropriateness of inferences *and actions* based on test scores (1988, p 42, emphasis added).

According to Messick (1981, 1988), there are four questions that should be addressed explicitly whenever a test is proposed for a particular purpose. These questions correspond to four aspects of a unified concept of test validity which obtains when two interrelated facets of validity are crossed. The first is the source of justification of the testing which will be based on appraisal of either evidence or consequence. The second facet is the function or outcome of the testing, which may be either its interpretation or use. The resulting matrix is shown in Table 5 and the four questions (reading from left to right, and top to bottom) are (Messick, 1981, p 9):

(i) 'What evidence justifies the proposed test interpretation as balanced against counterevidence or evidence supporting plausible rival interpretations?'

(ii) 'What evidence justifies the proposed test use on contrast to evidence supporting alternative proposals based on other measures or methods, including nontesting methods?'

(iii) 'What are the value implications of the preferred test interpretation and to what degree are its value implications and theoretical implications compatible or antagonistic?'

(iv) 'What are the potential social consequences of the proposed test use and to what degree are they facilitative or debilitative of the intended purpose?'

Though interrelated, the questions are nonetheless quite distinct in the sense that good answers to any one question do not constitute satisfactory answers to any of the others.

Table 5: Facets of Validity (Progressive-Matrix Formulation)

	TEST INTERPRETATION	**TEST USE**
EVIDENTIAL BASIS	Construct validity	Construct validity + Relevance/utility
CONSEQUENTIAL BASIS	Construct validity + Value implications	Construct validity + Social consequences

Source: Adapted from Messick, 1988.

As shown in Table 5, the evidential basis of test interpretation is construct validity. Construct validity also provides the cornerstone for the evidential basis of test use, but it is supplemented by evidence about the relevance of the test to the specific applied purpose and for the utility of the test in the applied setting. The consequential basis of test interpretation involves the appraisal of the value implications of the construct labels themselves, the value connotations of the nomological networks and still broader theories in which constructs are embedded. Finally, the consequential basis of test use involves the appraisal of the potential social consequences of the proposed use and the actual consequences of applied testing.

Messick's "extended" validity framework seems to us to have clear relevance to the assessment of health status and health-related quality of life, given the instrumental character of many such exercises. Take first the value implications associated with the measurement of health-related quality of life – the consequential basis of test interpretation. One leading implication, indeed in many respects the very point of departure for measuring health-related quality of life, is a recognition of the fact that patients are more than simply vectors of "objective" biochemical and physiological data; and that disease (the presence of measurable biophysical abnormality) and illness (the subjective psychological and social distress caused by symptoms) are two quite different concepts for purposes of treatment and patient management. A corollary of this development is an appreciation of the utility of "subjective" measurement[17] of health status, especially health profiles, and a tacit assumption that non-invasive research procedures are safe and pose no risks to patients (Carter & Deyo, 1982). According to McDowell & Newell (1989, p 15), subjective measurements hold several advantages, not the least of which is that 'they may offer a systematic way to record and present "the small, frantic voice of the patient" [Elinson, 1978].' Legitimation of data regarding the patients' experience of disease may help assert the patient's perspective vis a vis that of the physician where issues like non-compliance arise. It also underscores the conundrum associated with screening asymptomatic populations for risk factors (cf screening for early detection of disease).

In discussing the consequential basis of test interpretation, Messick (1981, p 12) argues that 'the process of test interpretation not only places test scores in a theoretical context of implied relations to other constructs … but also in a value context, possibly multiple value contexts, of implied relations of good and evil.' Such value connotations may be more or less subtle and seldom are they explicitly expressed in the development of health status measures or in reports of their application although clearly actions and consequences, in terms of either policy directions or patient treatment and management, should follow from the research findings.

One area where the underlying values and consequences have been clearly expressed is the development and application of quality-adjusted life years (QALYs) for use in resource allocation decisions. Here economists have been very explicit indeed in spelling out both the policy effects of adopting a cost-per-QALY decision criterion for choosing among alternative health care interventions and the underlying value assumptions, including that of the explicitness of the decision-making itself. For this reason, the cost-per-QALY metric in resource allocation decisions is used to address the general question of the consequential basis of validity[18] though it should be stressed that the question is pertinent to all approaches to health status assessment.

In their discussion of the proposal by economists to use the cost per quality-adjusted life years in this way, Mulkay, Ashmore and Pinch (1987) seem to argue that the credibility of the exercise is augmented by the rhetoric of economics, and its implied relations to good and evil:

[17] The term "subjective measurement" is used here to refer to measurements of individual health that rely on indicators in which a person (the patient *or* a proxy, including the physician) makes a judgement that forms an indicator of health (McDowell & Newell, 1989). Traditionally, the term "subjective" has been widely associated with the pejorative label "soft data" especially when the source of the observation about a patient is the patient rather than the physician. 'Thus, a doctor's observation of whether the patient has tenderness in a knee is regarded as objective and is therefore harder than the patient's observation and report of the pain experienced in the knee when he walks.' … [In fact, however,] 'the crucial attribute for hardness is reliability, which can be attained even when the observation is subjective, non-preservable, and non-dimensional' (Feinstein, 1977, p 489).

[18] Throughout the following discussion we have chosen to quote liberally from "primary sources" as it were, rather than to paraphrase and summarise the ideas and arguments put. Although this practice is considered inelegant by some, we consider it preferable to let researchers speak for themselves in the context of a discussion of value implications and social consequences of health status measurement.

> 'Although health service managers provide the starting point … the practical task of deciding on the allocation of health service resources is transformed by the economist into an exercise in cost/benefit analysis. The course of action required by the health service managers is treated in the course of economic analysis as following necessarily from the balance between costs and benefits as conceived in that analysis. The benefits or outcomes of various forms of medical treatment are defined by the economist in quantifiable terms as the effect on patients' expectation of life with adjustments made for variable effect on patients' quality of life. Costs are defined, in principle, to include monetary and intangible costs to patients and their families as well as monetary costs to the NHS. …
>
> 'Thus the [health service] managers' practical task is reconstituted in the economist's text in such a way that it becomes solvable on the basis of a simple economic metric. This metric generates recommendations because its basic terms carry strong normative weight. No rational actor, it is implied, would incur greater cost than was necessary; nor would s/he refuse additional benefit which was available at no further cost. These assumptions are built into the very meaning of the terms 'cost' and 'benefit'. They are part of the semantics of economic analysis. For instance, if actors chose not to accept what was thought to be additional benefit, it would follow necessarily either that this was in fact not a benefit for them after all or that other aspects of the situation, which had been ignored, constituted benefits or costs in their eyes. This is one reason why the economic metric is immensely powerful and persuasive. Once complex administrative decisions have been reduced to simple, and usually quantified, comparisons of cost and benefit, it comes to seem irrational (or improper, if individuals choose to pursue their private ends rather than the public good) not to act in accordance with the numbers' (p 544).

Another set of values is implicit in the (perhaps unconscious?) metaphor that at least some advocates of QALYs frequently use when they refer to the investment potential of health care and patients alike. Health care is apparently seen as an investment good rather than a consumption good; interventions generate a stream of QALYs, rather than cash flows, to be discounted. The state is entreated to act as a QALY maximizer akin to the profit-maximizing entrepreneur. Thus:

> 'Resources in such a system go to those patients whose care creates the largest volume of health (QALY) benefits. Those patients who are poor *investments* in terms of generating QALYs do not get care.' (Maynard, 1987b, p 113, emphasis added)
>
> 'When making decisions on allocation we need information about costs and benefits (outcomes). Before considering how to identify and measure these variables, however, it is necessary to decide what data are actually needed for decision making.
>
> The managers, administrators, and clinicians of a health care system typically generate data in terms of totals and averages (arithmetic means). The first rule for achieving economic efficiency with such data is:
>
> (1) (a) if total costs do not exceed total benefits do not *invest* in the procedure
> (b) if total benefits exceed total costs do *invest* in the procedure.
>
> 'This does not, however, tell us what amount to *invest* if rule 1(b) is met. The advice of the economist is that the decision on the level of *investment* should be decided by using

data about the margin. The margin is the increment added to either costs or outcomes by a small (one unit) change in the level of activity. Thus the decision maker needs data on the marginal cost of producing one more (or one fewer) hernia repair and the marginal benefit to the state of health of none more or fewer of such a procedure. From these data the second rule may be derived to decide the efficient level of *investment* in an activity:

2. (a) if marginal cost exceed marginal benefit reduce investment;
   (b) if marginal benefit exceeds marginal cost increase investment;
   (c) when marginal cost equals marginal benefit stop *investing* and maintain that efficient level of output.

> 'The logic of these rules is quite simple. Rule 2(a) indicates that if, as a result of increasing activity by one unit (the margin), the value of opportunity costs (foregone alternatives) exceeds the benefits of the procedure *investment* could be made more productively elsewhere and the level of activity should be reduced. On the other hand, if the benefits (in terms of incremental effects on the state of health) exceed the costs of one more operation *investment* in that procedure should increase. The efficient level of activity is where costs and benefits are equal at the margin.' (Maynard, 1987a, p 1154, emphasis added)

The potential social consequences of pursuing this analogy, as discussed below in connection with the consequential basis of test use, strike many as particularly unsalutary.

Similarly, value implications are evoked by the frequent references that advocates of the cost-effectiveness approach make to the public or *explicit* character of the rationing that they envisage (eg, Daniels, 1991; Dixon & Welch, 1991; Dowie, 1991; Eddy, 1991a, 1991b; Hadorn, 1991). The deprecatory, even contemptuous, regard in which implicit rationing is held is illustrated by the following quotations:

> 'Priorities have "to be determined by social judgements about need." At present, allocations are determined largely by doctors and the basis of their choices is implicit, incoherent, and inconsistent. The economic approach to "unpacking" this covert process is to measure (guesstimate?) the costs and outcomes of competing therapies in an explicit framework which is subject to challenge and debate. At its most provocative, and as an incentive to its opponents to do better (rather than throw out the baby with the bath water), it creates guesstimate cost-outcome "league tables". Such information informs the rationing process based on need or benefit to patients, and would direct resources away from relatively cost ineffective (ie, high cost per quality adjusted life year [or QALY] procedures such as liver transplants and breast cancer screening. Such choices are unavoidable and made every day in the NHS where patients are denied beneficial treatment because of resource constraints. The economic approach to rationing is explicit: it makes what is at present inconsistent, incoherent and implicit, open to democratic review' (Maynard, 1990, p 188).

> 'The article by Aaron and Schwartz [1990] does not point out the significant difference between the British and Oregon models of rationing health care. Rationing in Great Britain has been implicit, not explicit, as public input routinely has been excluded from the process. It is a silent conspiracy between a dense, obscurating bureaucracy, intentionally avoiding written policy for macroallocation (rationing), and a publicly unaccountable

> medical profession privately managing microallocation so as to conceal life and death decisions from patients.
>
> ‘The Oregon approach is open, starting with citizen values for health care and building through expert advice toward legislated, health care rationing policy. … Understanding the difference between the two systems is critical in obtaining active endorsement of health care rationing by the community, rather than passive acceptance. … This approach avoids both corrupting the medical profession with responsibilities antithetical to the profession’s ethical duty and burdening civil servants with life and death decisions fostered from the top down’ (Cranshaw, 1990, pp 661-662).
>
> ‘Whether or not we accept the word rationing, it must be acknowledged that underfunding of services, long queues, and “never-never lists” are an implicit means to limit services. The Oregon experiment offers an alternative; make the choices explicit, and base them on systematic rather than ad-hoc methods’ (Dixon & Welch, 1991, p 893).

These statements rely more or less heavily on the *“ipso facto* case” for explicit rationing, in particular, the generally favourable value implications of the term “explicit”. The upshot is that the question of whether the rules that govern rationing should be publicised or kept secret has been little debated – despite the fact that publicity *is* controversial (Calabresi & Bobbit, 1978; Friedman, 1986; Rhoads, 1980; Winslow, 1986). For example, it has been argued that “tragic choices” are best made out of the public view in order to sustain important symbolic values, such as the sanctity of life, and that the usefulness and social acceptability of many allocation schemes depend on the ‘charade that they serve the purposes that they say they do’ (Calabresi & Bobbit, 1978, p 24). Perhaps the increasing impetus for a switch from explicit to implicit rationing simply reflects a “strategy of cycles”?[19]

What is explicit? We need to be wary of claims of explicitness where the measurement of health-related quality of life is concerned. Not infrequently, there is a gap between what we are invited to accept at face value and what *can* be accepted at face value, and some claims are, frankly, quite disingenuous. The issue of what is and what is not explicit seems to explain in part the reluctance to put too much store in summary scores. Mosteller, for instance, intimates that,

[19] Calabresi and Bobbit (1978, pp 195-199) argue that the strategy of cycles is a way that society copes with tragic choices in the face of being forced to choose among competing values. To elaborate:

> ‘Why do approaches to tragic allocations change? Such changes are not mindlessly made; they have, in fact, represented quite rational responses preceded by discussions as rational as discussions termed rational usually are. The criticisms of the pre-existing system have described in generally accurate detail its fundamental flaws and have invoked the basic values which that system degrades. But the defenders of the pre-existing system are just as rational. They are penetrating in their recognition of the flaws inherent in the proposed reform. And when the reform is accepted and had become the vested method, it is eventually seen to display the very shortcomings which its critics had predicted (and to degrade those values they had sought to protect). Are these *mistakes*? If they are not, why do we move restlessly from one system which proves inadequate to another?
>
> ‘The answer is, we have come to think, that a society may limit the destructive impact of tragic choices by choosing to mix approaches over time. Endangered values are reaffirmed. The ultimate cost to other values is not immediately borne. Change itself brings two dividends, though all too often of an illusory kind we have associated with subterfuges. First, a reconceptualization of the problem arouses hope that its final price will not be exacted; the certainties of the discarded method are replaced. Second, the society is acting, and action has some palliative benefit since it too implies that necessity can somehow be evaded if only we try harder, plan better than those we followed, avoid their mistakes, and so forth. More important, because more honest, the deep knowledge that change will come again carries with it the hope that values currently degraded will not for all that be abandoned’ (pp 196-197).

> ‘My personal experience on committees with talented laypeople over the last 2 or 3 years has discouraged me quite a bit about such measures as quality adjusted life years (ie, not about years *per se* but about our approaches that give us a tricky summary like quality-adjusted life years). These laypeople were very willing to make choices between medical technologies. They wanted to know what different technologies will produce for different groups of people and what the benefits and losses would be, but they do not like to have these complicated problems summed up in single numbers. In using quality-adjusted life years or other cost-benefit analysis summaries, they felt something was being concealed from them, and they did not understand how the work was being done. They were right about that; they did not understand the cost-benefit approach. That does not mean that they did not understand the general idea or that they were unwilling to make hard choices. They were willing and able to look at one set of benefits and losses and to compare that with another set of benefits and losses and to try, in their own minds, to put those together and make a comparison. They were *not* willing to accept somebody else’s summary numbers. This practical experience raises difficulties for the application of some of our strong economic methods when they have to have public approval and satisfaction. I do not know about your experience, but you might want to think about mine.’ (Mosteller, 1989, pp S285-286)

It is not central to the present discussion whether the apprehension about the aggregation of health profile scores into summary indexes to which Mosteller alludes is reasonable, though it very well may be for the reasons Ware (1989) articulates:

> ‘When we aggregate to achieve the simplicity of a single health number, we must remembers what is lost. There are many different profiles of health concepts that, when combined, lead to the same number. Further, the same profile can produce many different aggregate scores depending on weights (values) that are used in aggregation. Pending much needed advanced in understanding of these issues, it may be best to analyze a profile of health measures and those measures in the aggregate using alternative weighting schemes. If the conclusions are sensitive to methods, that story should be told’ (p S287).

The point is that explicitness is an end-product of an understanding of “from whence the numbers come”. As Carr-Hill (1991, p 361, emphasis added) elaborates:

> ‘We might agree that all [assessments of individual activities, procedures, or programs] should relate to a common set of dimensions – for example, the ability to function without pain or anxiety – but not agree on how they should be combined. *The point is that index numbers are not an observation upon the world: they are generated and produced by a specific set of technical procedures*, and the QALY is no exception. In turn, technical procedures are not neutral: they serve different interests (Carr-Hill, 1982; Seers, 1975), and this should be made explicit.’

Lurking not far behind the suspicion that index scores may conceal more than they reveal is a foreboding that the resource allocation decisions may become subject to the tyranny of numbers. Smith (1987, p 189) for one has voiced the concern that ‘once complex administrative decisions have been reduced to simple, and usually quantified, comparison of costs and benefits, it would seem irrational … not to act in accordance with the numbers’ (Smith, 1987, p 1135); and Evans (1990, p 189) argues that the ‘attraction for the politician of outcome measures such as the

QALY is the prospect of turning difficult political decisions about the size and distribution of the health budget into routine technical procedures.'

A defensible analogy can be made to Gould's (1981) devastating criticism of the rise and rise of I.Q. testing in the first half of this century under the auspices of influential researchers such as Burk, Spearman, Terman and Thorndike. Despite the fact that "intelligence" was derived as a single number from exploratory factor analysis, one of the least "hard" statistical techniques, and was conspicuously devoid of physiological underpinnings, it was delivered to the public and policy-makers as an innate component of human physiology. The consequences were socially appalling: they included the early sorting of children for educational purposes (as in the 11+ examination in Britain), the drafting of the Immigration Restrictions Act (1924) in the U.S.A. and support for theories of inherent intellectual inferiority of certain races (especially blacks) and the poor (see, for example, Block & Dworking, 1976; Blum, 1978; Eckberg, 1979).

A point emphasised and re-emphasised by Gould was not merely the certitude with which the "science" of intelligence measurement was presented, but the *reification* of intelligence – the conversion of a mental or psychological test, a construct, into a material entity. Much of the literature on quality of life measurement is disturbingly analogous.

Although there are social consequences that follow from the use of other health-related quality of life measures, the link between test scores and social consequences that is associated with the consequential basis of test use, the last cell in Table 5, is most easily made with reference to QALYs. Given that the raison d'etre for the cost per QALY metric is to aid resource allocation decisions and decisions have consequences, the linkage is unmistakably direct.

To a very considerable degree the deleterious consequences that many commentators foresee with the application of QALYs stem from their fungibility across categories of patients, diseases and treatments. Thus, the very aspect of QALYs that makes them so attractive to health economists, viz. their usefulness for making such 'global' comparisons, detracts from their utility for their critics who see some value in using them to make 'local' comparisons. Thus, Smith (1987, p 1135) considers that

> 'the use of QALYs to assist decisions about which treatments are of most benefit to patients with a particular disease is a potentially useful refinement of techniques already used in the assessment of procedures having a possible bearing on the outcome. If the QALY technique can be developed so as to avoid the methodological difficulties described above [in particular, the assumption of reciprocal commensurability], decisions between alternative treatments might well be enhanced. However, it is quite another matter to apply the QALY or any other cost-effectiveness approach to the problem of deciding which patients to treat. Decisions between treatments that are relevant to quite different diseases are essentially decisions about whom we should treat.'
>
> Evans (1990, p 188) expresses the same kind of reservation:
>
> 'There are technical and philosophical problems which need to be resolved before it would be justifiable to institutionalise the QALY or any other measure of average outcomes in practical health service planning. Insofar as such a measure is to be used for choosing between two different treatments for the same disease in the same types of patients the problems are largely technical. The philosophical problems dominate if a

> measure of average outcome is used to decide, directly or indirectly, which patients are to be treated.' (Evans, 1990, p 188)

Up to a point this stand-off may seem to simply rehearse some points made previously in relation to the choice of disease-specific versus generic measures of health status. There is a difference, however. Instead of debating which kinds of instruments will best facilitate drawing inferences about the efficacy of treatments in a research context, this time we are talking about the real consequences that may follow from the application of these results, about the sequelae of an approach wherein 'the probability of treatment moves discontinuously from 0 to 1' (West, 1985, quoted in Carr-Hill, 1991, p 360). These consequences are well known and have been identified by many commentators in more or less emotive terms. Thus:

> 'In choosing which patient should receive treatment the average outcome approach will, other things being equal, give preference to those with most years of remaining expectation of life to offer. Not only will this mean treating younger patients rather than older; it should also mean treating women rather than men, white patients rather than black, and upper social class patients rather than lower' (Evans, 1990, p 189).

> 'A choice of whom to treat based on any form of cost-effectiveness assessment will always favour patients whose age or disease confers the prospect of longer and better-quality survival. Old and very sick patients will be placed by resource allocation decisions in a position of double jeopardy.' (Smith, 1987, p 1135).

> 'Triage in time of war can be represented as a pointed illustration of the implementation of the principle [of QALY maximization]. In general, priority is given to those needing a quick, straightforward life-saving operation and anyone with multiple wounds for whom little can be done may, in consequence, be left to die. This policy may often be motivated by the need to return as many men as possible to the front-line, but it will produce incomparably more QALYs than the selection of patients on a more equitable basis – for example, by drawing lots' (Cubbon, 1991, p 182).

> 'Since Cubbon invites a peacetime analogy, the most likely expression of a QALY-motivated health policy is as one designed to get the productive back to work and leave the useless to suffer and die. …
>
> 'Echoing Hobbes again, triage and QALYs are part of the philosophy of permanent war in which the good guys are the fortunate for whom long and healthy life-expectancy can be cheaply provided. The enemy are those unfortunates who stand between the fortunate and their survival by daring to make rival claims on our concern and our resources' (Harris, 1991, p 187).

In the fact of disquiet with such (presumably) adverse social consequences the reflex reaction of proponents of QALYs is to offer one of two defences. One is to acknowledge the difficulty, disavow throwing the baby out with the bathwater, and suggest that varying weightings may be attached to the lives of members of different population groups. A second tactic is to go on the offensive by contrasting their analysis, which they acknowledge to be imperfect, with the status quo of no analysis at all, and to offer to pit their approach against the present one. The latter rejoinder simply erects a straw man since there is no real argument that we need better outcome measures (Carr-Hill, 1991). The first response is also problematic because, apart from the underspecified offer to paper over a crack, it misses the point (as does the second retort) that many critics do not agree that there is a baby in the bathwater in the first place. Such critics query whether there is sufficient evidence for the construct validity of methods of measuring QALYs. Messick includes the concept of construct validity in each cell of the matrix of Table 5 advisedly.

As suggested above, the social consequences of non-QALY measures of health status are not as transparent as with QALYs. Nor have they exercised the minds of commentators as much. Still one area where there is the potential for adverse social consequences clearly exists. It relates to the fact that the patient is usually focal when, in fact, it is apparent that ‘the quality of life of many people other than the patient is determined in large parts by the care provided’ (Hollandsworth, 1988, p 430). It seems not inconceivable that cases may arise where the care setting and/or regimen that has the most favourable impact on the patient’s health-related quality of life may impose demands on family members that are either unreasonable or unsustainable in the long term. Certainly, the present predisposition toward cost-shifting between the informal and health care sectors favours the trailing of such interventions and we should be alert to the possibility.

7. SUMMARY AND CONCLUSION

The motivation to write this report was a deep concern about the quality of many if not all of the existing measures of health status and health-related quality of life that are being used increasingly as the primary outcome measures in evaluation studies. With some few exceptions, these measures have not been subjected to serious, let alone adequate validation and this should be a matter of concern both to those working in the area and to users of reports emanating from empirical studies.

The report is intentionally limited in scope. It has not attempted to address some areas of importance such as scaling and sampling design which affect both internal and external validity. Most notably, it has not concerned itself with the paucity of references to the literature on survey methodology and allied fields of cognitive behaviour that are important to the veridicality of elicited responses to questions and the interpretation of presented scenarios that constitute the health states construct in holistic responses to utility and time trade-off questions. This alone warrants a major report because the structure and wording of many questionnaires and health state scenarios suggest that Fischhoff’s (1990) comment on contingent valuations in the economics literature, that ‘These economists have done a remarkable job of performing surveys without really knowing the literature’ and ‘fall short of the standards acceptable to research professionals’, may have wider application.

The thrust of the research report is a plea for recognition of the central importance of the construct we choose to call health status or health-related quality of life. If validity may be defined simply as measuring what we say we measure, the specified characteristics of health or well-being we choose to include in the domain, and the manner in which they are included, are critical. Matters of scaling and the psychometric manipulations of data that follow the elicitation of responses are, of course, important. Perhaps these aspects of measurement are more familiar to those who develop and use health status instruments because a respectable proportion of the measures have recognised them as matters worthy of comment and at least some testing. But the construct itself is more fundamental. Unless its validity can be demonstrated, one must, prima facie, reject all that follows, irrespective of how much attention is devoted to reliability of responses and analysis of results.

The basis of validity testing is supporting or defending inferences about attributes of people, not about statistical properties of scores. Therefore, validation studies are grounded in theory and revolve around hypotheses and hypothesis testing. Validation must be looked on, too, as a continuing process wherein theory informs and is informed by validations studies. It requires consistent support across many studies, many researchers and a variety of theoretically-derived variables.

The conventional approach to validation, a comparison of correlations among two operational definitions of constructs, is methodologically flawed because it involves circular reasoning and unsubstantiated assumptions about confounding variables in both definitions.

The standard methodology of carrying out the extended process of hypothesis testing is more complex and time-consuming. It recognises that constructs such as good health and well-being are necessarily imprecise and methods of eliciting and measuring responses are affected by extraneous factors. Moreover, it is seldom possible to obtain samples of individuals who unambiguously represent target populations and otherwise conform neatly with the requirements of statistical sampling theory. For all these reasons, classical test statistics are of limited value and the many recent validation exercises that report simple correlations among a few (sometimes two) alternative measures probably do more harm than good by lending credibility to poor research.

The multitrait-multimethod paradigm, which should be looked on as a methodological approach and not a set of steps to be followed, recognises the fuzziness of both construct and measurement method and simultaneously tests both by looking for convergent and divergent validity in constructs and theoretical predictions; and by seeking to establish both the boundaries of the construct and redundancies within those constructs. In so doing, it utilises a variety of statistical methods that yield results that do not permit the degree of certainty associated with classical test theory that are so conducive to peace of mind.

Nevertheless, it is this approach to continuous multiple testing of both methods and constructs that can lead legitimately to increased confidence in the validity of health status measures and, consequently, in the implications that one can draw from the results of empirical investigations in health services research and program evaluations.

Although the emphasis in the research report is on explaining and defending this basis of construct validation, for completeness three other related matters have been addressed.

The first concerns the obvious gap in much of the methodology literature and in many research applications: the question of the longitudinal validity of constructs. Where the research question is concerned with measurement of changes in health status over time, as it often is, the ability of the chosen instrument to measure change rather than discriminate cross-sectionally is critical and few measures have been so validated.

The second addresses cultural considerations in generalising results across populations. This is a matter of considerable importance in many countries, including our own, and one that is usually ignored entirely (with consequent diminution in external validity) or finessed by excluding potentially “troublesome” segments of the population.

The third extends the concept of validity beyond the simple notion of measuring what one purports to measure to encompass potential consequences of the inferences that may be drawn from test results and the actions that follow from such inferences. Because they are constructs, measures of health status are not value-free and the inferences drawn from research results necessarily reflect underlying value assumptions. The research report concludes with an examination of the assumptions, both implicit and explicit, embedded in the use of quality-adjusted life years for resource allocation, arguably the most value-laden application of health status measurement in the current research literature.

It is hoped that the foregoing will promote more awareness of the importance of validating health status measures before they are rushed into use for policy-related investigations. It is hoped, too, that it will provide some useful guidance to methodologies for carrying out such validation exercises and to the abundant validation literature.

8. REFERENCES

Aaron H & Schwartz WB 1990, Rationing health care: the choice before us. *Science*; vol 247, pp 418-422.

Aday LA, Chui GY & Anderson R 1980, Methodological issues in health care surveys of the Spanish heritage population, *American Journal of Public Health*, vol 70, pp 367-374.

Althauser RP & Herberlein TA 1970, 'Validity and the multitrait-multimethod matrix', in Borgatta EF & Bohrnstedt GW (Eds), *Sociological methodology*, pp 151-169. San Francisco: Jossey-Bass.

Anastasi A 1986, Evolving concepts of validation, *Annual Review of Psychology*, vol 37, pp 1-15.

Andrews FM 1984, Construct validity and error components of survey measures: a structural modelling approach, *Public Opinion Quarterly*, vol 48, pp 409-442.

Angel R & Cleary PD 1984, The impact of culture on the cognitive structure of illness, *Culture, Medicine and Psychiatry*, vol 11, pp 465-484.

Angoff WH 1988, 'Validity: an evolving concept', in Wainer H & Braum HI (Eds), *Test validity*, pp 19-32. Hillsdale, NJ: Erlbaum.

Arpin K, Fitch M, Browne GB & Corey P 1990, Prevalence and correlates of family dysfunction and poor adjustment to chronic illness in speciality clinics, *Journal of Clinical Epidemiology*, vol 43, pp 373-383.

Bauernfeind RH 1968, The need for replication in educational research, *Phi Delta Kappa*, vol 50, pp 126-128.

Baum FE & Cooke RD 1989, Community-health needs assessment: use of the Nottingham Health profile in Australia, *Medical Journal of Australia*, vol 150, pp 581-590.

Bentler PM 1985, *Theory and implementation of EQS: a structural equations program*. Los Angeles: BMDP Statistical Software.

Bentler PM 1990, Comparative fit indexes in structural models, *Psychological Bulletin,* vol 107, p 238.

Bentler PM 1990, 'Latent variable structural models for separating specific from general effects', in Sechrest L, Perrin E & Bunker J (Eds), *Research methodology: strengthening causal interpretations of nonexperimental data*, pp 61-83. Rockville MD: US Department of Health and Human Services, Agency for Health Care Policy Research.

Bentler PM & Bonnet DG 1980, Significance tests and goodness of fit in the analysis of covariance structures, *Psychological Bulletin*, vol 88, pp 588-606.

Bergner M 1987, 'Development, testing, and use of the Sickness Impact Profile', in Walker SR & Rosser RM (Eds), *Quality of life: assessment and application*, pp 79-94. Lancaster: MTP Press.

Bergner M 1990, 'Comment: Latent variable structural equation modelling in health services research', in Sechrest L, Perrin E & Bunker J (Eds), *Research methodology: strengthening causal interpretations of nonexperimental data*, pp 81-84. Rockville MD: US Department of Health and Human Services, Agency for Health Care Policy Research.

Bergner M, Bobbitt RA, Carter WB & Gilson BS 1981, The Sickness Impact Profile: development and final revision of a health status measure', *Medical Care*, vol 19, pp 787-805.

Bergner M, Bobbitt RA, Kressel S, Pollard WE, Gilson BS & Morris JR 1976, The Sickness Impact Profile: conceptual formulation and methodology for the development of a health status measure, *International Journal of Health Services,* vol 6, pp 393-415.

Bergner M & Rothman ML 1987, Health status measures: an overview and guide for selection, *Annual Review of Public Health*, vol 8, pp 191-210.

Bindman AB, Keane D & Lurie N 1990, measuring health changes among severely ill patients: the floor phenomenon, *Medical Care*, vol 28, pp 1142-1152.

Block NJ & Dworking G (Eds) 1976, *The IQ Controversy*. New York: Pantheon Press.

Blum JM 1978, *Pseudoscience and mental ability. The origins and fallacies of the IQ Controversy*. New York: Monthly Review Press.

Bombardier C, Ware JE, Russell IJ, Larson M, Chalmers A & Read JL 1986, Auranofin therapy and quality of life in patients with rheumatoid arthritis: Results of a multicenter trial, *American Journal of Medicine*, vol 81, pp 565-578.

Boyd NF, Sutherland HJ, Heasman KZ, Tritchler DL & Cummings BJ 1990, Whose utilities for decision analysis? *Medicine Decision Making*, vol 10, pp 58-67.

Bremer BA & McCauley CR 1986, Quality-of-life measures: hospital interview versus home questionnaire, *Health Psychology*, vol 5, pp 171-177.

Brinberg D & McGrath JE 1985, *Validity and the Research Process*. Beverly Hills: Sage.

Brislin R 1976, ‘Translation research and its applications: an introduction’, in Brislin R (Ed), *Translation: applications and research*. New York: Wiley/Halstead.

Browne GB, Arpin K, Corey P, Fitch M & Gafni A 1990, Individual correlates of health service utilization and the cost of poor adjustment to chronic illness, *Medical Care,* vol 28, pp 43-58.

Bush J 1984, ‘Relative preferences versus relative frequencies in health-related quality of life evaluation’, in Wenger N, Mattson M, Furberg C & Elinson J (Eds), *Assessment of quality of life in clinical trials of cardiovascular therapies*, pp 118-139. New York: Le Jacq Publishing.

Calabresi G & Bobbitt P 1978, *Tragic Choices*. New York: Norton.

Campbell DT 1960, Recommendations for APA test standards regarding construct, trait, or discriminant validity, *American Psychologist*, vol 15, pp 546-553.

Campbell DT & Fiske DW 1959, Convergent and discriminant validation by the multitrait-multimethod matrix, *Psychological Bulletin,* vol 56, pp 81-105.

Campos SS & Johnson TM 1990, ‘Cultural considerations’, in Spilker BF (Ed), *Quality of Life Assessments in Clinical Trials*, pp 163-170. New York: Raven Press.

Carmines EG & Zeller RA 1979, *Reliability and Validity Assessment*. Beverly Hills: Sage.

Carr-Hill RA 1982, ‘Social indicators for basic needs: who benefits from which numbers?’ in Cole S & Lucas H (Eds), *Models, Planning and Basic Needs*. Oxford: Pergamon Press.

Carr-Hill RA 1991, Allocating resources to health care: is the QALY (quality adjusted life year) a technical solution to a political problem? *International Journal of Health Services*, vol 21, pp 351-363.

Carter WB & Deyo RA 1982, The impact of questionnaire research on clinical populations: a dilemma in review of human subjects research resolved by a study of a study, *Evaluation Studies Review Annual*, vol 7, pp 287-295.

Carver RP 1974, Two dimensions of tests: psychometric and edumetric, *American Psychologist*, vol 29, pp 512-518.

Carver RP 1978, The case against statistical significance testing, *Harvard Educational Review,* vol 48(3), pp 378-399.

Casagrande J 1954, The ends of translation, *International Journal of American Linguistics*, vol 20, pp 335-340.

Chambers LW 1988, 'The McMaster Health Index Questionnaire: an update', in Walker SR & Rosser RM (Eds), *Quality of Life: Assessment and Application*, pp 113-131. Lancaster: MTP Press.

Chambers LW, Haight M, Norman G & MacDonald L 1987, Sensitivity to change and the effect of mode of administration on health status measurement, *Medical Care*, vol 25, pp 470-480.

Chapman DW & Carter JF 1979, Translation procedures for the cross cultural use of measurement instruments, *Educational Evaluation and Policy Analysis*, vol 1, pp 71-76.

Charny MC, Lewis PA & Farrow SC 1989, Choosing who shall not be treated in the NHA, *Social Science & Medicine*, vol 28, pp 1331-1338.

Christensen-Szlanski JJJ 1984, Discount functions and the measurement of patients' values, *Medical Decision Making*, vol 4, pp 47-48.

Christensen-Szlanski JJJ & Northcraft GB 1985, patient compliance behavior: the effects of time on patients' values of treatment regimens, *Social Science and Medicine*, vol 21, pp 263-273.

Churchill DN, Wallace JE, Ludwin D, Beecroft ML & Taylor DW 1991, A comparison of evaluative indices of quality of life and cognitive function in hemodialysis patients, *Controlled Clinical Trials*, vol 12 (Supplement), pp 159S-167S.

Cole DA 1987, utility of confirmatory factor analysis in test validation research, *Journal of Consulting and Clinical Psychology*, vol 55, pp 584-594.

Comrey AL 1978, Common methodological problems in factor analytic studies, *Journal of Consulting and Clinical Psychology*, vol 46, pp 648-659.

Conrad P 1985, The meaning of medications: another look at compliance, *Social Science and Medicine*, vol 20, pp 29-37.

Cook DT & Campbell TD 1979, *Quasi-experimentation. Design & Analysis Issues for Field Settings*. Chicago: Rand McNally.

Cranshaw R 1990, Health care rationing [Letter], *Science*, vol 247, pp 662-663.

Cronbach LJ 1951, Coefficient alpha and the internal structure of tests, *Psychometrika*, vol 16, pp 297-334.

Cronbach LJ 1971, 'Test validation', in Thorndike RL (Ed), *Educational Measurement*, 2nd edition. Washington, DC: American Council on Education.

Cronbach LJ 1990, 'Construct validation after thirty years', in Linn RL (Ed), *Intelligence: Measurement Theory and Public Policy*, pp 147-171. Urbana, IL: University of Illinois Press.

Cronbach LJ, Gleser GC, Nanda H & Rajaratnam N 1972, *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. New York: Wiley.

Cronbach LJ & Meehl PE 1955, Construct validity in psychological tests, *Psychological Bulletin*, vol 52, pp 281-302.

Cubbon JE 1991, The principle of QALY maximisation as the basis for allocating health care resources, *Journal of Medical Ethics*, vol 17, pp 181-184.

Daniels N 1991, Is the Oregon rationing plan fair? *Journal of the American Medical Association*, vol 265, pp 2232-2235.

Davies AR & Ware JE 1981, *Measuring Health Perceptions in the Health Insurance Experiment* (R-2711-HHS). Santa Monica: The RAND Corporation.

De Groot AD 1986, 'An analysis of the concept of "quality of life",' in Ventafridda, V, van Dam FSAM, Yancik R & Tamburini M (Eds), *Assessment of Quality of Life and Cancer Treatment*. (Proceedings of International Workshop on Quality of Life Assessment and Cancer Treatment, Milan, 11-13 December, 1985). Amsterdam: Excerpta Medica.

Deniston OL, Carpentier-Alting P, Kneisley J, Hawthorn VM & Port FK 1989, Assessment of quality of life in end-stage renal disease, *Health Services Research*, vol 24, pp 555-578.

Derogatis LR 1986, The psychological adjustment to illness scale (PAIS), *Journal of Psychosomatic Research*, vol 30, pp 77-91.

Deyo RA 1984, Pitfalls in measuring the health status of Mexican Americans: comparative validity of the English and Spanish Sickness Impact Profile, *American Journal of Public Health*, vol 74, p 569.

Deyo RA & Centor RM 1986, Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance, *Journal of Chronic Diseases* vol 39, pp 897-906.

Deyo RA, Diehr P & Patrick DL 1991, Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation, *Controlled Clinical Trials*, vol 12 (Supplement), pp 142S-158S.

Deyo RA & Inui TS 1984, Toward clinical applications of health status measures: sensitivity of scales to clinically important changes, *Health Services Research*, vol 19, pp 275-289.

Deyo RA, Inui TS, Leininger J & Overman S 1982, Physical and psychosocial function in rheumatoid arthritis: clinical use of a self-administered health status instrument, *Archives of Internal Medicine*, vol 142, pp 879-.

Deyo RA & Patrick DL 1989, Barriers to the use of health status measures in clinical investigation, patient care, and policy research*, Medical Care*, vol 27 (Supplement), pp S233-S253.

Dixon J & Welch HG 1991, Priority setting: lessons from Oregon, *Lancet*, vol 337, pp 891-894.

Donabedian A, Elinson J, Spitzer W & Tarlov A 1987, Discussion: Advances in Health Assessment Conference, *Journal of Chronic Diseases*, vol 40 (Supplement), pp 183S-191S.

Donaldson C, Atkinson A, Bond J & Wright K 1988, Should QALYs be programme-specific? *Journal of Health Economics*, vol 7, pp 239-257.

Dowie J 1991, ‘Health outcomes and evaluation: a short and slightly impolite paper about health status and health service outcome measurement’. Paper presented to the ‘Forum on Priorities for National Health Statistics’, Canberra.

Eckberg DL 1979, *Intelligence and Race. The Origins and Dimensions of the IQ Controversy*. New York: Praeger Publishers.

Eddy DM 1991a, What’s going on in Oregon? *Journal of the American Medical Association*, vol 266, pp 417-420.

Eddy DM 1991b, Oregon’s methods: did cost-effectiveness fail? *Journal of the American Medical Association*, vol 266, pp 2135-2141.

Edwards AL 1976, *An introduction to Linear Regression and Correlation*. San Francisco: Freeman.

Einhorn HJ & Hogarth RM 1978, Confidence in judgment: persistence of the illusion of validity, *Psychological Review*, vol 85, pp 395-416.

Elinson J 1979, Introduction to the theme: sociomedical health indicators, *International Journal of Health Services*, vol 6, pp 385-391.

Evans JG 1990, Symposium proceedings: the ethics of resource allocation, *Journal of Epidemiology and Community Health*, vol 44, pp 187-190.

Evans WJ, Cayten CG & Green PA 1981, Determining the generalizability of rating scales in clinical settings, *Medical Care*, vol 19, pp 1211-1220.

Fabrega H 1975, The need for an ethnomedical science, *Science*, vol 189, pp 969-.

Feeny D & Torrance GW 1989, Incorporating utility-based quality-of-life assessment measures in clinical trials: two examples, *Medical Care*, vol 27 (Supplement), pp S190-S204.

Feeny D, Labelle R & Torrance GW 1990, ‘Integrating economic evaluations and quality of life assessments’, in Spilker B (Ed), *Quality of Life Assessments in Clinical Trials*, pp 71-83. New York: Raven Press.

Feinstein AR 1977, Clinical biostatistics XLI: Hard science, soft data, and the challenges of choosing clinical variables in research, *Clinical Pharmacology and Therapeutics*, vol 22, pp 485-498.

Feinstein AR 1987a, Clinimetric perspectives, *Journal of Chronic Diseases*, vol 40, pp 635-640.

Feinstein AR 1987b, *Clinimetrics*. New Haven: Yale University Press.

Feinstein AR, Josephy BR & Wells CK 1986, Scientific and clinical problems in indexes of functional disability, *Annals of Internal Medicine*, vol 105, pp 413-420.

Feinstein AR & Kramer MS 1980, Clinical biostatistics LII. A primer on quantitative indexes of association, *Clinical Pharmacology and Therapeutics*, vol 28, pp 130-145.

Felton BJ & Revenson TA 1984, Coping with chronic illness: a study of illness controllability and the influence of coping strategies on psychological adjustment, *Journal of Consulting and Clinical Psychology*, vol 52, pp 343-353.

Felton BJ, Revenson TA & Hinrichsen GA 1984, Stress and coping in the explanation of psychological adjustment among chronically ill adults, *Social Science and Medicine*, vol 18, pp 889-898.

Fineman H 1991, The social construction of noncompliance: a study of health care and social service providers in everyday practice, *Sociology of Health & Illness*, vol 13, pp 354-374.

Fischhoff B 1991, Value elicitation: is there anything in there? *American Psychologist*, vol 46, pp 835-847.

Fiske DW 1982, 'Convergent-discriminant validation in measurements and research strategies', in Brindberg D & Kidder LH (Eds), *Forms of Validity in Research*, pp 77-92. San Francisco: Jossey-Bass.

Folkman S, Lazarus R, Gruen RJ & DeLongis A 1986, Appraisal, coping, health status and psychological symptoms, *Journal of Personality and Social Psychology*, vol 50, pp 571-579.

Friedman DD 1986, 'Comments on "rationing and publicity"', in Agich GJ & Begley CE (Eds), *The Price of Health*, pp 217-224. Dordrecht, Holland: D Reidel.

Froberg D & Kane R 1989, Methodology for measuring health-state preferences – IV: Progress and a research agenda, *Journal of Clinical Epidemiology*, vol 42, pp 675-685.

Ganz PA, Lee JJ & Siau J 1991, Quality of life assessment. An independent prognostic variable for survival in lung cancer, *Cancer*, vol 67, pp 3131-3135.

Ghiselli EE, Campbell JP & Zedeck S 1981, *Measurement Theory in the Behavioral Sciences*. San Francisco: WD Freeman.

Gilson BS, Erickson D, Chavez CT, Bobbitt RA, Bergner M & Carter WB 1980, A Chicano version of the Sickness Impact Profile (SIP), *Culture, Medicine and Psychiatry*, vol 4, pp 137-150.

Goldsmith SB 1972, The status of health status indicators, *Health Services Reports*, vol 87, pp 212-220.

Gould SJ 1981, *The Mismeasurement of Man*. New York: WW Norton.

Green LW & Lewis FM 1986, *Measurement and Evaluation in Health Education and Health Promotion*. Palo Alto, CA: Mayfield.

Greenwald HP 1987, The specificity of quality-of-life measures among the seriously ill, *Medical Care*, vol 25, pp 642-651.

Guildford JP 1946, New standards for test evaluation, *Educational & Psychological Measurement*, vol 6, pp 427-438.

Guion RM 1980, On Trinitarian doctrines of validity, *Professional Psychology*, vol 11, pp 385-398.

Guyatt GH, Bombardier C & Tugwell PX 1986, Measuring disease-specific quality of life in clinical trials, *Canadian Medical Association Journal*, vol 134, pp 889-895.

Guyatt GH & Jaeschke R 1990, 'Measurements in clinical trials: choosing the appropriate approach', in Spilker BF (Ed), *Quality of Life Assessments in Clinical Trials*, pp 37-46. New York: Raven Press.

Guyatt GH, Mitchell A, Irvine EJ, Singer J, Williams N, Goodcare R & Tompkins C 1989b, A new measure of health status for clinical trials in inflammatory bowel disease, *Gastroenterology*, vol 96, pp 804-810.

Guyatt GH, Veldhuyzen van Zanten SJO, Feeny DH & Patrick DL 1989c, Measuring quality of life in clinical trials: a taxonomy and review, *Canadian Medical Association Journal*, vol 140, pp 1441-1448.

Guyatt GH, Walter S & Norman G 1987, Measuring change over time: assessing the usefulness of evaluative instruments, *Journal of Chronic Diseases*, vol 40, pp 171-178.

Hadorn DC 1991, The Oregon priority-setting exercise: quality of life and public policy, *Hastings Center Report*, (Supplement), pp 11-16.

Hadorn DC & Hays RD 1991, Multitrait-multimethod analysis of health-related quality-of-life measures, *Medical Care*, vol 29, pp 829-840.

Harris J 1991, Unprincipled QALYs: a response to Cubborn, *Journal of Medical Ethics*, vol 17, pp 185-188.

Hart LG & Evans RW 1987, The functional status of ESRD patients as measured by the Sickness Impact Profile, *Journal of Chronic Diseases*, vol 40 (Supplement), pp 117S-130S.

Hays RD & Stewart AL 1990, The structure of self-reported health in chronic conditions, *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, vol 2, pp 22-30.

Hendricson WD, Russell IJ, Prihoda TJ, Jacobson JM, Rogan A & Bishop GD 1989, An approach to developing a valid Spanish language translation of a health-status questionnaire, *Medical Care*, vol 27, pp 959-966.

Hogarth RM 1980, *Judgment and Choice: the Psychology of Decision*, 1st edition. Chichester: Wiley.

Hollandsworth JG 1988, Evaluating the impact of medical treatment on the quality of life: a 5-year update, *Social Science and Medicine*, vol 26, pp 425-434.

Hollenberg NK, Testa M & Williams GH 1991, Quality of life as a therapeutic end-point. An analysis of therapeutic trials of hypertension, *Drug Safety*, vol 6, pp 83-93.

Hui CH & Triandis HC 1985, Measurement in cross-cultural psychology: a review and comparison of strategies, *Journal of Cross-Cultural Psychology*, vol 16, pp 131-152.

Hui CH & Triandis HC 1989, Effects of culture and response format on extreme response style, *Journal of Cross-Cultural Psychology*, vol 20, pp 296-309.

Hulin CL 1987, A psychometric theory of evaluations of item and scale translations: fidelity across languages, *Journal of Cross-Cultural Psychology*, vol 18, pp 115-142.

Hunt SM 1986, Cross-cultural issues in the use of socio-medical indicators, *Health Policy*, vol 6, pp 149-158.

Hunt SM 1988, ‘Measuring health in clinical care and clinical trials’, in Teeling-Smith G (Ed), *Measuring Health: A Practical Approach*, pp 7-21. Chichester: Wiley.

Hunt S, McEwen J & McKenna SP 1986, *Measuring Health Status*. London: Croom Helm.

Hunt S & Wiklund I 1987, Cross-cultural variation in the weighting of health statements: a comparison of English and Swedish valuations, *Health Policy*, vol 8, pp 227-235.

Isaac S & Michael WB 1981, *Handbook in Research and Evaluation*, 2nd edition. San Diego: EdITS.

Jackson KG 1969, Multimethod factor analysis in the evaluation of convergent and discriminant validity, *Psychological Bulletin*, vol 72, pp 30-49.

Jaeschke R & Guyatt G 1990, ‘How to develop and validate a new quality of life measure’, in Spilker BF (Ed), *Quality of Life Assessments in Clinical Trials*, pp 47-57. New York: Raven Press.

Jenkins CD, Jono RT, Stanton B-A & Stroup-Benham CA 1990, The measurement of health-related quality of life: major dimensions identified by factor analysis, *Social Science and Medicine*, vol 31, pp 925-931.

Jette AM 1980, Health status indicators: their utility in chronic disease evaluation research, *Journal of Chronic Diseases*, vol 61, pp 85-89.

Joreskog KG & Sorbom D 1984, *LISREL VI: Analysis of Linear Structural Relationships by Maximum Likelihood, Instrumental Variables, and Least Squares Methods*, 3rd edition. Moresvill IN: Scientific Software Inc.

Kaplan RM, Bush JW & Berry CC 1976, Health status: types of validity for an index of well-being, *Health Services Research*, vol 11, pp 478-507.

Kaplan RM 1985, 'Social support and social health. Is it time to rethink the WHO definition of health?' in Sarason IG & Sarason BR (Eds), *Social Support: Theory, Research and Applications*. Dordrecht: Martinus Nijhoff.

Kaplan RM & Anderson JP 1988, 'The Quality of Well-Being Scale: rationale for a single quality of life index', in Walker SR & Rosser RM (Eds), *Quality of Life: Assessment and Application*, pp 51-77. Lancaster: MTP Press.

Kaplan RM & Anderson JP 1990, 'The general health policy model: an integrated approach', in Spilker B (Ed), *Quality of Life Assessments in Clinical Trials*, pp 131-149. New York: Raven Press.

Kaplan RM, Anderson JP, Wu AW, Mathews WC, Kozin F & Orenstein D 1989, The Quality of Well-Being Scale: applications in AIDS, cystic fibrosis, and arthritis, *Medical Care*, vol 27 (Supplement), pp S27-S43.

Kaplan RM & Bush JW 1982, Health-related quality of life measurement for evaluation research and policy analysis, *Health Psychology*, vol 1, pp 61-80.

Kaplan SH 1987, patient reports of health status as predictors of physiologic health measures, *Journal of Chronic Diseases*, vol 40 (Supplement 1), pp 27S-35S.

Kidder LH & Judd CM 1986, *Research Methods in Social Relations*, 5th edition. New York: CBS College Publishing.

Kind P, Rosser RM & Williams AR 1982, 'Valuation of quality of life: some psychometric evidence', in Jones-Lee MW (Ed), *The Value of Life and Safety*. Amsterdam: North Holland.

Kirschner B & Guyatt G 1985, A methodological framework for assessing health indices, *Journal of Chronic Diseases*, vol 38, pp 27-36.

Kleinman A, Eisenberg L & Good B 1978, Culture, illness and care: clinical lessons from anthropologic and cross-cultural research, *Annals of Internal Medicine*, vol 88, pp 251-258.

Kramer MS & Feinstein AR 1981, Clinical biostatistics LIV: the biostatistics of concordance, *Clinical Pharmacology and Therapeutics*, vol 29, pp 111-123.

Krause MS 1972, The implications of convergent and discriminant validity data for instrument validation, *Psychometrika*, vol 37, pp 179-186.

Kuder GF & Richardsojn MW 1937, The theory of the estimation of test reliability, *Psychometrika*, vol 2, pp 151-160.

Labuhn KT 1984, ‘An analysis of self-reported depressed mood in chronic obstructive pulmonary disease’, Doctoral Dissertation, University of Michigan.

Landy FJ 1986, Stamp collecting versus science, *American Psychologist*, vol 41, pp 1183-1192.

Lewis PA & Charny M 1989, Which of two individuals do you treat when only their ages are different and you can’t treat both? *Journal of Medical Ethics*, vol 15, pp 28-32.

Llewellyn-Thomas H, Sutherland HJ, Tibshirani R, Ciampi A, Till JE & Boyd NF 1982, The measurement of patients’ values in medicine, *Medical Decision Making*, vol 2, pp 449-462.

Lomas J, Pickard L & Mohide A 1987, Patient versus clinician item generation for quality-of-life measures: the case of language-disabled adults, *Medical Care*, vol 25, pp 764-769.

Long JS 1983, *Confirmatory Factor Analysis*. Beverly Hills: Sage.

Lykken DT 1968, Statistical significance in psychological research, *Psychological Bulletin*, vol 70, pp 151-159.

MacKenzie CR, Charlson ME, DiGioia D & Lelley K 1986a, A patient-specific measure of change in maximal function, *Archives of Internal Medicine*, vol 146, pp 1325-1329.

MacKenzie CR, Charlson ME, DiGioia D & Lelley K 1986b, Can the Sickness Impact Profile measure change? An example of scale assessment, *Journal of Chronic Diseases*, vol 39, pp 429-438.

Maynard A 1987a, Logic in medicine: an economic perspective, *British Medical Journal*, vol 295, pp 1537-1541.

Maynard A 1987b, The inevitability of outcome management: how should QALYs be used? *Journal of Management in Medicine*, vol 2, pp 107-114.

Maynard A 1990, Symposium proceedings: the ethics of resource allocation, *Journal of Epidemiology and Community Health*, vol 44, pp 187-190.

McClellan WM, Anson C, Birkeli K & Tuttle E 1991, Functional status and quality of life: predictors of early mortality among patients entering treatment for end stage renal disease, *Journal of Clinical Epidemiology*, vol 44, pp 83-89.

McDowell I & Newell C 1987, *Measuring Health: a Guide to Rating Scales and Questionnaires*. New York: Oxford University Press.

McEwen J 1988, ‘The Nottingham Health Profile’, in Walker SR & Rosser RM (Eds), *Quality of Life: Assessment and Application*, pp 95-111. Lancaster: MTP Press.

McFarlance A, Norman G, Streiner D, Roy R & Scott D 1980, Longitudinal study of the influence of the psychosocial environment on health states: a preliminary report, *Journal of Health and Social Behavior*¸ vol 21, pp 124-133.

McGrath JE 1982, ‘Dilemmatics. The study of research choices and dilemmas’, in McGrath JE, Martin J & Kulka RA (Eds), *Judgment Calls in Research*. Berverly Hills: Sage.

McGuire WJ 1983, ‘A contextualist theory of knowledge’, in Berkowitz L (Ed), *Advances in Experimental Social Psychology*. Orlando FL: Academic Press.

McKinlay JB 1981, From “promising report” to “standard procedure”: seven stages in the career of a medical innovation, *Milbank Memorial Fund Quarterly*, vol 59, pp 374-411.

McSweeny AJ & Labuhn KT 1990, ‘Chronic obstructive pulmonary disease’, in Spilker BF (Ed), *Quality of Life Assessments in Clinical Trials*. New York: Raven Press.

Marsh HW 1989, Confirmatory factor analyses of multitrait-multimethod data: many problems and a few solutions, *Applied Psychological Measurement*, vol 13, pp 335-361.

Meehl PE 1986, 'What social scientists don't understand', in Fiske DW & Schweder RA (Eds), *Metatheory in Social Science: Pluralism and Subjectivities*, pp 315-338. Chicago: Chicago University Press.

Meehl PE 1978, Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology, *Journal of Consulting and Clinical Psychology*, vol 46, pp 806-834.

Mehrez A & Gafni A 1989a, 'Healthy years equivalent: how to measure them using the standard gamble approach'. Working paper no 22. Hamilton: Centre for Health Economics and Policy Analysis, McMaster University.

Mahrez A & Gafni A 1989b, Quality-adjusted life years, utility theory, and health-years equivalents, *Medical Decision Making*, vol 9, pp 142-149.

Messick S 1975, Meaning and values in measurement and evaluation, *American Psychologist*, vol 30, pp 955-966.

Messick S 1980, Test validity and the ethics of assessment, *American Psychologist*, vol 35, pp 1012-1027.

Messick S 1981, Evidence and ethics in the evaluation of tests, *Educational Researcher*, vol 10, pp 9-20.

Messick S 1988, 'The once and future issues of validity: assessing the meaning and consequences of measurement', in Wainer H & Braun H (Eds), *Test Validity*, pp 33-45. Hillsdale NJ: Lawrence Erlbaum.

Messick S 1989, 'Validity', in Linn RL (Ed), *Educational Measurement*, 3rd edition, pp 13-103. New York: MacMillan/American Council on Education.

Miettinen OS 1987, Quality of life from the epidemiologic perspective, *Journal of Chronic Diseases*, vol 40, pp 641-643.

Morrow GP, Chiarello RJ & Derogatis L 1978, A new scale of assessing patients' psychosocial adjustment to medical illness, *Psychological Medicine*, vol 9, pp 605-610.

Mosteller F 1989, Final panel: comments on the Conference on Advances in Health Status Assessment, *Medical Care*, vol 20, pp 117-139.

Moum T 1988, Yea-saying and mood-of-the-day effects in self-reported quality of life, *Social Indicators Research*, vol 20, pp 117-139.

Mulkay M, Ashmore M & Pinch T 1987, Measuring the quality of life: a sociological invention concerning the application of economics to health care, *Sociology*¸ vol 21, pp 541-564.

Newcomb MD & Bentler PM 1987, Self-report methods of assessing health status and health service utilization: a hierarchical confirmatory analysis, *Multivariate Behavioral Research*, vol 22, pp 415-436.

Nord E (in press) The validity of a visual analogue scale in determining social utility weights for health states, *International Journal of Health Planning and Management*.

Nord E 1991, 'Methods for establishing quality weights for life years', Working paper no 8, Melbourne: NHMRC National Centre for Health Program Evaluation.

Norman GR 1989, Issues in the use of change scores in randomised trials, *Journal of Clinical Epidemiology*, vol 42, pp 1097-1105.

Nunally JC 1975, 'The study of change in evaluation research: principles concerning measurement, experimental design, and analysis', in Struening EL & Guttentag M (Eds), *Handbook of Evaluation Research, Vol 1*, pp 101-138. Beverly Hills: Sage.

Nunally JC 1978, *Psychometric Theory*, 2nd edition. New York: McGraw-Hill.

Paterson M 1988, 'Assessment of treatment in rheumatoid arthritis', in Teeling-Smith G (Ed), *Measuring Health: A Practical Approach*, pp 157-189. Chichester: Wiley.

Patrick DL 1987, Commentary: patient reports of health status as predictors of physiologic health measures, *Journal of Chronic Diseases*, vol 40 (Supplement 1), pp 37S-40S.

Patrick DL 1981, 'Standardization of comparative health status measures: using scales developed in America in an English-speaking country', in *Survey Research Methods: Third Biennial Conference*. Hyattsville MD: US Dept of Health and Human Services, pub No (PHS) 81-3268, pp 216-220.

Patrick DL & Bergner M 1990, Measurement of health status in the 1990s, *Annual Review of Public Health*, vol 11, pp 165-183.

Patrick DL & Deyo RA 1989, Generic and disease-specific measures in assessing health status and quality of life, *Medical Care*, vol 27 (Supplement), pp S217-S232.

Patrick DL & Erickson P 1988a, 'Assessing health-related quality of life for clinical decision making', in Walker SR & Rosser RM (Eds), *Quality of Life: Assessment and Application*, pp 9-49. Lancaster: MTP Press.

Patrick DL & Erickson P 1988b, What constitutes quality of life? Concepts and dimensions, *Quality of Life and Cardiovascular Disease*, vol 5, pp 103-127.

Patrick DL, Sittampalam Y, Somerville SM, Carter WB & Bergner M 1985, A cross-cultural comparison of health status values, *American Journal of Public Health*, vol 75, pp 1402-1407.

Peak H 1953, 'Problems of observation', in Festinger L & Katz D (Eds), *Research Methods in the Behavioral Sciences*, pp 243-299. Hinsdale IL: Dryden Press.

Priestman T 1986, Measuring quality of life during cancer therapy, *Update*, vol 1, pp 987-998.

Read JL, Quinn RJ, Berwick DM, Fineberg HV & Weinstein MC 1984, Preferences for health outcomes: comparison of assessment methods, *Medical Decision Making*, vol 4, pp 315-329.

Read JL, Quinn RJ & Hoefer MA 1987, Measuring overall health: an evaluation of three important approaches, *Journal of Chronic Diseases*, vol 40 (Supplement), pp 7S-22S.

Rhoads S 1980, 'How much should we spend to save a life?' in *Valuing Life: Public Policy Dilemmas*. Boulder CO: Westview Press.

Roberts J, Browne G, Brown B, Byrne C & Love B 1987a, Coping revisited: the relation between appraised seriousness of an event, coping responses and adjustment to illness, *Nursing Papers*, vol 19, pp 45-54.

Roberts J, Browne G, Streiner D, Byrne C, Brown B & Love B 1987b, Analyses of coping responses and adjustment: stability of conclusions, *Nursing Research*, vol 36, pp 94-97.

Rockey PH & Griep RJ 1980, Behavioural dysfunction in hyperthyroidism: improvement with treatment, *Archives of Internal Medicine*, vol 140, pp 1194-1197.

Rosser RM 1987a, ‘A health index and output measure’, in Walker SR & Rosser RM (Eds), *Quality of Life: Assessment and Application*, pp 133-160. Lancaster: MTP Press.

Rosser RM 1987b, ‘Quality of life: consensus, controversy and concern’, in Walker SR & Rosser RM (Eds), *Quality of Life: Assessment and Application*, pp 297-304. Lancaster: MTP Press.

Rosser RM 1990, ‘From health indicators to quality adjusted life years: technical and ethical issues’, in Hopkins A & Costain D (Eds), *Measuring the Outcomes of Medical Care*, pp 1-17. London: Royal College of Physicians of London, King’s Fund for Health Services Development.

Rosser RM & Watts VC 1972, The measurement of hospital output, *International Journal of Epidemiology*, vol 1, pp 361-368.

Rummell RJ 1970, ‘Dimensions of error in cross-national data’, in Narrol R & Cohen R (Eds), *A Handbook of Method in Cultural Anthropology*. New York: American Museum of Natural History.

Runkel PJ & McGrath JE 1972, *Research on Human Behavior: a Systematic Guide to Method.* New York: Holt, Rinehart & Winston.

Schmitt N & Stults DM 1986, Methodology review: analysis of multitrait-multimethod matrices, *Applied Psychological Measurement*, vol 10, pp 1-22.

Schriesheim CA 1981, Leniency effects on convergent and discriminant validity for grouped questionnaire items: a further investigation, *Educational and Psychological Measurement*, vol 41, pp 1093-1099.

Sechrest L, Fay TL & Zaidi SMH 1972, Problems of translation in cross-cultural research, *Journal of Cross-Cultural Psychology*, vol 3, pp 41-56.

Seers D 1975, ‘The political economy of national accounting’, in Cameron A & Puri M (Eds), *Employment, Income Distribution and Development Strategy*. London: Macmillan.

Segovia J, Bartlett RF & Edwards AC 1989, An empirical analysis of the dimensions of health status measures, *Social Science and Medicine*, vol 29, pp 761-768.

Sherbourne CD & Stewart AL 1991, The MOS social support survey, *Social Science & Medicine*, vol 32, pp 705-714.

Shortell SM & Richardson WC 1978, *Health Program Evaluation*. St Louis: CV Mosby.

Spitzer WO 1987a, Discussion: Advances in Health Assessment Conference, *Journal of Chronic Diseases,* vol 40 (Supplement 1), pp 187-189.

Spitzer WO 1987b, State of science 1986: quality of life and functional status as target variables for research, *Journal of Chronic Diseases*, vol 40, pp 465-471.

Stanley JC 1961, Analysis of unreplicated three-way classifications, with applications to rater bias and trait independence, *Psychometrika*, vol 26, pp 205-219.

Stanley JC 1971, ‘Reliability’, in Thorndike RD (Ed), *Educational Measurement*, pp 356-442. Washington DC: American Council on Education.

Stevens SS 1971, Issues in psychophysical measurement, *Psychological Review*, vol 78, pp 426-450.

Streiner DL & Norman GR 1989, *Health Measurement Scales: a Guide to their Development and Use*. Oxford: Oxford University Press.

Torrance GW 1976, Social preferences for health states: an empirical evaluation of three measurement techniques, *Socio-Economic Planning Sciences*, vol 10, pp 129-136.

Torrance GW 1986, Measurement of health state utilities for economic appraisal: a review article, *Journal of Health Economics*, vol 5, pp 1-30.

Torrance GW 1987, Utility approach to measuring health-related quality of life, *Journal of Chronic Diseases*, vol 40, pp 593-600.

Tucker LR 1966, Some mathematical notes on three-mode factor analysis, *Psychometrika*, vol 31, pp 279-311.

Viney LL & Westbrook MT 1984, Coping with chronic illness: strategy preferences, changes in preferences and associated emotional reactions, *Journal of Chronic Diseases*, vol 36, pp 489-502.

Ware JE 1984a, Conceptualizing disease impact and treatment outcomes, *Cancer*, vol 53 (Supplement), pp 2316-2323.

Ware JE 1984b, 'General Health Rating Index', in Wenger NK, Mattson ME, Furberg CE & Elinson J (Eds), *Quality of Life in Clinical Trials of Cardiovascular Therapies*, pp 184-188. New York: Le Jacq.

Ware JE 1984c, 'Methodological considerations in the selection of health status assessment procedures', in Wenger NK, Mattson ME, Furberg CE & Elinson J (Eds), *Quality of Life in Clinical Trials of Cardiovascular Therapies*, pp 87-111. New York: Le Jacq.

Ware JE 1986, 'The assessment of health status', in Aiken LH & Mechanic D (Eds), *Applications of Social Science to Clinical Medicine and Health Policy*. New Brunswick: Rutgers University Press.

Ware JE 1989, Final panel: comments on the Conference on Advances in Health Status Assessment', *Medical Care*, vol 27 (Supplement), pp S286-S290.

Ware JE, Davies-Avery A & Brook RH 1980, *Conceptualization and Measurement of Health for Adults in the Health Insurance Study: Vol VI. Analysis of Relationships Among Health Status Measures* (R-1987/6-HEW). Santa Monica: the RAND Corporation.

Ware JE, Brook RH, Davies-Avery A, Williams KA, Stewart AL, Rogers WH, Donald CA & Johnston SA 1980, *Conceptualization and Measurement of Health for Adults in the Health Insurance Study: Vol I. Model of Health and Methodology* (R-1987/1-HEW). Santa Monica: The RAND Corporation.

West PA 1985, 'What is wrong with QALYs?' Paper presented to the Health Economists' Study Group, December, 1985.

Whitelaw NA & Liang J 1991, The structure of the OARS physical measures, *Medical Care*, vol 29, pp 332-347.

Wiklund I & Karlberg J 1991, Evaluation of quality of life in clinical trials: selecting quality-of-life measures, *Controlled Clinical Trials*, vol 12 (Supplement), pp 204S-216S.

Wiklund I, Romanus B & Hunt SM 1988, Self-assessed disability in patients with arthrosis of the hip joint: reliability of the Swedish version of the Nottingham Health Profile, *International Disability Studies*, vol 10, pp 159-163.

Williams AR 1985, Economics of coronary artery bypass grafting, *British Medical Journal*, vol 291, pp 325-329.

Winslow GR 1986, ‘Rationing and publicity’, in Agich GJ & Begley CE (Eds), *The Price of Health*, pp 199-216. Dordrecht, Holland: D Reidel.

World Health Organization 1958, *The First Ten Years of the World Health Organization*. Geneva: World Health Organization.

Wright SJ 1986, ‘Age, sex and health: a summary of findings from the York health evaluation survey’. Discussion paper no 15. Centre for Health Economics, University of York.

Zeller RA & Carmines EG 1980, *Measurement in the Social Sciences*. Cambridge: Cambridge University Press.