CENTRE FOR HEALTH PROGRAM EVALUATION

WORKING PAPER 30

ISSUES IN THE MEASUREMENT OF HEALTH-RELATED QUALITY OF LIFE

Rod O'Connor
Senior Research Associate

July, 1993
ISBN 1038-9547
ISSN 1 875677 26 7

CENTRE PROFILE

The Centre for Health Program Evaluation (CHPE) is a research and teaching organisation established in 1990 to:

- undertake academic and applied research into health programs, health systems and current policy issues;
- develop appropriate evaluation methodologies; and
- promote the teaching of health economics and health program evaluation, in order to increase the supply of trained specialists and to improve the level of understanding in the health community.

The Centre comprises two independent research units, the Health Economics Unit (HEU) which is part of the Faculty of Business and Economics at Monash University, and the Program Evaluation Unit (PEU) which is part of the Department of General Practice and Public Health at The University of Melbourne. The two units undertake their own individual work programs as well as collaborative research and teaching activities.

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 National Health and Medical Research Council and Monash University.

The Program Evaluation Unit of the CHPE is supported by The University of Melbourne.

Both units obtain supplementary funding through national competitive grants and contract research.

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

TABLE OF CONTENTS

Contents (Cont'd)

Contents (Cont'd)

Issues in the Measurement of Health-Related Quality of Life

A Introduction

This paper reviews the literature and discusses the major issues regarding the development of a reliable, valid and practicable instrument that could comprehensively measure a patient's quality of life.

Issues of definition are first considered, with special attention to the notion that quality of life could be defined as a patient's subjective well-being. The complexities of measuring such a psychological construct are noted, followed by consideration of the major issues that are entailed. These concern the problem of interpreting such measurements as interval or ratio data, and the means of development and assessing test reliability and validity. Next issues of test construction and interpretation of specific importance to health status measures are investigated. Finally there is an examination of the conceptualisation, development and psychometric status of four of the major instruments proposed to provide health status measures (those developed by Bergner; Bush, Kaplan et al.; Torrance; and Rosser).

The aim of the review is to provide basic information and analysis that is central to anyone considering the development of a health status assessment instrument from a psychometric perspective, or who wishes to assess existing instruments. It should not be concluded that the review is definitive: there are a number of issues that are only hinted at in this review (eg. the need for a systematic comparison of rating versus trade-off measures of health state value). However it is hoped that the information, analyses and thoughts presented may be of assistance to those working in what is a very complex and increasingly important inter-disciplinary field.

To aid assimilation of the information, and to permit a rapid overview, the concluding sections of each of the first four chapters may be read (ie. the final summary sections of `Concepts and Definitions'; `Scaling'; `Reliability'; and `Validity'). The final two chapters, `Issues when constructing health-related quality of life measures', and `Some current tools', are best read in their entirety.

B CONCEPTS AND DEFINITIONS IN HEALTH-RELATED QUALITY OF LIFE

1 The Motivation for Developing Quality of Life (QOL) Measures in Health

The development of Quality of Life measures in health has been encouraged both by the need to assess the relative merits of rival health programs in a context of increasing pressure on health resources, and a desire to be able to comprehensively assess the impact of clinical therapies.

(a) Resource allocation

The need to have a measure of health program effect that goes beyond traditional output measures (such as number of patients treated) has been a major motivation in the development of generic health status assessment instruments. There has been an increasing need to rationally allocate health service resources across diverse health programs. Arising from this a number of measures of the broad effects of illness state on a patient's life have been developed with the declared aim of assisting health policy decisions. The Rosser Index has been seen to allow evaluation of health service funding. Gudex (1986) and Torrance (1972) described the development of his health utility approach as a means of measuring health improvement that is disease and program independent, the aim being to facilitate decisions regarding program funding. Similarly Kaplan (1988a) has applied the QWB to develop a General Health Policy Model.

Even the Sickness Impact Profile (or SIP) developed by Bergner and co-workers (eg. see Bergner et al. 1976a, 1976b), seemingly developed outside the context of cost-utility theory, had this as a declared aim. Bergner stated that the SIP was developed with the specific aim of providing information on the efficacy of health programs to assist decisions regarding the appropriate allocation of the government's resources. It was aimed to provide a `fiscally and logistically practical measure of health status' (Bergner et al. 1976a, p. 393).

(b) Assessing clinical outcomes

Revicki (1989) notes that advances in medical research and therapy have shifted health care resources from the diagnosis and treatment of infectious disease to the prevention and control of chronic disease: with this has come an increased emphasis on changes in functional status and quality of life outcomes. This move is assisted by the undesirable aspects of many modern treatments. Bergner (1989) notes that the consequences of treatment and treatment-related side effects may affect all of a patient's life. eg. becoming bald and nauseous, being on a restricted diet, being tied to a machine 12 hours out of 24, etc. and hence it is important to assess all aspects of a treatment's effects. In this context Revicki (1989) cited a 1986 study by Croog et al. which compared three anti-hypertensive drugs, and selected measures to indicate QOL dimensions of general well-being, sleep dysfunction, sexual problems, work performance, social activity participation, physical

distress, and cognitive function. While the drugs were found to have comparable efficacy in decreasing blood pressure, and there were no differences between the treatment groups on measures of sleep dysfunction, social participation, and visual memory, one drug did show differences on general well being and physical distress.

Deyo and Patrick (1989) have also pointed out that medical interventions may result in improved functional health status without evidence of physiologic improvement eg. pulmonary rehabilitation programs may improve exercise capacity without altering pulmonary function tests, while on the other hand therapy may result in physiologic improvement without discernible clinical benefit to patients eg. nitrate therapy may alter haemodynamics in patient with heart disease without improving exercise capacity.

Cancer is an area which has been noted as particularly relevant to QOL considerations. Donovan et al. (1989) note evidence that the emotional suffering produced by cancer exceeds the physical suffering it causes, while at the same time pointing out that QOL measures have generally not been included in clinical trials of cancer therapy. This is at least in part due to physicians not being comfortable working with social scientists, whose tools have been seen as `soft' and cumbersome (Skeel 1989), and lacking credibility (Deyo & Patrick 1989).

2 Types of Health-related QOL Measure

A large number of measures of health status and associated notions are available (eg. see McDowell & Newell, 1987). Most can be characterised in terms of three continuums: disease specific versus generic measures; single dimension versus broad spectrum measures; and the range of values output (see also Bergner 1989; Donovan et al. 1989).

(a) Disease specific versus generic measures

Measures vary in the degree to which they are developed to measure a specific disease or to be capable of application to many or all illness states. As noted by Deyo and Patrick (1989), disease specific measures have greater salience for physicians, better focus on functional areas of particular concern, and may possess greater responsiveness to disease-specific interventions. On the other hand generic measures permit comparisons across interventions and diagnostic conditions, which is particularly important for policy makers [resource allocation]. They also allow dysfunction to be quantified for an individual experiencing several disease conditions (Bombardier et al. 1986; Temkin et al. 1989).

There is some evidence that generic measures can be as responsive in some settings as disease-specific measures. Kaplan et al. (1989) criticise disease-specific methods on the grounds that all diseases and disabilities affect overall quality of life, and the purpose of QOL measures is not to identify clinical information relevant to the disease but to determine the impact of the disease on general function. General QOL measures are proposed as better as they can capture a wide variety of dysfunction that might be in different systems,

ie. not specific to the disease condition (eg. confusion, tiredness, sexual impotence, depression).

Certainly general measures have the ability to capture side effects and benefits that might not have been anticipated (Kaplan & Anderson 1988), although once identified a disease-specific measure could be prepared to more exactly assess the dimension of interest.

(b) Single dimension versus broad spectrum measures

Measures vary according to the degree to which they focus on particular activities, or attempt to encompass the full range of aspects of living that may influence personal contentment or satisfaction for a person with the condition.

As noted by Bergner (1989), there are measures that focus on particular activities such as walking, eating and dressing (eg. Activities of Daily Living Index), while others measure physical functioning plus other health related aspects such as symptoms, emotional status, cognition, perceptions of health etc. The latter group consists of both measures specific to a disease and general measures (such as Bergner's Sickness Impact Profile).

(c) Range of values output

Measures also vary in that they may output a single value, a series of sub-scale values, or a series of sub-scale values plus an aggregate value. Examples of scales that both measure multiple dimensions and provide an aggregate measure are the QWB and SIP.

3 The Notion of QOL as Used in Health

While there is general agreement on the potential value of QOL measures as key evaluation variables, there is an absence of clear agreement on a definition of QOL: definitions of QOL in the health context are mostly vague or absent. As noted by Deyo and Patrick (1989), conceptions relevant to health and QOL are diverse, scattered through many disciplines, and use many different labels (eg. health status, functional status, disability scale, quality of life). Bergner (1989) notes that the notion of Quality of Life (QOL) has been a category in Index Medicus since 1966, yet QOL is usually not defined in the reports of clinical trials, and `definitions must be deduced from the dimensions assessed', and that `each investigator that purports to address quality of life actually examines a very narrow and specific set of factors'. Generally notions of quality of life are not specified, but are considered to be implicit in the measure used, ie. they are more inferred than explained.

None the less there seems to be acceptance that health-related QOL is a `multidimensional concept that encompasses the physical, emotional, and social components associated with an illness or treatment' (Revicki 1989). Which precise dimensions to include is less agreed.

For example, Torrance (1987) states that physiological and emotional functioning contribute directly to quality of life, and `taken together these two constitute health-related quality of life' (Torrance 1987, p. 593), with social functioning (eg. social role and social contacts) outside the scope of health-related quality of life. On the other hand Kaplan et al. (1989) use the term health-related quality of life to refer to the impact of health conditions on function, but include social role, although suggesting that health-related quality of life may be independent of quality of life relevant to work setting, housing, or similar factors.

The sampling of the proper dimensions when estimating QOL is central to the validity of QOL measures, and is considered further in Chapter F of this report.

4 QOL as an Individual's Subjective Well-being

To develop a clear conceptual base it is useful to attempt to clarify a notion of `quality of life'. The term `quality of life' (QOL) can have several meanings. It may be used to refer to outward material circumstances, such that good quality of life is represented by good physical health, material security, supportive family and friends, etc. Alternatively it can refer to subjective well-being, or SWB, by this being meant an individual's sense of happiness or satisfaction, typically reflecting a global assessment of all aspects of their life. (McCauley and Bremer (1991) make a similar distinction between outward circumstances and personal assessment in their proposal of `objective well-being' versus `subjective well-being').

Both emotional and cognitive factors may be referred to as part of subjective well-being, while objective conditions such as health, wealth, and comfort are seen to be potential influences but not inherently or necessarily part of the notion. As noted by Diener (1984), the literature on SWB broadly concerns notions such as happiness, morale, positive affect, etc. and covers both positive judgement and affective reactions; it has been concerned either with what leads people to evaluate their lives in positive terms (a global judgement regarding life satisfaction), or happiness in terms of a preponderance of positive affect over negative affect.

The work of Campbell is often referred to when interpreting quality of life as subjective well-being. For example Donovan et al. (1989) cited Campbell (1976) in suggesting that an accepted general definition of quality of life is `a persons subjective sense of well-being, derived from current experience of life as a whole'. In the context of treatment selection, Goodinson and Singleton (1989) propose a definition of quality of life as `the degree of satisfaction with perceived present life circumstances' (citing Young & Longman 1983), this being seen to encompass the `physical, social, and material well-being of an individual', and to concern an evaluation of the physical, psychological and social impact of disease treatment on patients lives.

Within this framework, QOL is seen to be influenced by quite idiosyncratic factors, with a major determinant of an individual's quality of life being the perceived discrepancy between

what is and what could have been. Skeel (1989), considering quality of life from the context of cancer research, cites Calman (1984) in that quality of life `is the extent to which a persons hopes and ambitions are matched and fulfilled by experience'. In further support of this interpretation, Campbell (1981) reported that when questioned about the quality of their lives, apparently healthy individuals respond in terms of life satisfaction, usually in relation to specific domains, where satisfaction is proportional to the closeness between aspiration and achievement. Bergner (1989) also reports the notion that QOL is enhanced as the distance between attained and desired goals diminishes.

The implication is that changing expectation can lead to altered perception of QOL in similar circumstances, and different experiences may have different quality of life implications for different individuals. The notion may also help to explain why some people appear to adapt to changed circumstances very rapidly, ie. by reducing their aspirations (see B 6(c)).

5 Objections to QOL as Subjective Well-being

There is no doubt that objective external factors such as income, length of survival, change in tumour volume, etc. influence quality of life. Generally such factors are assessed to be influences on QOL, not the QOL itself, however there are those who appear to argue that quality of life should be identified with physical conditions only. In the context of health status measurement, Kaplan et al. (1989) stated that `most investigators believe that symptoms and mortality do represent quality of life' (Kaplan et al. 1989, p. S31), contrasting this approach with those who regarded quality of life as `subjective appraisals of life satisfaction' (citing Hunt & McEwen 1983), or those who combine a **patient's** subjective evaluation of well being with physical symptoms, sexual function, work performance, emotional status, etc. (citing Croog et al. 1986).

There would seem to be at least two versions of a position where objective, externally observable measures are exclusively made use of when assessing quality of life. These might be termed a `non observables are banned' position, and a `only in development' argument.

(a) *`Non observables are banned'*

The `non observables are banned' position states that any behaviour that cannot be directly observed and confirmed by an independent observer is unworthy of analysis. In the development of the SIP, Bergner et al. (1976a) reported deciding that of a feeling state, clinical, and performance conception of an individuals own health state appraisal, only the last of these, the performance conception, was suitable. The feeling state conception was ruled out on the grounds of being inaccessible to external validation, and the clinical conception was seen as unsuitable as it required medical interpretation and hence was reliant on the definitions of physicians and not the person concerned. The performance conception was adopted as it could be based on respondent report, but could also be

easily observed and reported by an untrained observer, and also allowed easy comparison between different diseases and dysfunctions.

The difficulty with this approach as that it makes determination of the relative importance of different forms of physical quality of life/objective well-being exceptionally difficult, as the subject's own view of relative desirability would be precluded. Either one does not develop a global index, or one arbitrarily assigns weightings so that dimensions/sub-scales can be combined to form a global index. Kaplan, Bush and Berry (1979) have referred to this issue in suggesting that the category rating task allows a single global rating to be given to total case descriptions so that the subject can consider the multiple dimensions of health jointly and simultaneously, and argue that this is necessary if arbitrary rules for combining attributes into a total case rating are to be avoided. A different means of using patient report to weight dimensions has been noted by Goodinson and Singleton (1989), who refer to a 1985 study by Ferrans and Power where Likert scales were used to measure satisfaction and then measures were obtained on the relevance of each item/domain to the individual, an aggregate QOL index formed by weighting each item/domain according to its reported relevance value and then adding together.

(b) `*Only in development'*

This alternative position states that any measure of subjective well-being reliant upon the report of an individual is liable to random measurement error, and hence it is preferable to develop an instrument which can be used to predict subjective well-being independently of the subject's own report. While subject report is useful and possibly essential for instrument development, it is to be avoided in instrument application, ie. when making a specific assessment.

This latter approach seems more reasonable. Basically it allows subject report measures to play a role as dependent variables when developing a test instrument, and grants subjective well-being an important role in the development of QOL measures.

An approach that attempts to minimise the role of a direct measure of subjective well-being in a test situation may be sensible as there is considerable evidence that direct report can be misleading. Evidence relating to this issue is discussed in the next section.

6 Factors that Influence Reported Subjective Well-being

(a) Life events and experiences

There is little doubt that subjective well-being is influenced by major life events and experiences, eg. housing, employment, health, marriage etc. and a great deal of research has been concerned with the notion that the major cause of change in SWB are major life events and experiences (Diener 1984; Heady et al. 1985; Heady & Wearing 1989).

The relationship between life events and SWB is not simple. As well as issues regarding the relative effect of different types of event, there are questions concerning the effect of overall SWB on satisfaction within a given domain. Among the variables commonly treated as affecting SWB are domain satisfactions (eg. with marriage, health, work etc.), major life events, and reference standards (eg. expectations, aspirations, sense of equity). Furthermore satisfaction within a given domain could conceivably be a consequence of SWB. For example, Heady et al. (1991) have argued that satisfaction with work, standard of living, and leisure satisfaction, are largely the result of overall life satisfaction, and that satisfaction with friendship and general fitness (as opposed to illness) appear to be explicable solely on the basis of personality; on the other hand satisfaction with marriage appeared to both influence and be influenced by overall SWB.

In terms of the effects of illness, Diener (1984) concluded that objective health is significantly related to SWB, although the relationship appeared to be much weaker than that between self-rated health and SWB. The relationship between health and SWB may be also be bi-directional: Hughes (1985) found that depression following lung cancer radiotherapy may exacerbate symptom distress (tiredness, anorexia, pain). Donovan et al. (1989) also pointed out that cancer patients experience positive impacts of the disease on their life as well as negative, eg. increased closeness to spouse.

(b) Personality variables as mediators

Evidence has been reported that personality traits of extraversion and neuroticism are highly stable and can predict SWB 20 years later (Costa & McCrae 1980, 1984, cited Heady & Wearing 1989). It has been argued that personality can heavily mediate the impact of exogenous life events, with each person having a `normal' equilibrium level of life events and SWB, predictable on the basis of age and personality; only when events deviate from equilibrium levels is SWB seen to change (Heady & Wearing 1989).

Individuals have also been shown to vary in coping strategies, which in turn can affect physical factors such as health outcomes. Greer (1979) and Pettingale (1984) showed that recurrence free survival at 5 and 10 years after surgery for breast cancer was related to psychological approach at three months (fighting spirit or denial were better than helpless/hopeless responses).

(c) The effects of adaptation

In addition to individual-specific variables, there are general psychological mechanisms that act to increase SWB independently of direct physical effects. For example patients frequently experience release of anxiety and stress in the initial stages of recovery following surgery (Cohen 1982, see Goodinson & Singleton 1989). Other effects can develop more steadily, for example Cassillet (1984, cited Breetvelt & Van Dam 1991) reported that patients with newly diagnosed illness had greater anxiety and depression than patients who had been living with the illness for longer periods.

This adaptation to illness has been much reported, with many studies suggesting that patients may differ from controls markedly in physical complaints while differing little or not at all in terms of psychological complaints. For example Cassillet et al. (1982, cited Breetvelt & Van Dam 1991) found that melanoma patients had superior psychologic well being to other patients suffering dermatological disorders, and moreover the mean score for patients was not different from that of the normal public.

Adaptation can be so great as to apparently eliminate SWB differences between people chronically ill and controls, or even those who have recently had very positive experiences. Brickman et al. (1984) found that lottery winners and quadriplegics differed little from normal controls in SWB, and De Haes and van Knippenburg (1984, cited Goodinson & Singleton 1989), reported how in many studies of QOL in cancer patients, no differences are found compared to benign controls.

These findings may sometimes reflect inadequacies in the QOL measurement instruments. However they also suggest fundamental homeostatic processes, such as the re-setting of expectations, change in reference standards, etc. Diener (1984) concluded that health does seem correlated with SWB, but that adaptation markedly reduces its influence.

How long it takes to adapt, to what extent people can and do adapt, and the factors determining this, seems still to be broadly unknown. However Breetvelt and Van Dam (1991) have reported interesting findings that suggest how adaptation may be measured independently of SWB. First observing that many studies which employ patient self-report suggest that cancer patients are not more anxious or unhappy than other patient groups or even the normal healthy population, they suggest that this seems to conflict with the everyday experience of physicians and other care takers. A recent paper by Epstein et al. (1989) has provided evidence of this, where family/friend care givers were asked to act as proxies for older chronically ill patients. Although proxy and subject-own responses were generally similar for overall health, functional status, and social activity, proxies rated subjects' emotional health and satisfaction significantly lower than did the subjects themselves (of course this could indicate inaccuracy of the proxies).

Breetvelt and Van Dam (1991) argue that the appropriate control for a patient's report of well being is not the report of healthy subjects, but `retrospective pre-test' (citing Hoogstaten 1985). They propose evidence that while patents may not rate their current level of well being differently to that of controls, patients may give a much higher rating to the state that they experienced prior to their illness or accident. In other words, patients rate themselves as being considerably happier in the past than do control groups.

Breetvelt and Van Dam (1991) attributed estimates of current SWB (placed at similar levels to non-patients) to a subjective rescaling of what constitutes happiness, ie. a criterion shift in terms of the quantitative level that constitutes normal well being. It was suggested there may additionally be a change in the relative weighting of psychological components versus physical components, ie. the dimensions patients use to assess well being.

This line of investigation also suggests that there may be real problems in assuming that self-report is a valid measure of SWB. For example if patients make judgements relative to an internal criterion that is in some way adjusted to bring about a positive report (eg. by making downward comparisons with patients even less well-off; Taylor 1983, cited Breetvelt & Van Dam 1991; Diener 1984), then cognitive factors may lead to what Breetvelt and Van Dam call `under reporting', ie. that `patients report less emotional distress, satisfaction or the like than is actually present' (Breetvelt & Van Dam 1991, p. 983).

That self report may be unreliable has been suggested by other studies. Bombardier et al. (1986) found when examining the ability of the QWB scale and other measures to detect the effects of auranofin (oral gold) on arthritis that self-ratings by patients failed to detect significant treatment effects that were indicated by other instruments. Tests of self-versus-interviewer test administration have also suggested the self report can reduce test validity (see Chapter F, section 6).

7 Implications of SWB Findings for QOL Measurement

Subjective well-being is influenced by factors other than external events or physical conditions. For the purposes of developing an instrument to measure health-related QOL, personality effects could possibly be ignored and treated as a random variable. However adaptation effects are a different issue: self-reported QOL may well differ from underlying QOL (ie. represent underreporting). The apparently malleable and relative nature of QOL self-assessments means self-reported QOL cannot be taken as a criterion measure for the QOL of a given health state (in the sense of validity assessment). The fact that an individual's report indicates acceptance of their state does not mean that they would not greatly prefer an alternative one if the choice was available.

The consequence of this analysis is that in developing a measure of health-related QOL subjective self-report data should be treated very carefully, and alternative methods of estimating current QOL need to be explored. Investigations are needed into:

(a) the factors and conditions determining the rate and extent of adaptation;

(b) the nature of adaptation, ie. can one accept reported SWB at face value, or does it represent under-reporting, where `actual' or underlying SWB is less than reported SWB;

(c) the value of techniques such as `retrospective pre-test', which may provide a more valid measure of `actual' SWB:

(d) whether care givers also experience adaptation (this may in part explain the diminishing relationship between patient report and doctor, as doctor's become more senior and experienced; see Chapter F).

While such basic research is conducted, instrument development that makes use of patient report information needs to proceed with caution, and the following guidelines are suggested:

1 Include a measure such as the `retrospective pre-test' as part of the test battery when assessing QOL.

2 Attempt to include the measurement of variables relevant to the assessment of adaptation. Even in the event that an adaptation-free measure of illness impact is possible, it is likely that adaptation has an effect on `underlying' as well as `surface' (ie. current reported) SWB. Furthermore illnesses may differ in the extent to which they allow adaptation: eg. illnesses where acuity fluctuates or is uncertain may be less readily adapted to, compared to stable conditions where the prognosis is clear.

3 Recognise that subjective well-being is not a criterion measure of QOL, but it may still be the single best indicator available.

It may be wise to include self-report wherever possible as a general catch-all in the event that there is a key dimension that is of specific but unexpected relevance to a given condition or program that is overlooked (also see Chapter E for Bergner's notion of descriptive validity, as well as Kaplan's criticisms of factor analysis in test formation).

4 Clearly separate the steps involved in the development of a test measure, from the steps involved in the application of the developed test. Self-reported SWB might play a central role in test development, but a different role, or no role at all, in the final instrument.

8 General Problems in the Measurement of a Construct such as HQOL

If QOL is identified with subjective well-being alone, then a simple approach might have been to ask the patient. However, as described, QOL estimates via simple self report may be misleading. Adaptive cognitive mechanisms can influence the report, resulting in the judgement leading to the QOL rating being in some way comparative, producing an over-estimation of QOL (eg. through a diminished threshold of what constitutes an acceptable QOL).

Even if the adaptive nature of SWB had not been made evident, some procedure for estimating QOL beyond a simple inquiry would still be necessary. A single response can be a rather unstable, complex, and possibly misleading indicator of an attribute, and in most behavioural measurement it is necessary to combine several responses and types of response if the estimate is to be reliable.

There seems to be no simple measure of health-related QOL and reported SWB is likely to

be an imperfect measure of `actual' SWB. A test that aims to measure health-related QOL (or SWB) is endeavouring to measure a construct. By construct is meant a hypothetical concept which can never be directly measured or absolutely confirmed (unlike physical attributes such as height or weight), but only inferred from observations of behaviour. To estimate the value of a construct it is necessary to establish an operational definition, ie. a rule or rules of correspondence between the construct and behaviours that indicate it. A test of the construct is then a procedure for obtaining samples of this/these behaviour(s) that allows the value of the construct to be estimated (see Crocker & Algina 1986; Anastasi 1990).

To construct a test to measure QOL requires consideration of the conditions under which numbers may be reliably and validly assigned to represent the magnitude or amount of a psychological attribute. It subsumes issues such as reliability, standardisation, validity, and scaling. In addition, and as pointed out by Donovan et al. (1989) in the context of health-related QOL, for a measure to be meaningful it needs to have the psychometric properties of reliability and validity.

The problems facing the measurement of constructs have been outlined by Crocker and Algina (1986). These are stated below, along with a brief reference to where they are addressed in this report.

1 No single approach to the measurement of any psychological construct is universally accepted. Measures of psychological constructs are always indirect, hence theorists who talk about the construct may select very different behaviours to define the construct operationally.

See Chapter E, Validity.

2 Psychological measurements are usually based on limited samples of behaviour - determining the number of items and the variety of content necessary to provide an adequate sample of the behavioural domain is a major problem in developing a sound measurement procedure.

See Chapter E, Content validation.

3 The measurement obtained is always subject to error. It is very unlikely that re-testing of the same individuals would ever be identical, due to fatigue, boredom, guessing, carelessness, etc. If a different form of the test is applied, scores may also change because of variation in content.

See Chapter D, Reliability.

4 The measurement scales will tend to lack well-defined units - the properties of the measurement scale, the labelling of the units, and the interpretation of the values derived are complex issues.

See Chapter C, Scaling.

5 The psychological construct measured by the test must be both operationally defined (ie. defined in terms of observable behaviour) and hence capable of being empirically demonstrated, and defined in terms of its relationship to other constructs or events in the real world.

See Chapter E, Construct Validity.

C SCALING

A necessary consideration when developing instruments for assessing QOL is the nature of the data that is input to form the instrument. The development of an instrument for measuring a psychological construct involves the hypothesis that the construct is a property occurring in varying amounts that can be quantified using a scaling rule or a theoretical unidimensional continuum, and entails determining the real-number properties the scale values on this continuum possess, ie. nominal/ordinal/interval/ratio. This in turn determines which statements concerning values on a scale are meaningful, and which mathematical analyses can be legitimately applied to them.

The relevant area in psychology is known as scaling, scaling being `concerned with the theory and practice of associating numbers with psychological objects' (Eyfuth 1972). Scaling has been heavily influenced by psychophysics, which in turn concerns the `manner in which living organisms respond to the energetic configurations of the environment' (Stevens 1972). Efforts to determine the functional relationship between a physical stimulus and perceptual experience produced methods which have been applied to areas away from the simple physical qualities of the environment (eg. to the scaling of attitudes).

1 Stevens: the Effects of Task on Number Properties

The early psychophysicists were concerned to study the relationship between measurement obtained in two different ways of what were presumed to be the same property, eg. they studied the relationship between weight, length and temperature defined by the response of human subjects as instruments, and weight, length and temperature defined by other measuring instruments such as scales, foot rules, and thermometers. A psychophysical law is a statement of the relationship between measurements obtained by these two methods. The experimental methods and statistical processes developed by the early psychophysicists have since been used in psychological testing and in the study of human ability, and have been developed and applied to measure human ability, personality, attitudes, interests, and many other aspects of behaviour.

In what is now classical work, Stevens (eg. see Stevens and Galanter 1957) divided stimuli into those forming prothetic and metathetic continua. Prothetic continua where seen to be concerned with `how much', ie. quantitative aspects; an example is loudness, where different levels were formed by adding more of the same. On the other hand metathetic continua were concerned with `what kind', or `where', ie. qualitative aspects; an example is pitch, where differences are due to the substitution of new frequencies for old.

Stevens described the main differences between prothetic and metathetic continua as being exhibited in the formal relations that could be observed among three primary kinds of scaling measures, ie. magnitude, partition, and confusion measures

In *magnitude* scales, the observer is asked to directly assign numbers to stimuli in

proportion to their apparent magnitude, or of ratios among apparent magnitudes. The function relating stimulus magnitude to subjective magnitude is generally determined to be a power one. Cross-modality matching can be used to validate the scaling produced in this way; in this process the functions relating, say, loudness and brightness are first determined, and then it is shown that the exponents of the two functions can be used to predict the third where brightness is directly matched to loudness.

With *partition* scales, on the other hand, the observer assigns one of a finite set of numbers to each stimulus, eg. the numbers 1 to 5, or adjectives such as small, medium, large. This is seen to represent judgements of subjective differences, or distances (also see Eisler 1962).

On quantitative (prothetic) continua, partition scaling methods (eg. interval, category scales) are usually curved relative to the magnitude scale, ie. the partition scale gives a smaller exponent. With qualitative (metathetic) continua, scaling methods are linearly related. The loss of linearity between methods was seen by Stevens as being a product of forcing the observer to partition the continuum, and prevent the making of a proportional number assignment that would preserve ratios: this restriction causes the dramatic curvature in the scale.

Confusion scales includes scales such as JND (Just Noticeable Difference), discrimination, paired comparisons, and successive intervals. These tasks tend to be concerned with issues of determining thresholds of no difference/difference. The common feature is that some measure of variability or confusion is taken as the unit, in the sense that if there was no noise or confusion in human judgements then the JND would become infinitely small. With prothetic continua confusion scales were reported to be logarithmically related to magnitude scales, while being linear on metathetic continua.

2 No Scaling is Direct

Since the early work of Stevens, there has been a retreat from the view that saw magnitude estimation as the method of choice for scaling psychophysical functions. In a major review of psychophysical scaling, Gescheider (1988) concluded that it is now clear that sensation magnitude cannot be measured directly by any method, including the method of Stevens which was argued to be direct, ie. methods that require subjects to assign numbers to stimuli that presumably represent sensation magnitude (eg. magnitude estimation or category scaling).

Responses in the scaling task are now seen to be a joint function of cognitive and sensory factors (first a sensory stage and then a cognitive stage), with the validity of a psychophysical scale able to be established only through an examination of how well the psychological responses of subjects can be predicted from theories of sensory and cognitive processes.

Attacks on magnitude estimation have come from Shephard (1981, cited Gescheider 1988), who noted that the subject response in magnitude estimation is really only a discrete verbal response that itself possesses no definite qualitative magnitude, and that evidence from cross-modality matches does not guarantee validity as subjects could simply be assigning numbers to sensation in a nonlinear but consistent way. Zwislocki (1983, cited Gescheider 1988) also found that individual subjects had unique but varying characteristic nonlinear functions when assigning numbers to sensation magnitudes.

It has also been observed that while average magnitude estimations are approximately a power function of stimulus intensity, subjects responses can be influenced by many stimulus variables other than individual stimulus magnitude (eg. the range of stimuli presented in the experiment; Marby & Cook 1986, cited Gescheider 1988). Models to explain magnitude estimation results now suggest the subject rehearses the psychological continuum in which the stimulus is presented, with the subjects response determined by the location of the stimulus perceived on the continuum relative to various possible anchors.

There is now overwhelming evidence that responses in psychophysical experiments can be biased in many ways, with effects reported in both magnitude estimation and category rating tasks. Which effects appear (some biases are contradictory) seem dependent upon subtle conditions of instructions, stimuli, and response. Gescheider (1988) reviewed numerous examples, including:

- sequential bias effects, a bias to report values similar to the last reported (successive measures are correlated). The result of this is seen to reduce the response range, through assimilation of the responses toward the centre.

- contraction bias, where the subjects response are closer to the centre of the response range than they should be (possibly a product of sequential dependencies)

- instruction or `framing' effects, eg. where numerical examples are given in magnitude estimation tasks, the function has been found to vary depending on the size of the numerical ratio given as example

- the tendency to use categories equally often in category rating

- stimulus frequency bias, where stimuli presented that are a little larger or smaller than a frequently presented stimulus are judged to be excessively different

- stimulus equalisation bias, a tendency to use the full range of responses whatever the size of the range of the stimuli

On the other hand it has been claimed (eg. by Zwislocki and others; see Gescheider 1988), that subjects can make judgements of sensation magnitude that are relatively immune to context effects if specifically instructed to assign numbers to stimuli in such a way that the

impression of the size of the number matches the impression of the sensation magnitude of the stimulus. When subjects are so instructed and asked to judge each stimulus independently, stimulus context effects are claimed to be relatively small (as opposed to where, say, example stimuli are given with an assigned number standard). However others have disputed this (eg. Mellers 1983 cited Gescheider 1988, claimed all judgements to be relative and occurring in context, with instructions and stimulus factors exerting a controlling influence), and at the time of Gescheider's review the issue was in doubt.

It is apparent that the defining of a category rating or magnitude estimation task requires a much more complex model of cognitive processing than the psychophysical one.

3 Magnitude Estimation and Category Rating may Measure Different Things

As noted, a new perspective on the interval scaling debate arose with the growth of cognitive psychology, with its emphasis upon cognitive processes as determinants of perception and response. In this tradition may be placed work by Anderson (1976), who addressed the problem of the `equal-interval' scale in psychological measurement as part of an attempt to develop a general theory of information integration.

Anderson's `method of functional measurement' proposed that whether or not a given task produces an interval measure of an underlying variable or construct may be determined through an examination of the effect of two or more variables on the response scale. If the response measure is an interval scale, and two stimulus variables combine linearly, then this should result in a nonsignificant F ratio for interaction and with a two-way graph of the data appearing as a set of parallel lines. If parallelism was found to be present, three goals were seen to be simultaneously realised:

1) a linear model was supported;
2) the response was shown to be on an interval scale; and
3) interval scales of the stimulus variables were indicated.

Alternatively, if the plotting of data produced a `linear fan', then

1) a multiplying model was supported;
2) the response is indicated to be on an interval scale; and
3) interval scales of the stimulus variables are demonstrated.

The functional measurement approach can be applied to any scaling procedure. Anderson's own results (Anderson 1976) suggested that category judgements did yield interval scales, and as category ratings are almost always non-linearly related to magnitude estimation, magnitude estimation must not. However, since then, the work of others (eg. Marks 1974, cited Gescheider 1988) has been claimed to indicate that magnitude estimation, too, satisfies functional assessment requirements.

Gescheider (1988) suggested that both category scaling and magnitude estimation appear to pass a variety of interval-scaling tests, and that perhaps each procedure produces valid

measurements of different psychological processes. Magnitude estimation tasks require subjects to give a response in terms of apparent magnitude, while category judgements may entail judgements of difference: the sensory magnitudes and sensory dissimilarities of stimuli are equally meaningful but different dimensions of experience. Magnitude estimation and category rating are equally valid because they measure different things.

Gescheider concluded that the observed non-linear relation between scales obtained by different methods was the most perplexing and one of the oldest problems in psychophysics, and whether the non-linearities are due to cognitive-judgement factors or sensory-perceptual factors was yet to be determined.

4 Scaling in Health-state Assessment

When testing, say, illumination, one can present a range of stimuli and determine a subject's ability to discriminate among stimuli and estimate differences. But for health states, one cannot present the experience of ill health to subjects. Either one uses a range of patients who have experienced different states and use a common measuring tool, or a single group of subjects and ask them to `imagine' different states. Neither situation could be said to correspond to a classical psychophysical experiment. None the less, psychophysics has been the model for scaling studies, and while the sensation-measurement aspects are likely to be doubtful, health status scaling can be expected to be effected by cognitive factors of task, instruction, and stimulus in a similar way to that found in psychophysical measurement. Experimental findings regarding rating versus magnitude estimation illustrate this. Findings relevant to the issue of category rating versus magnitude estimation in health-related quality of life (HQOL) measurement are discussed below.

4.1 Does category rating provide an interval scale

The developers of the Quality of Well-Being (QWB) Index have invested significant effort in attempting to determine the characteristics of different scaling procedures. Patrick, Bush and Chen (1973) examined category rating, magnitude estimation, and an equivalence task as methods of measuring preference for health status scenarios, employing states that covered the preference continuum from 1.0 for complete well being, to 0.0 for death (each scenario containing a functional description in terms of mobility, physical and social activity; age; and a symptom-problem complex or CPX - for more details of the QWB, see Chapter G).

The category rating task required subjects to score scenarios on an 11 point scale, where `most desirable' was stated as above 11, and `least desirable' as below 1. The magnitude estimation task required subjects to assess scenarios on a 1000 point scale, where 1000 equalled a standard scenario of a `as healthy as possible' person, and each scenario was to be rated in terms of fractions of this standard healthy person, ie. if a scenario described a person half as healthy as the standard, then a score of 500 should be given. Equivalence

measured trade-offs in numbers of individuals for a given state against a standard population and health state.

The results were that no significant differences were found in the values assigned to items by comparable groups, and the relation between category and magnitude scales was found to be linear. However, as noted by Patrick et al., the magnitude estimation scale was anchored at the upper end as perfectly well function, value 1000, and did not allow the response to be arbitrarily large (as would be allowed if, eg. a central value was given as an example) and this could have converted it to a 0 to 1000 category rating scale. The functional measurement test of Anderson was also applied and revealed results consistent with the separate scales possessing interval properties.

In contrast to Patrick et al., Kaplan, Bush and Berry (1979) did report finding a logarithmic relationship between values obtained via category rating versus magnitude estimation (as also did Kind & Rosser 1988). In Kaplan et al's experiment the magnitude estimation task differed from that of Patrick et al. in that the standard was selected from the middle of the scale, with the result that the scale was now unbounded. Kaplan et al. (1979) pointed out the notion that on logical grounds either category rating or magnitude estimation could be providing interval data, but not both, and concluded in favour of category rating as the magnitude estimation results were seen to provide `intuitively unreasonable' results when interpreted in terms of a 0 to 1 scale. Furthermore Kaplan et al. reported that Anderson's functional measurement approach applied to category rating data in a 100 subject test (hence possessing much statistical power), produced highly significant main effects of social activity and level of well-being, but no interaction between the effects (F-ratio < 1, and the graphed lines per effect being parallel). This was consistent with category rating providing an interval response scale.

As reported in the previous section, however, others have suggested that magnitude estimation tasks can satisfy Anderson's test, and it cannot be concluded that the presence of defined endpoints on the response scale, ie. a `bounded' scale, category rating' as defined by Kaplan et al. (1979), uniquely provides interval data. On the other hand it does seem that defining a range of allowed responses when estimating stimulus size causes a fundamental change in the nature of the task, and this may be the most useful distinguishing feature of a category rating versus magnitude estimation task.

4.2 What is the effect of providing defined versus vague endpoints to a scale

Not withstanding the `reasonableness' intuitions of Kaplan et al. (1979), the argument that a scale with well-defined endpoints uniquely possesses the ability to provide an interval scale for health state assessment cannot be upheld. However there are grounds for supporting such a scale beyond Andersons test of functional assessment. This evidence comes from investigations by Kaplan and Ernst (1983) of the reported tendency for distribution effects in category rating tasks, ie. that subjects tend to spread their responses across all the allowable categories (Steven & Galanter 1957; Parducci 1968), Kaplan and Ernst conducting, experiments to investigate whether and under what conditions distribution effects occurred.

In their first experiment, Kaplan and Ernst used a 10 point rating scale, and instructions that included descriptions of a completely well person and a person in a coma so as to define clearly the end-points of the scale. Different groups of subjects were presented with four different groups of health state scenarios: all high scale items, all low, all medium, and mixed.

Analysis of the ratings of health state descriptions by subjects who saw only high scale values produced some evidence of distribution effects. There was a slight trend for these subjects to spread their ratings of high items across the response range (from 0 `as bad as dying' to 10 `completely well' relative to the assignment of the same sub-set of high scale items by different subjects who were presented with the high sub-set in the context of a broader overall range of items. This trend was not supported in a second experiment.

However a further experiment suggested that such effects could occur under conditions where subjects did not have or were not given clear information regarding the types of items that should define the end points of the scale. In this experiment subjects were to assess acts of immorality, and those subjects given instructions which clearly defined the endpoints did produce responses distributed differently to those given minimal instructions.

Kaplan and Ernst concluded from this that health state assessments using a scale with poles of death to complete health are naturally readily understood and hence well defined, and therefore tend to be resistant to context effects. On the other hand, to minimise context effects of the distribution type it was suggested that:

1 the continuum along which states are to be rated should be well defined
2 the end points should be clearly defined
3 the stimuli should not be available for inspection prior to their rating

To this might be added the provision of enough scale points to allow maximum discrimination between judgements.

The provision of endpoints to a scale, but uncertain ones, may in fact lead category rating to produce ordinal scales. For example, Read et al. (1984) appeared to require subjects to define the extremes of the scale themselves, picking the worst and best outcomes (where the best was not a state of perfect health, and then filling in the medium values. This explicit requirement to use all the values of the scale would seem to counteract its `absolute' nature, ie. it encourages subjects to make relative selections of scale points.

A related issue concerns that of the zero point, and the defining of death on a scale. It has been clearly demonstrated that there is a need for estimates corresponding to states worse than death (eg. Read et al. 1984; Kind & Rosser 1988; Sutherland, Dunn & Boyd 1983). This poses problems both for advocates of category scaling (problems in clearly defining the poles, so encourage distribution effects), and also for those favouring magnitude estimation. Sutherland, Dunn and Boyd (1983) raise the question of can you have ratio

scale, as opposed to interval or ordinal, when the zero point is indeterminate.

This latter point was indirectly addressed by Haig et al. (1986), who inverted the traditional poles of the illness-health continuum such that now the absence of dysfunction or discomfort (corresponding to perfect health?) became zero, with the other end of the scale open-ended. This variant of magnitude estimation entailed subjects being given the zero state as the first state, which by definition was zero. The second state was randomly selected, and subjects (inpatients of a surgery ward) were apparently free to assign any number to it that appeared to correspond to the magnitude of its undesirability, with all subsequent states in proportion to the earlier judgements.

Haig et al. reported finding a linear relationship between this scale and earlier results reported by Bush and co-workers using category **rating,** while at the same time claiming that the scale was a truly ratio one because it employed magnitude estimation after Stevens, and death did not represent what they (as have others) found to be an inappropriate zero. The failure to find a logarithmic relationship between this and Bush's scale was not explained.

4.3 Effects of task complexity

If one wishes to minimise the influence of complex cognitive variables, eg. task complexity, etc. and focus on the ostensible aspect of interest, ie. desirability or otherwise of a health state, then it is sensible to use a task that is as simple as possible. Patrick et al. (1973) found that both category and magnitude estimation procedures were easy to use, as opposed to the method of equivalent stimuli which was complex, and reported to be `unrealistic, emotive, confusing, and offensive to some judges' (consistent with this, the equivalence [tradeoff] method also tended to give the largest standard deviations).

In contrast, Torrance (1976) found a test that employed category scaling as the hardest to use, compared to standard gamble or time trade-off. In Torrance's category scaling the subject had to indicate the relative desirability of three-paragraph health scenarios on a 0 to 100 continuum (using a visual analogue scale), being asked for each of a series of health states to mark three lines - one for the case where the subject has three months left to live, one for 8 years, and one for normal life expectancy. This seems a very complex task, with the subject having to estimate and combine QOL per state and length of time simultaneously. It illustrates the importance of looking at the whole task, not just some abstracted element of it (Kaplan, Bush & Berry 1976, suggested the result was due to item complexity and because the category rating task was presented first).

Kaplan and Ernst (1983) have also argued that the magnitude estimation task, by its very nature, involves response scales that are not clearly understood by subjects, and hence are likely to be particularly prone to distribution bias.

4.4 Effects of stimulus materials

Llewelyn-Thomas et al. (1984) examined the effect on numerical values assigned to health states of presenting a written description in formalised point form as used in the development of the Index of Well-being (Patrick, Bush & Chen 1973) versus scenarios similar in style to that used by Torrance (written in detail in the first person singular in common language). The subjects were outpatients with malignant disease, and both standard gamble and a category rating task were used. The point form scenarios were found to produce higher values than the narrative form standard, and standard gamble provided substantially different and systematically higher scores than did category rating (although there was a significant interaction effect of task order). Also effects of instruction have been demonstrated by Tversky and Kahneman (1981), and McNeil et al. (1982).

4.5 What is a category rating task?

It is apparent that what constitutes a category rating task means different things to different people, and as pointed out in 4.1 and 4.2 it may be that clearly defining the boundaries of the scale is a more important aspect of a task than whether the subject is requested to estimate stimulus values in terms of ratios of a standard or differences. Brooks (1991) in a very useful review exhibits another confusion in commenting upon `perhaps its .. [ie. the category rating task] .. most attractive feature being the visual nature of the approach, which can help raters conceptualise what is required of them in evaluating health states' (Brooks 1991, p. 19). This is in one sense a good point, but it confuses the use of a visual aid with what constitutes category scaling. A visual scale can be used for the explicit ordinal scaling of states, and a task can require interval-type judgements on a clearly bounded scale without reference to a visual aid. It seems time to abandon simplifying terms such as `magnitude estimation' and `rating scales' and clearly describe the task in all its features.

5 How Important are Interval Properties for Statistical Operations on Health States

It has been assumed so far that a response scale must possess interval properties if it is to be useful in health status measurement. This is by no means clear.

Firstly, it can be argued that most if not all psychological variables are in fact ordinal, although for statistical purposes they are, quite *justifiably,* commonly treated as if they were interval or ration variables.

Ferguson (1966), for example, suggests in the analysis of statistical data in psychology and education it is common for the information to be superimposed on the data, eg. to assume that a set of ordinal numbers can be replaced by the cardinal numbers, and to then proceed to apply arithmetic operations to the numbers. This involves assumptions about the equality of intervals, when in fact the measuring operation does not yield this. Ferguson suggests that scores on intelligence tests, attitude tests and personality tests are in all effect ordinal variables, for no aspect of the operation of measuring intelligence, say, permits the making of meaningful statements about the equality of intervals. It cannot be

said that the differences in intelligence between a person with an IQ of 80 and one with an IQ of 90 is in any sense equal to the difference in intelligence between a person with an IQ of 110 and another with an IQ of 120. Much the same argument could be applied to a subjective well-being scale. Although a logical purist may conclude from this that the testing of data in this way should be discontinued, Ferguson disagrees. Practical necessity can dictate a procedure, although it should be understood what assumptions are being made.

A second and related argument, is that most parametric statistical tests do not require data to be measured on an interval scale, but rather that tests require assumptions about the distribution of the data: parametric statistics can be used as long as the data meets these distribution assumptions For example Anderson (1976) proposed that the present concern with interval or linear scales is in many applications unnecessary, as the power of a statistical test has no necessary relation to the nature of the response scale. A nonlinear or ordinal response may well provide normal distributions.

Some have responded to this view by testing whether the distribution is normal, and then correcting non-normal data using eg. log transformations (cf. Hall et al. 1989). Others have been even more pragmatic. Crocker and Algina (1986) concluded that psychology test data should be treated as interval scale data as long as it can be demonstrated empirically that usefulness of the scores for prediction or description is enhanced by this treatment. Even Stevens who popularised the ordinal/interval debate was quoted as sanctioning the use of parametric statistics on the pragmatic grounds that it can lead to fruitful results. Shavelson (1988) in a recent text on statistical reasoning for behavioural science, illustrated the hands-off approach of most to this issue, in stating that `The problem, then, is not so much the match of statistical method measurement scales. Rather the problem is one of interpreting the results of a statistical analysis .... ', and then quoting Hays in that `the experimenting psychologist [sociologist, educational researcher] must face the problem of the interpretation of statistical results within psychology and on extramathematical grounds' (Hays 1973, p. 88, cited Shavelson 1988).

It would seem that as far as inferential statistics is concerned, it is not necessary to show evidence of the measurements being interval in nature when measuring the effect of a treatment. To a large extent procedures may be used if it can be shown they are useful, ie. on the balance of the evidence they appear to have practical, predictive value.

While this approach may legitimate tests which suggest tests of significance, it could be argued that this only supports attempts to test the relative order of different HQOL estimates, and that one still cannot assume that equal differences on the scale represent equal quantities of health. When concerned with the relative allocation of quantities of resources, then a ratio scale may be needed (see Kind & Rosser 1988). Countering this Anderson (1976) has argued that when using linear regression for prediction, which can be the aim of a test to predict HQOL predictive accuracy will tend not to be affected by nonlinearity in the stimulus values employed in the regression. This is a considerable advantage for practical prediction (although considerable caution still need to be applied when interpreting parameters estimated from the regression equation).

6 Conclusions and Observations

It is now clear that sensation magnitude cannot be measured directly by any method, including magnitude estimation and category scaling. Responses in any scaling task are now recognised to be a joint function of cognitive and sensory factors. It is apparent that the defining of a category rating or magnitude estimation task requires a much more complex model of cognitive processing than the psychophysical one.

Furthermore both category scaling and magnitude estimation appear to pass a variety of interval-scaling tests, and perhaps each procedure produces valid measurements of different psychological processes. However Kaplan and Ernst have provided evidence that a form of category rating task does seem more resistant to inter-item context effects, namely one where the both the continuum along which states are to be rated and the end points are clearly defined. Task complexity is also likely to have a major effect in producing undesirable error variance (ie. variance related to task factors as opposed to health state factors), and standardisation of stimulus materials and instructions is essential.

It also needs to be clear that no psychological scale may be interval, in the sense that equal quantities at different parts of the scale cannot be treated as meaningfully equivalent (eg. intelligence). The question is what do we want to measure, and how do we measure it so that responses are reliable. Task factors will change results, be they the product of the particular set of stimuli presented, examples given, whether a scale with defined endpoints is presented, or only one standard: people appear to use all available information to define the task not just the single sentence or paragraph statement of task as given by the experimenter. In scientific research into human behaviour (which includes preferences), then the aim is to examine differences between situations, not the absolute level of something. Is there an underlying value to be tapped - or is it created especially for the experiment? Assuming that the former is at least partly the case, then the issue is not so much does its measure have interval or ratio properties, but if they are at least ordinal, then in what way can we legitimately aid decision making in that area. One should aim to clarify the purposes of the investigation as much as possible, defining the type of underlying variable that is to be estimated, then do all one can to minimise undesired sources of variance.

This introduces us to the issues of reliability and validity, addressed in the following chapters.

D RELIABILITY

For any test or measure of health status to be useful it must be reliable, that is, repeat measurements made under constant conditions need to give the same result. As put by Anastasi (1990), measures of test reliability allow an estimate of the proportion of test variability that is due to error variance, where error variance is change in scores due to anything other than the characteristic of interest. The need to minimise error variance is the reason why psychological tests typically specify all aspects of the test environment, ie. instructions, time limits, mode of subject-tester interaction, etc. the aim being to eliminate or control extraneous sources of variance that would otherwise effect test results.

The basic measure of test reliability or reliability coefficient is the correlation coefficient, which indicates the consistency between two independently derived sets of scores. The most common of these is Pearson's Product Moment Correlation, which measures the location of items in the two variables to be compared in terms of the amount of deviation each item displays above or below their respective group means (ie. determines the standard scores for all test scores in each variable), and calculates the product of the paired scores. The Pearson correlation coefficient is then the mean of these products (computational formulas simplify this process). Significance tests then determine the probability of the observed correlation occurring by chance alone.

It is possible to estimate the reliability coefficient for an instrument either through repeated presentation of a test or presentation of parallel forms of a test, or through a single test administration. The advantages and disadvantages of these methods are briefly reviewed below.

1 Types of Reliability

1.1 Test-retest, or measure of stability

The most obvious means for estimating the reliability of a test is through representing the identical test on a later occasion. The correlation coefficient for a test-retest procedure is termed the coefficient of stability, and the Pearson product moment formula can be used. Crocker and Algina (1988) state that few if any standards exist for judging the minimally acceptable value for a coefficient of stability, but that commercially published individually administered aptitude test are amongst the highest. Subsets of the WAIS (Weschler Adult Intelligence Scale) have coefficients in the .70s, .80s, and low .90s. Personality, interest or attitude measures are often lower than these, but Crocker and Algina propose that well constructed test should still have test-retest coefficients in the .80s.

Although apparently simple and straightforward, the problems attendant the test-retest method are major. Basically the two tests can not be considered as independent due to practice and/or recall, and while the longer the interval between test and retest the less the risk of memory effects, the greater is the risk of intervening events causing respondents to

change their views. The obtaining of a low coefficient may mean either the test is an unreliable measure of the trait, or the trait itself may be unstable. Alternatively the testee's behaviour may have been altered by the first administration, and the second test may reflect effects of memory, practice, learning, boredom, sensitisation etc.

In assessing the correlation it is hard to decide if it has been inflated (due to memory effects), or deflated (due to change in views).

1.2 Alternate form method, or measure of equivalence

A way of avoiding the difficulties of test-retest is through the use of alternate forms of a test. In the development of alternate forms care is needed to determine that the forms are truly parallel, and are independently constructed to meet the same specifications (Anastasi 1990).

Crocker and Algina (1986) propose that any tests that have multiple forms should have some evidence of their equivalence. To test this equivalence the two forms are administered within a very short time period, allowing only enough time between testings so that the testees are not fatigued. The order of administration should be balanced. Pearson product moment correlation coefficient could then be computed between the two forms, to form the coefficient of equivalence.

Crocker and Algina also state that there are no hard and fast rules for what constitutes a minimally acceptable value for alternate form reliability estimates, but that many standardised achievement test manuals regard coefficients varying in the .80s to .90s for this type of reliability. In addition means, SDs, and standard errors of measurement should be reported for each form and these should be `quite similar'.

Anastasi (1990) concluded that while more widely applicable than test-retest reliability, alternate form reliability also has limitations. Alternate form tests can still be subject to practice effects, and differences between the two sets of answers will be a mixture of differences between the items used and other sources of error. Also there may be real problems for many tests of constructing truly equivalent forms, and for these reasons other techniques for estimating reliability may be required.

1.3 Measures of internal consistency - Split-half method

If only one administration of a single form is involved, then changes in test result due to changes in the construct under examination are less likely to occur (although some changes could occur within the period of the test itself). Procedures designed to estimate reliability in these circumstances are called *measures of internal consistency,* and are primarily concerned with errors caused by content sampling (although errors of measurement because of faulty administration and scoring, guessing, and temporary fluctuations of individual performance within the testing session may also affect the internal consistency coefficient). Crocker and Algina (1986) note that measures of internal

consistency are very important in many tests, as the aim of a test is generally not to estimate how the testee would score on the items presented but how the testee would score on a larger content domain of possible items that might have been asked

The original measure of this type was the *split-half* method, once the most widely used way of estimating how consistently the testee performed across items or subsets of items on the single test form (Moser & Kalton 1979). This method involve the division of the test into two sub-tests, each half the length of the original test, the two half-tests then scored and the correlation coefficient computed between the two scores.

The coefficient so obtained will be an underestimate of the reliability coefficient for the full-length test, for longer tests are generally more reliable than shorter tests because errors of measurement due to content sampling are reduced (see section 2.2 this chapter). To correct this the Spearman Brown prophecy formula can be employed, or other correction methods (eg. Rulon method), to give the stepped-up reliability of the whole test or r_w (see Anastasi 1990; Crocker & Algina 1986).

Split-half methods have been criticised on the grounds that there are many ways of dividing a test into halves, and these can result in different reliability estimates. For this reason, methods based on item covariance may be preferred. This will not apply to a highly heterogeneous test, for which split-half or even test-retest may be the most appropriate method.

1.4 Measures of internal consistency - Methods based on item covariance, or coefficient alpha.

A fourth method for finding reliability, also utilising a single test administration, is based on the consistency of responses to all items in the test.

Anastasi (1990) notes that this *interitem consistency is* influenced by two sources of error variance, namely content sampling (as also applies with alternate form and split half tests), and heterogeneity of the behaviour domain sampled. The more homogeneous the domain, the higher the interitem consistency.

A highly relevant question in this context is whether the criterion that the test is trying to predict is itself relatively homogeneous or heterogeneous. `A single homogeneous test is obviously not an adequate predictor of a highly heterogeneous criterion' (Anastasi 1990, p. 123).

The most common procedure for finding interitem consistency is the Kuder-Richardson reliability coefficient, which is equivalent to the mean of all split-half coefficients resulting from different splittings of a test. This is in contrast to the ordinary split half, which is a planned split designed to yield equivalent sets of items. Hence unless the test items are highly homogeneous, measures of interitem consistency will yield much lower coefficients than the split-half method (Anastasi 1990). Cronbach's Alpha and Hoyts Analysis of

Variance are said to yield identical results to Kuder-Richardson, (Crocker & Algina 1986), all determining the ratio of the sum of the item covariances to the total observed score variance.

2 Factors that Affect Reliability Coefficients

2.1 Characteristics of the subjects: variation in the behaviour, and ability to perform the measurement task

An important factor influencing the size of a reliability coefficient is the range of individual differences in the group, or subject homogeneity.

The magnitude of a reliability coefficient depends on variation among individuals on both their `true' scores (ie. their score on the underlying behaviour of interest), and error scores. Thus the homogeneity of the test group is a major consideration. If members of the test group are similar with respect to the trait being measured then the reliability coefficient **will be much lower than if** they varied markedly regarding the trait, because the random error variance will tend to be constant for the two groups (if of equal size), while `true' score variance will be much less and hence account for a much smaller proportion of the observed score variance.

In other words a test is not reliable or unreliable, rather reliability is a property of the scores on a test for a particular group of subjects. A consequence of this is that to compare tests it is essential to determine whether reported reliability estimates were based on samples similar in composition.

A further factor influencing the size of the reliability coefficient is the ability level of the group upon which the test was developed. For example, if the subjects are confused or pressured by the demands of the task then there is likely to be increased error variance, and hence reduced test reliability. On the other hand it would be difficult to justify eliminating such subjects, particularly if they over-represent a particular group, eg. patients or patient-relatives, as opposed to health professionals. Patients and their relatives may possess less formal education and be relatively unfamiliar with formal testing procedures but also possess relevant and valuable knowledge of health status variables and their effects. Eliminating such subjects because they fail to master the task is likely to reduce the test's validity. Even when no rater elimination occurs it is important to assess consistency across raters, ie. inter-rater reliability, particularly where measures are made of complex judgements.

Examiner variance as a source of error, ie. variation in test scores as a product of experimenter/examiner factors (Anastasi 1990, p. 125), should also be considered, and, if necessary, controlled for.

2.2 Test items: number (test length) and homogeneity

As already noted, test length affects both true score variance and observed score variance. Longer tests have greater test reliability than shorter test composed of similar items (errors of measurement due to content sampling are reduced).

The Spearman-Brown prophecy formula can be used to estimate the effects on reliability of increasing or decreasing test size. Thus if two equivalent items could be used in a scale, but only one is in fact used, the correlation between the two items is a measure of the internal consistency reliability for the one item. As pointed out by Moser and Kalton (1979), if the correlation between the two items was 0.5, and both items were used in the test, the Spearman-Brown formula calculates their reliability at 0.67. If further items all intercorrelated 0.5 were added the reliability would increase further. The higher the intercorrelation between items, the less the number needed to reach a given level of reliability. Thus only 4 items intercorrelated at 0.7 are needed to reach an r_w = 0-9, but 9 if the items were intercorrelated 0.5.

Moser and Kalton (1979) note that with attribute measurement the set of items rarely intercorrelates highly, and in order to attain an adequate level of reliability multiple items are needed. The higher the intercorrelation between items, ie. the greater the item homogeneity, the less items are needed. However item homogeneity may only be obtained by restricting the breadth of the scale, and this may have a direct effect on reducing validity.

3 Reporting and Interpreting Reliability

The reliability of a test may be expressed in terms of the *standard error of measurement or SEM,* also termed the standard error of the score. The SEM is particularly suited to the interpretation of individual scores, and Anastasi recommends that when the score for a test is reported an indication of its expected error should be provided. The standard error of measurement may be considered as the average standard deviation of examinees' individual error distributions for a large number of repeated testings, and allows the establishment of a confidence interval in which the true score is expected to lie (for the formula see Anastasi 1990; Crocker & Algina 1986).

The standard error of measurement and the reliability are alternative ways of expressing test reliability, although the reliability coefficient is best for comparing the reliability of different tests, while the standard error of measurement is recommended for the interpretation of individual scores. The interpretation of score differences should be made via measures of the standard error of score differences (see Anastasi 1990).

Finally, it is generally accepted that the developer of any test has an obligation not only to investigate the reliability of the test, but to report this information. Guidelines for this, derived from the American `Standard for Educational and Psychological Testing' (1985) are given by Crocker and Algina (1986, p. 152).

4 Conclusions

To summarise reliability measures, for any test or measure of health status to be useful it must be reliable, that is, repeat measurements made under constant conditions need to give the same result. The test-retest method of forming a reliability coefficient has the disadvantage that the obtained value may be inflated (due to memory effects), or deflated (due to change in views). The alternate form method has similar if diminished limitations. Of internal consistency measures, the ordinary split-half has the advantage of being a planned split, which may be appropriate if the items were planned to be highly heterogeneous. If the test items are planned to be homogeneous, then an interitem consistency measure such as Kuder-Richardson 20 or Chronbach's Alpha would be more appropriate, noting that unless the test items are highly homogeneous measures of interitem consistency will yield much lower coefficients than the split-half method.

Whatever the method chosen, much care needs to be taken when interpreting reliability coefficients, for a test is not reliable or unreliable, rather reliability is a property of the scores on a test for a particular group of subjects. A consequence of this is that to compare tests it is essential to determine whether reliability estimates were based on similar subjects, for the size of a reliability coefficient will tend to be proportional to the *variability between subjects* on the characteristic being assessed.

It also needs to be borne in mind that test reliability is directly related to *item homogeneity.* The greater the item homogeneity, the less items are needed for a given correlation size. However item homogeneity may only be obtained by restricting the breadth of the scale, and this may have a direct effect on reducing validity. Thus when a test is assessing a highly heterogeneous construct such as QOL there is likely to be a trade-off between test reliability and test validity.

Ultimately, a high reliability coefficient indicates there is consistency in a testee's scores, but it does not ensure that the inference to be drawn for the test is correct. Validity refers to the ability of the scale to measure what it sets out to measure, so differences between individuals truly reflect differences in the characteristic under study. Reliability is a necessary but not sufficient condition for validity. When examining reliability coefficients healthy scepticism is recommended regarding what is purported to be demonstrated, combined with a close examination of the item and subject conditions.

E VALIDITY

Cronbach (1971, cited in Crocker & Algina 1986) describes validation as the process by which a test developer or test user collects evidence to support the type of inferences that are to be drawn from test scores. To plan a validation study, the desired inference must be clearly identified and then evidence gathered.

Anastasi (1990) states that fundamentally all procedures for determining test validity are concerned with the relationships between performance on the test and other independently observable facts about the behaviour characteristics under consideration.

1 Content Validation

A major priority in constructing any test is to ensure that all the elements necessary to measure the construct have been included. With a multidimensional construct there is a need to develop an appropriate weighting of these dimensions. This is the area of content validation.

Content validation involves the systematic examination of the test content to determine whether it covers a representative sample of the behaviour domain to be measured (Anastasi 1990). The items that form the test should convey the attribute, and also cover the full range of the attribute in a balanced way. The items need to be a representative sample of the universe of content. The purpose of content validation is to assess whether the items truly represent the performance domain or construct of specific interest.

In conducting content validation, Crocker and Algina (1986) note that considerations include how should separate elements that make up a test be weighted; and how we should determine whether all elements necessary to represent the construct have been included. Anastasi (1990) makes the point that the domain should be defined in advance rather than after the test has been prepared.

Kaplan, Bush and Berry (1976) have argued strongly for content validity as the basis of health status validation. This is issued is addressed in much greater detail in Section F.

2 Criterion-related Validation

Criterion-related validation procedures indicate the effectiveness of a test in predicting an individual's performance in specified activities. It entails performance on the test being checked against a *criterion,* that is a direct and independent measure of that which the test is designed to predict (Anastasi 1990), and validity can be estimated based on the correlation coefficient between predictor and criterion scores.

With criterion -related validity the scale is developed as an indicator of some observable criterion, ie. the behaviour of interest can be directly measured, but either the criterion is available at the time of testing but a test is quicker, simpler, or less expensive *(concurrent*

validity), or the behaviour could only be determined in the future *(predictive validity).*

A test may be validated against as many criteria as there are uses for it (Anastasi 1990). Such criteria may come from personal judgement by one considered expert in the area, eg. as with a psychiatric diagnosis, assuming that the diagnosis has been based on prolonged observation and detailed case history, etc. and is itself valid. Anastasi (1990) points out that ratings have been employed in the validation of almost every type of test, and while subject to judgemental errors, they represent a valuable source of criterion data (when obtained under carefully controlled conditions, and raters are adequately trained, etc.; see Anastasi 1990, pp. 645-647).

Kaplan, Bush and Berry (1976) argue that criterion validity is not possible for a broad health status measure because no criterion exists that accurately measures the phenomena of interest, and indeed the lack of such a measure is the reason why effort has gone into the development of health status measures (if a criteria exists, only greater practicability or less cost justifies the use of some other measure). This is doubtless true, but it should not obscure the fact that while no criterion may exist, there may (as discussed in Chapter B, and see 3.2 following) be a measure (or measures) that is close to such a criterion.

3 Construct Validity

The construct-related validity of a test is the extent to which the test may be said to measure a theoretical construct or trait (Anastasi 1990), where the construct is manifested in a variety of behaviours and where there is no single behaviour that is seen to represent it comprehensively and be measured. Crocker and Algina (1986) describe a psychological construct as `a product of informed scientific imagination', which is not directly observable. Examples are intelligence, creativity, neuroticism, etc. To be useable a construct needs to be defined (operationally or semantically), and its relationships with other constructs and measures of specific real-world criteria specified.

According to Crocker and Algina, the process of construct validation might involve:

a) defining the construct
b) formulating a hypothesis as to how differences on the construct should be associated with differences on other characteristics
c) measuring the construct
d) gathering empirical data on the hypothesised characteristics
e) determining consistency between the construct levels and the characteristic levels.

If the hypothesised relationship(s) are found as predicted, then both the construct and the test that measures it are useful.

Details of examinations that might be carried out to assess construct validity are as follows.

3.1 Correlations between the test and other tests

Correlations between a new test and similar tests can be cited as evidence that the same general area of behaviour is being assessed. Anastasi (1990) points out that such correlations should be moderately high but not too high, for otherwise the new test represents needless duplication (unless it is eg. briefer, or easier to administer).

3.2 Correlation between the test and selected variables

The construct measure may be used to see if individuals or populations hypothesised to differ on the construct do so. If expected differences are found, the test is supported. If not there may be a fault in the theory underlying the construct, the measure of the construct, or the treatment assumed to provide the difference.

In essence measures that fail as criterion measures of validity are input as proxy-criteria. For example Donovan et al. (1989) in reviewing QOL-cancer scales argued that the best form of validation was assessment of the tests' capacity to predict external criterions such as medical or psychological stress indicators, eg. number of requests for medical help, use of psychiatric services etc.

3.3 Convergent and Discriminant validation

A framework within which to conduct construct validation was proposed by Campbell and colleagues, who pointed out that in order to demonstrate construct validity it should be shown not only that a test correlates highly with other variables it would be theoretically expected to, but also that it does not correlate highly with variables with which it would be expected to differ (Campbell & Fiske 1959; also see Anastasi 1990). Campbell and Fiske (1959) proposed a systematic method for exploring this, the Multitrait-Multimethod Matrix method, which entails the assessment of two or more constructs by two or more methods.

Campbell and Fiske's method proceeds from the notion that each test or task employed for measurement purposes is a `trait-method unit', a union of a particular trait content with measurement procedures not specific to the trait being measured. To examine discriminant validity, ie. that the test does not correlate highly with other tests that it should differ from, it is proposed that the researcher must identify two or more ways of measuring the construct of interest, and at least one further construct which can be measures by the same methods.

Using one sample of subjects, measurements are made on each construct by each method, and correlations computed between each pair of measurements. The matrix of all of the intercorrelations is the multitrait-multimethod matrix (or MM Matrix).

Three types of correlation are formed:

- Reliabilities (or monotrait, monomethod).

- Convergent validity coefficients - correlations between measures of the same construct using different measurement methods (monotrait heteromethod).

- Discriminant validity coefficients - correlations between measures of different constructs using different methods (heterotrait monomethod), or correlations between measures of different constructs using different methods (heterotrait heteromethod).

For satisfactory construct validity, the scores obtained for the same trait by different methods *(validity coefficients),* should be higher than the correlations between different traits measured by different methods, and the correlations measured between different traits using the same method (if the latter is high, a person's scores may be being affected by an irrelevant common factor such as ability to understand the questions; Anastasi 1990).

Campbell and Fiske argue that a careful examination of the MM Matrix will indicate what the next steps should be: whether methods should be discarded or replaced, or concepts sharpened in definition, and which concepts are poorly measured because of excessive or confounding method variance. They also caution that many M M Matrices will show no convergent validation: no relationship may be found between two methods of measuring a trait. In this situation alternative propositions are:

(a) neither method is adequate for measuring the trait;
(b) one of the two methods does not really measure the trait; or,
(c) the response tendencies are specific to the non-trait aspects of the test.

3.4 Construct representation

Embretson (1983, 1986, cited Anastasi 1990) has proposed a new approach to the assessment of construct validity arising from the process orientation of cognitive psychology, as opposed to the focus on outcome measures (correlation between test result and another measure) that arose from psychophysics. In Embretson's approach the study of *construct representation* is to `identify specific information processing elements and knowledge stores needed to perform the tasks set by the test' (Anastasi 1990, p. 160). Various task decomposition procedures and methods of experimental manipulation are advised so as to measure the contribution of different response components to test performance, and `to determine what theoretical constructs are assessed by the test'.

An approach of this type could well be useful when considering the output of some of the complex tasks demanded of subjects when developing health status assessment measures. As in the discussion of reliability earlier, it is difficult not to believe that the performance of many tasks is more concerned with a subject's capacity to conduct a complex cognitive manipulation. than the declared content of the task. Are the results of tradeoff tasks determined more by values about health, the ability to mentally juggle complex notions, or moral principles regarding the value of life.

3.5 Sensitivity/responsiveness

As noted by Donovan et al. (1989), a QOL measure needs to be able to discriminate between conditions and/or within conditions as the disease course changes, and Deyo and Patrick (1989) have suggested that a test should be tested for `responsiveness', meaning sensitivity to change, in addition to reliability and validity. It is equally possible to consider sensitivity as a special aspect of validity, which is the approach favoured in this review.

3.6 `Descriptive validity'

Bergner has proposed the term `descriptive validity' to refer to the ability of an instrument to comprehensively characterise a patient's health status. Bergner et al. (1981) proposed that item categories should be retained in the SIP even if they fail to account for additional variance, on the grounds that they contribute to the descriptive capacity of the SIP.

4 Techniques Used in the Measurement and Development of Validity

4.1 Correlation

Calculating the correlation between a test score and a criterion measure forms a *validity coefficient.* The Pearson Product-Moment correlation can be used for this, and as with correlation coefficients generally (and as with reliability coefficients), sample heterogeneity is a major factor: the wider the range of scores, the higher will be the correlation. Crocker and Algina (1986) suggest that whenever a low correlation coefficient is obtained, the researcher should determine whether a restriction in variance has occurred because of sample selection or some aspect of the measurement process obscuring a possible relationship between the variables of interest. Examination of the scatterplot is recommended.

Attention to the form of the relationship between test and criterion has other uses. The Pearson Product-Moment correlation assumes that the relationship is linear and uniform throughout the range, and it may be useful to examine the bivariate distribution via a scatter plot of to determine the relationship, eg. it may not be of equal variability throughout the range (ie. not homeoscedastic; see Shavelson 1988).

The validity coefficient may also be interpreted as the *standard error of estimate,* analogous to the error of measurement in connection with reliability. Even with a validity of .80. the error of predicted scores may be considerable, but a test may improve predictive efficiency if it shows any significant correlation with the criterion, however low. Anastasi (1990) suggests that even validities as low as .20 or .30 may justify inclusion, depending on the relative benefit from having the test.

In interpreting correlation coefficients the magnitude of the coefficient needs to be checked against the two criteria of whether it is significantly different from .00, and what percentage of variance in one variable is shared with variance in the other (see Hays 1981; Shavelson 1988). Remember also (as noted in the previous chapter) that high correlations between

variables does not mean that they are causally related, as some further intervening variable may affect both variables. Also restriction of variance in the scores of X or Y variables can reduce the maximum correlation value that can be obtained.

It is particularly important to be aware that high correlations can conceal important differences. Anderson, Bush and Berry (1986) compared dysfunction scores on the QWB scale using self and interviewer modes of administration, against a measure of dysfunction gained from applying the QWB to a detailed examination of ancillary clinical information. While finding a very high correlation between the two measures for all subjects (Pearson product moment r = 0.98), tests of sensitivity and specificity showed appreciable differences. Thus for those with actual dysfunction on the physical activity and social activity scales the self-administered QWB was reported to result in the accurate classification of only 45%, ie. sensitivity = .45 (note that the misclassifications may still have been of dysfunction, but not at the same level), while a sensitivity level of .86 was determined for the interviewer administered mode. Looking at the dysfunction subjects alone, the correlation between self and interviewer mode was .90. High correlations can mask major differences, and there is a need to look at specificity, sensitivity wherever possible (but criterion against which to check are not often available - here it was a carefully calculated QWB value, not dysfunction per se).

Bergner et al. (1976b) also produced data to suggest that an overall moderate or high correlation may mask shortcomings in sensitivity for particular sub groups. They found the correlation for a total subject sample between self-assessment of dysfunction and SIP score to be 0.52. This was more than each of several subgroups alone, with the correlations for one sub-group (speech pathology patients) being-.01.

4.2 Multiple regression

Multiple Regression, or MR, is increasingly being used in the development and validation of health status measures for the prediction of health status (eg. see Llewellyn et al. 1992; Lipscomb 1989; Hall et al. 1989). The degree of error likely in such prediction can again be estimated (for calculating the Standard Error of the Estimate to determine confidence intervals, see Crocker & Algina 1986).

MR is of particular use in health status tests because of the multi-dimensionality of the construct. Anastasi (1990) proposes that for the prediction of most constructs several (sub)tests are likely to be required. A single test designed to measure an appropriate criterion would have to be highly heterogeneous, and Anastasi recommends the combination of several relatively homogeneous tests rather than a single test consisting of many different sorts of items (on the grounds that tests, or subtests, should be homogeneous in the interest of reliability).

The combination of such subtests to allow a decision can be achieved by multiple regression. In the computation of multiple regression each test is weighted in direct proportion to its correlation with the criterion, and in inverse proportion to its correlations

with the other tests. Thus the highest weight, will be assigned to the test with the highest validity and the least amount of overlap with the rest of the battery.

These weights are optimal only for the particular sample in which they were derived, and it is important that the test algorithm is cross-validated by correlating the predicted criterion scores with the actual criterion scores in a new sample (although formulas are available to estimate shrinkage in a multiple correlation, the larger the original sample the smaller the shrinkage, empirical verification is considered preferable).

Anastasi (1990) also points out that sometimes a negatively correlated variable is needed to obtain the best correlation, due to the need to eliminate some influential variable that is uncorrelated with the criterion but would otherwise introduce irrelevant variance into the test. For example reading comprehension may correlate highly with scores on a test because the test problems require the ability to understand the instructions. Inserting a measure of reading comprehension in the regression equation will eliminate this error variance and raise the validity of the battery (although it is better to redesign the test to eliminate the undesired variable).

As described by Cohen (1968), multiple regression (MR) and analysis of Variance/analysis of covariance (AV/ACV) are essentially identical systems. In fact MR was developed in the course of the study of natural variation, while AV/ACV came out of artificial or [experimentally manipulated variation, but they are both general linear models. Both are equally robust to violations of normality assumptions. In essence AV is a special simplified case of MR particularly suited to neat experimental layouts where qualitative treatments are manipulated in appropriate orthogonal relationships. It has far less flexibility than MR as it leads to the dichotomisation of variables (so they can be examined as treatments) with a consequent loss of information and associated statistical power. In contrast to the constraints of AV/ACV programs, the very general MR program can accommodate any given design by coding those independent variables of interest.

Experimentalists often criticise MR as an inferior statistic, particularly compared to Analysis of Variance (ANOVA). However Cohen (1968) argued that MR is far more powerful and flexible an analytic system. Dummy variables allow the coding of nominal scale data, subtle variables can be captured via contrast coding, and curvilinear relationships can be examined by means of a polynomial form in power terms so that non-linear regression can be represented within the linear multiple regression framework. The preference for ANOVA was proposed as in part reflecting the original non-availability of MR, because it requires the computation and inversion of a matrix of correlations (or sums of squares and products) among the independent variables, which require major computation for even few independent variables. With electronic data processing facilities, there is no longer a barrier.

However caution is needed when using MR. For example many independent variables can be readily generated, with associated loss of statistical, power (as df increases) and the need to be aware of type I error rates. For multiple comparisons, the significance level that

gives an appropriate overall error rate of alpha is approximated by alpha/n, where n is the number of simultaneous comparisons (Patrick, Bush & Chen 1973). However organisation of the independent variables, and step-wise admission to the analysis (testing for significant increases in R^2) can control this.

Anderson (1976) has also pointed out the caution that is necessary when using MR to test theoretical models. Anderson describes how assuming a linear model could lead to high correlations of the order of 0.98, even when plotting the independent variables reveals curves that are strictly non-parallel, and as demonstrated by a significant ANOVA interaction term. Anderson argued that regression -correlation methodology can be useful in applied prediction, but can be misleading when it comes to testing theoretical models. Indeed the great usefulness of regression -correlation analysis in applied prediction stems largely from its insensitivity to real deviations from linear summation models.

Anderson (1976) also points out the fallacy that some researchers have engaged in of interpreting the importance of independent variables according to the magnitude of their factor loadings, as it is apparent that such correlations are confounded with a number of factors (such as the range of the relevant variable).

4.3 Factor analysis

Factor analysis began as an attempt by Spearman to examine the question of whether intelligence is the expression of a single major factor or whether there are multiple `intelligences' (Cattell 1972). Anastasi (1990) describes factor analysis as `particularly relevant to construct validation'.
The aim of factor analysis is to explain the correlations among a large number of variables as reflecting variation on a smaller number of underlying inferred factors, to go beyond appearances to basic concepts. Its role is both to generate hypotheses, and to test them, and proceeds from the supposition that many psychological attributes (traits) can be measured only by a whole pattern of variables and not any single variable.

Factor analysis (along with multiple regression; see Cohen 1968) are the major expressions of correlation analysis, and Cattell (1972) characterises factor analysis as the principal tool for examining the significance and magnitude of relations among variables when a large number of variables need to be examined simultaneously, just as ANOVA (Analysis of Variance) dominates analysis where variables can be manipulated under strictly controlled conditions.

Factor analysis involves obtaining n measures on the same testees, computing an n * n correlation matrix, and then using factor analysis techniques to identify a number of underlying variables (factors) that account for variation in the n variables.

The n measures may be either items, in which case it may be determined whether the items cluster together as predicted by the theoretical structure of the construct, or tests/measures (that may be made up of sets of items). Again the initial issue is whether

the subtests or tests which are supposed to measure the same element are identified as measuring a common factor.

Kaplan, Bush and Berry (1976) attacked factor analysis as a tool in the development of health status indexes. The reason proposed was that once factor analysis derived underlying factors (such as sociability, physical distress, etc.), then it was likely that items that are checked rarely or are poorly correlated with other items (and hence contributing little to explaining variance) would be considered unimportant and excluded. Kaplan et al. argue that such items could correspond to eg. rare conditions, which while not loading significantly on any of the larger factors and because rarely selected not representing a substantial unique factor, were none the less extremely important for the proper assessment of those rare cases. Hence infrequently used items should not be excluded.

The issue here seems to bear on the distinction between construct exploration and data reduction. Infrequent items should not be excluded solely on factor analysis grounds, for as Anastasi (1986) has emphasised, both logic and empiricism must play roles in developing construct measures. On the other hand Hall et al. (1989) have demonstrated the usefulness of factor analysis in health status instrument assessment (for more details, see Chapter F).

5 Conclusions and Observations: The Case for Concept Validity

Anastasi (1986, 1990) has argued that content-, criterion-, and construct-related validation no longer correspond to meaningfully distinct categories, but are products of the developmental history of validation testing. In Anastasi's framework, statistical methodology leads first to the analysing of items against total test scores or external criterion measures, and then to factor analysis, and so on, and construct validity should now be seen as a comprehensive concept that includes all the other types. `Content validation and criteria-related validation can be more appropriately regarded as stages in the construct validation of all tests'. Test scores are seen to be always based on constructs, with even in a simple test the factor being measured not corresponding to any single empirical measure. (For example, a test to measure an individuals walking speed there would be a need to take representative measurements to obtain a distribution of speeds depending on context, purpose, persons condition at the time etc.). After Messick (1980, cited Anastasi 1990), Anastasi argues that the term validity, insofar as it refers to the interpretive meaningfulness of a test, should be reserved for construct validity. Content validity should be labelled content-relevance and content-coverage, and Criterion-related validity termed predictive utility and diagnostic utility (corresponding to predictive and concurrent validation).

In terms of the test construction process, validity is seen to be built in to a test from the outset, rather than being limited to the last stage of test development. Almost any information gathered in the process of developing a test is related to validity. It begins with the formation of the construct definition, derived from psychological theory or prior research, then follows item preparation and analyses to select the most valid items, followed by internal analyses that may include factor analysis of item cluster or subtests (an

item needs to be shown to belong in a scale based on both logic, the construct definition, and through the results of factor analysis or other procedures of item analysis).

Finally, and most importantly, there should be correlation of scores with external real-life criteria. This needs to be combined with *very* close examination of the results of such correlation analyses, to ensure that an overall high correlations is not masking a major failure of the instrument for important sub-groups. This is particularly relevant when validating generic health status measures, given the range of patient types that they are required to handle, and the evidence of Bergner and Anderson et al. discussed earlier (section 4.1).

F SPECIFIC ISSUES WHEN CONSTRUCTING HEALTH-RELATED QOL MEASURES

It was argued in Chapter B that the measurement of health-related quality of life (HQOL) is best achieved via an instrument that assesses the dimensions found to affect HQOL and combines them into a single index. The process of forming such an instrument involves many steps, many of which are general to all tests (item analyses, subject analyses, etc. see O'Connor 1991, 1992a). The following comments are restricted to elements that deserve special mention in the context of HQOL measurement.

1 The Importance of Clearly Defining the Purpose to which the HQOL Test will be put and the QOL Concept to be Used

A clear conceptualisation of HQOL is essential. Earlier it was argued that Subjective Well-being was an appropriate notion to be used. The perspective inherent in this is that it is the *patient's experience* of the health state (current and anticipated) that should be the basis for determining the benefits or otherwise to be gained from a given health program.

The purpose of the HQOL test also needs to be detailed, for example: `Within a period of five years and subsequently on a comprehensive annual basis, to aid decisions regarding the allocation of state health funding between all health programs, at a level which would allow decisions regarding the relative funding of hospital-based individual clinical specialties, community-based and domiciliary services, and health promotion/disease prevention.' Such specification of purpose is necessary to ensure that the instrument is capable of performing the task demanded of it. Thus if the purpose is as above, then the instrument may need to be based on data that is readily and routinely accessible through existing information systems.

2 The Structure and Outputs Required of the HQOL Instrument

As discussed in Chapter B, it is desirable that an instrument for measuring HQOL assesses a number of dimensions: a measure which is derived solely from a single global self-report of QOL is likely to be inaccurate and unreliable, given the multi-dimensional nature of QOL and the role of adaptation.

As well as possessing a sub-scale structure it is desirable that the test outputs the sub-scale values (in addition to a global measure). The presence of sub-scales in the test instrument makes explicit which dimensions/areas are being assessed, and allows measures of dimensions to be validated individually.

The presence of sub-scales also allows the instrument to play a diagnostic role. As noted by Revicki (1989), in many situations there may be no clear advantage for one therapy compared with another when the therapies are assessed globally. However there could be differences in their profiles of effect, eg. some dimensions improving (such as functional ability) and others becoming worse (eg. more symptoms). A single global score would not

reveal this, hence it is important to be able to examine subscale/domain scores (see also Deyo & Patrick 1989; Goodinson & Singleton 1989; Donovan et al. 1989).

It is also of considerable value to pre-specify the criteria that a particular health status instrument will require to meet if it is to be considered valid. These criteria should concern steps in the development of the instrument, as well as the final properties of the completed instrument. Such a set of criteria developed for a work-disability measuring instrument is presented as illustration in **Appendix 1.**

3 Selecting the Task Used to Develop and Scale the Test

Issues concerning the selection of the most appropriate task for gathering assessments of health states when developing a test were discussed in Chapter C. Other than rating methods, psychologists have traditionally used a range of formats when developing attitude and personality inventories, including the Thurstone method of equal-appearing intervals, Likert, and Guttman (eg. see Moser & Kalton 1979).

4 Determining the Content of the Test

As noted in Chapters B and E, while developing and validating a QOL measure is continuing and incremental, emphasis needs to be placed in the first instance on ensuring that the instrument adequately assesses all the areas of the patients life that are significantly affected by the condition or program (ie. the test possesses content validity, or in Anastasi's terms, content relevance and content coverage). Bergner (1989) states there is a need to identify the domains or categories of variables that are relevant before starting to assess the appropriateness of existing measures, and that in general assessment should examine factors that are likely to be affected by the intervention, or have been troubling to patients in the past. Donovan et al. (1989) noted that any failure to assess all relevant dimensions may lead to an inability to detect the impact of a treatment on QOL or to record no difference where one exists. Donovan also provided analyses to suggest that many current QOL measurement instruments are likely to fail tests of content validity: of 17 scales concerned with the assessment of the QOL of cancer patients, only 2 were judged to have weighted items to adequately represent the importance of social and psychological areas, despite a background of reports indicating both major negative and positive effects in these areas. Donovan et al. cited a 1982 study by Peters-Golden indicating that 72% of a group of cancer patients said that people treated them differently after they knew their diagnosis (with a perception of being misunderstood and avoided), as well as some reports of positive affects (eg. increased closeness to spouse).

Donovan et al. further suggested that the sensitivity of instruments to effects in different domains can be roughly indexed by the number of questions asked per area and/or the number of response categories per question. For example, the Spitzer instrument was seen to fail this test as it only had 5 items in total, and 3 options per item, and could not discriminate between individuals who were seriously ill (eg. between those with late stage cancer and other conditions), or between those who were without physical symptoms.

4.1 Forming content materials

The process of developing appropriate content may variously entail:

(a) Content analysis, where open ended questions are posed to subjects about the constructs of interest, and their responses are sorted into topical categories. Those topics that occur predominantly may be taken as major components of the construct.

Moser and Kalton (1979) note that statements made in unstructured interviews can be a valuable source of items for an item pool, and they have the advantage of being expressed in the everyday language of the respondents themselves as opposed to being contrived by the researcher. Haig et al. (1989) asked subjects to describe in narrative form there most recent and most severe illness, and in addition to write upon the types of discomfort or unpleasant sensations they experienced during the course of a day for one particular illness (to bring discomfort formally into the illness equation).

(b) Review of research, ie. the examination of reports of relevant investigations.

(c) Direct observation of patients and their circumstances, and the examination of their clinical records.

(d) Expert judgement, ie. obtaining input from those with first hand professional experience of the effects of illness.

(e) The checking of candidate test materials and their contained dimensions to ensure they are meaningful and relevant, eg. ask patients which items/materials/descriptions seem most relevant (Donovan et al. 1989, reported finding no studies that had examined the perceived relevance of the items for any of the scales included in their review).

4.2 Who should provide the health state assessments

While various sources may be consulted to determine the range of factors that are relevant to comprehensively assessing health state, it is essential (as discussed in Chapter B) that candidate areas be assessed to determine how they should be combined to form a single index. This entails the issue of who should supply global assessments of health state. In Chapter B it was argued that patient report was the primary source of data, although adaptation may require that the judgements of others also be included.

It is clear that the description and assessment of a given health state may differ according to whether the patients, health professionals, or the general public make the appraisal. Differences exist both in terms of overall level of valuation (there is some evidence of a

tendency for the general public to appraise a given illness state more negatively), and in the relative emphasis placed on different aspects of the condition.

For example it is apparent that the assessments of patients and doctors differ markedly. Skeel (1989) cites a 1982 study by Jachuck et al. examining the effects of anti-hypertensive therapy. It was found that although 100% of physicians thought the patients QOL had improved following the beginning of therapy, only 44% of the patients themselves reported feeling improved, 44% felt no change, and 8% felt worse. At the same time assessment by relatives was that QOL had decreased for more than 90% of the patients.

Hall, Epstein and McNeil (1989) also have reported substantial differences between patients' appraisals and that of their doctors. When the patients were asked to give global ratings of their own health they drew on aspects relating to emotional. functional, and physiologic health. In contrast, their physicians ratings appeared to be primarily based on physiologic health, with relatively little regard for emotional health. This emphasis by doctors upon physical much more than psycho-social aspects was previously reported by Martin et al. (1976) when examining the Sickness Impact Profile.

The same studies suggest that the separation between doctor and patient judgements increase as the doctor becomes more senior. Estimates made by first year residents, second and third year residents, and staff physicians had steadily decreasing correlations with patients total SIP scores, which Hall et ai. interpreted to reflect a steady narrowing of the physician's perspective on the patient's health (see also Bergner et al. 1976).

Patients also differ from the uninformed public. It is apparent that the experience of the health state causes a change in its assessment. In a study examining the attitudes of 18 pregnant women towards anaesthesia during child birth, Christensen-Szalanski (1984, cited Tsevat et al. 1990) found that experienced mothers differed from first-time mothers in feeling more strongly about avoiding anaesthesia during as well as outside of the delivery period. Tsevat et al. (1990) also note that patients with end-stage renal disease value life more than the public believes.

Kind and Rosser (1988) reported that ill subjects (medical and psychiatric patients) and nurses tended to give higher magnitude valuations for severe illness states than did healthy subjects, and Sackett and Torrance (1978) found valuations of health states differed between healthy volunteers and patients. Carter et al. (1976) also reported that non-health workers tended to give higher ratings of dysfunction to illness behaviour statements than did health professionals and students.

It is evident that patients differ from both doctors and the general public in their assessments of illness state, and it has been claimed that the development of QOL tools upon members of the general public makes the instruments potentially irrelevant to patient populations. This criticism has applied both to the use of aspects of condition derived from studies of normal people, and the relative emphasis (or weights) placed on the aspects.

Donovan et al. (1989) concluded that a 1982 study by Irwin et al. of the QOL of cancer survivors was of little value as it used an instrument that had little relevance in the context of cancer, the instrument being designed to measure QOL in healthy populations. Deyo and Patrick (1989) expressed concern regarding the fact that many health status questionnaires employ a `weighting system' that assigns values to dysfunction based on the rankings of lay persons, noting that clinicians may be justifiably concerned that such standard weighting schemes may not reflect a particular patient's values. It is interesting to note that Kaplan and Anderson note criticism of the Quality of Well Being (QWB) scale for being developed on the basis of community rather than specific-population weights, and in part defend this aspect by stating that Balaban et al. (1986) found that weights obtained with arthritis sufferers to be `remarkably similar to those we obtained from members of the general population' (Kaplan & Anderson 1988).

It may be pointed out that there are those who believe that while patients should be the source of information regarding which aspects of life to measure, the general population may be employed to provide estimates when determining the relative weight to be placed on each aspect. While this issue is a matter of debate, it would be difficult to make a case that someone other the patient be accepted as the primary repository of information about the effects of any condition, particularly in the light of evidence that patients emphasise elements of their condition differently to others. As argued by Donovan et al. (1989), if it is accepted that QOL is an individual's subjective sense of well-being, then it is necessarily the result of personal perception of circumstances and any scale assessing QOL must be constructed using information from patients themselves. Otherwise there is the risk of omitting information of central importance. At the very least, patient information should play a central role in the formation of any QOL instrument.

On the other hand the evidence discussed in Chapter B on adaptation questions the wisdom of relying solely upon patient report, given the as yet poorly understood role and nature of adaptation. Health care givers, be they professionals or voluntary, know to a greater or lesser extent the impact of factors related to health, while those intimate to the patient may know more fully of other factors and their inter-relationships. Professional care givers can also have a broader knowledge of the comparative effects of different health states.

A related issue concerns who should complete tests that concern individual patient assessment. Donovan et al. (1989) argue that such measures must be completed by patients, and not observers, for only the patient is fully aware of their own condition etc. On the other hand this is an empirical issue, and if it could be demonstrated that a test was a valid predictor of patient report then the patient's involvement would not be necessary.

As noted in Chapter B, there are many uncertainties yet to be resolved in this area. O'Connor (1992b) may be consulted for further discussion.

4.3 Which dimensions should be assessed for a comprehensive HQOL test?

Revicki (1989) refers to physical functioning, psychological functioning, social functioning, cognitive functioning, and general well-being. Donovan et al. (1989) proposed physical, psychological, social, and spiritual. Goodinson and Singleton (1989) suggest that research has shown the most important factors in QOL to be social support, family or marriage. Haig et al. (1986, 1989) noted that discomfort (pain, nausea etc.) was not centrally represented in the QWB, and endeavoured to develop a generic discomfort sub-scale (manipulating quality, intensity and duration). Skeel (1989) suggested that elements that patients, families, and health care providers (eg. oncologists) commonly accept as important to a patients QOL experience are functional capacity, self-perception of wellness or its absence, and symptoms of disease or treatment.

It is apparent that there has been considerable speculation about which aspects of life are most affected by illness. It might even be necessary to consider examining the role/interactions of variables such as accommodation type, eg. there is some evidence that hospital stay can be associated with higher mortality rates for old people, and domiciliary support programs may remove the need for hospital admission. On the other hand assessment of the benefits of domiciliary programs possibly should also consider the QOL of other family members, ie. those who may have to provide the major care for such chronically ill home-based patients (Gudex 1986).

Which domains are most important may also vary with age and circumstances. Cutler (1969, cited Diener 1984) found that the structure of domain satisfaction (satisfaction with work vs. marriage vs. clothes, etc.) varied for different age groups, and generally the domains that were closest and most immediate to persons were those that most influenced SWB.

While many have expressed opinions in this area, few have conducted detailed research. However studies that have systematically examined the importance of different aspects of quality of life have produced interesting results. A study worth recounting in some detail is noted earlier, ie. that of Hall, Epstein and McNeil (1989). Hall et al. explored the perceptions of elderly people to determine the relationship among various domains of health, in the sense of health being perceived as a multidimensional construct, and the degree to which patient perceptions of their own health corresponded to the perceptions of their physicians. This investigation was cast in the context of examining the construct validity of measures of health, ie. the degree to which any one instrument measures the desired trait.

To carry out this exploration Hall et al. interviewed 590 patients to gather information on 13 different variables, most assessed by multiple items. Four variables concerned functional status, three emotional experience ie. depression, anxiety, and overall well-being, two social activity, one cognitive function and one self-perceived health. They also gathered patient-specific physician ratings on current physical and mental health, and likelihood of deterioration. Finally they examined the medical records of each patient interviewed to determine the number of diagnoses per patient of 24 different medical conditions, this

indexing a variable termed medical complexity.

Scales producing data with non normal distributions were treated with log or arsine transformations. This data was then input to factor analysis, using both principal component analysis to assess the strength of one common underlying dimension, and rotated solutions to determine the presence and nature of a number of independent dimensions. The 13 variables correlated with the first principal components factor to account for 35% of the variance, this seen to indicate a common underlying dimension of health. This was not however equally reflected in each variable, eg. family social activity correlated only .18 with the first principal component factor.

Rotated solutions with three factors were then developed, leading to the four functional health scales and the medical complexity scale all loading on one factor, while the three emotional experience scales loaded on a separate factor. Self-perceived health loaded almost equally on both these factors, while in contrast the physician rating scale loaded only on the first factor of functional health and medical complexity. The third factor consisted of family social activity and cognitive status only. A six factor rotated solution produced distinct factors relating to functional status, emotional status, physiologic status, family social activity, non-family social activity, and cognitive status, with only patients self-ratings of health loading on multiple factors, namely on functional status, emotional status, and physiologic status.

It is also of interest to note that the original table of correlations between variables revealed that patients rating of well-being with other variables fell roughly into four categories, highest correlations being with the depression and anxiety variables (.65), followed by self-perceived health (.54). Functional health variables and non-family social activity correlated .31 or greater, with the remaining four variables of physician health rating, medical complexity, family social activity, and cognitive status correlated .25 or less. Such values are of course confounded by all the other variables.

While all values should be interpreted carefully because of possible confoundings between variables, the data suggests that `social support of the non-family kind' may be a variable of the order of importance of functional health as an indicator of subjective well-being. Moreover since restricted social activity can be both a consequence of illness conditions and (possibly) play a major role in determining subjective well-being, variables such `non-family social support' may be necessary elements of tests estimating the effects of health on quality of life.

There is a need for basic research into the dimensions relevant to quality of life under different patient and illness conditions, and as a first step it would be beneficial to systematically explore the aspects of quality of life that patients perceive to be most relevant at different stages of their illness(es). Such information would be of immediate relevance to those forms of QOL investigation that entail the preparation of patient scenarios (descriptions of the condition and circumstances of a patient or condition). QOL investigations frequently use scenarios to obtain ratings of illness conditions on a scale of

desirability/undesirability, as a step in the development of QOL scales. A scenario formed to represent a given condition as experienced by the patient is likely to differ markedly from a scenario of the same condition formed without first hand knowledge of the subjective experience of the condition, with likely major effects on any ratings elicited and the QOL instrument that may be developed from those ratings.

In summary, there is a need for the systematic exploration of the range of domains for a given condition, the relative importance of different domains, and the identification of any interactions between domains. As discussed in Chapter B, there is also a need to identify variables that may predict conditions under which psychological adaptation will fail and lead to diminished well-being. Such understanding may be a prerequisite for the development of valid QOL measuring instruments.

5 The Treatment of Future Events and Mortality

Kaplan and Anderson (1989b) note that most health status assessments are essentially morbidity indicators (eg. SIP, RAND), and argue that the benefits of medical care, behavioural interventions, and preventive programs need to be expressed in terms of output in years of life adjusted by the quality of life which has been lost because of disease or disability (ie. well years, quality adjusted life years, QALY's, discounted well years, etc.) if there respective merits are to be usefully evaluated.

This raises the issue of how to evaluate future health status from the present perspective. As noted by Kaplan et al. (1989, but see Lipscomb 1989), it is important not to neglect what will happen in the future. Many health programs affect the probability of occurrence of future dysfunction rather than altering current status. A person who is very functional and asymptomatic today may harbour a disease giving him a poor prognosis eg. many individuals are at high risk of dying from heart disease due to poor diet etc. Should these people be called `healthy'? Kaplan points out that the term `severity of illness' needs to take into account both dysfunction and prognosis, and health needs to include current and future components. As noted by Goodinson and Singleton, the use of satisfaction in defining QOL raises the issue of how to treat future satisfactions in relation to present ones. Is it rational to discount these, or should they be adjusted using some alternative model.

Consciousness of the developing/possible future condition may complicate this analysis. Such events may not be apparent to the person involved. As noted by Bergner (1989), external factors may contribute to diminished quality of life (at least in the future) even if the individual is not aware of it. For example individuals could be considered to be experiencing a lesser quality of life by virtue of living in an area of major environmental pollution, with QOL seen to be diminished even if the effects of the pollution are not perceived or realised by those resident in the area. Perhaps Kaplan and Anderson (1988b) are intending to incorporate such `invisible to the individual' factors in their symptom problem complex of `breathing polluted air' (see Chapter G). One way of dealing with such factors is to assume QOL state A for length of time A', then QOL state B for length of time B', etc. the sum of

these fractions providing total well-years, ie. the effects of pollution would be counted when they affect perceived QOL and/or length of life. The calculation of values is complicated however by the fact that consciousness of likely future ill-health is likely to affect current QOL: for many individuals plans and aspirations which make a pivotal contribution to QOL would not be worth undertaking without a good chance of their fulfilment (Glover 1977, cited Goodinson & Singleton 1989).

6 Method of Test Administration

Test administration is important both from the perspective of ease of use/practicality of developing and applying the instrument, and the validity of the test itself

6.1 The need for practicality

Donovan et al. (1989) refer to the need for measurement tasks to be capable of being readily understood, and administered in a reasonable time, and note that a question formed where there are a limited number of fixed alternatives tends to be easier to administer and score than an instrument employing continuous or visual analogue scales (but see Chapter C). Deyo and Patrick (1989) suggest that the most usual problem is the length of the instrument placing too great a burden on the respondent, and for older or disabled persons administration time can be much greater than in younger populations. This can be a major issue when used in busy clinics or when population based surveys are conducted. Revicki (1989) also suggests that measures should be judged on the basis of mode of administration (respondent burden, clarity, self-report versus interview, etc.).

6.2 Self-administration can produce less valid measures

Anderson, Bush and Berry (1986) compared QWB dysfunction scores using self and interviewer modes of administration against a measure of dysfunction gained from applying the QWB to a detailed examination of ancillary clinical information. For those with actual dysfunction on the physical activity and social activity scales, the self-administered QWB was reported to result in the accurate classification of only 45% of cases. This was contrasted to 86% being accurately classified with the interviewer administered mode. Anderson et al. suggested that self administration leads to a substantial and persistent problem of `false function' errors, i. failing to detect actually dysfunctional persons and erroneously classifying them as fully functional.

That self-administration may lead to less accurate measures is also suggested by work of Bergner et al. (1981). Bergner et al. also found appreciable effects of test administration mode when considering the reliability of three types of administration:

a) interviewer-administered;
b) interviewer delivered and explained, and then self-administered; and
c) mail delivered and selfadministered.

Bergner et al. found that mail delivered, self administered SIPs had the lowest internal

consistency reliability (measured by Cronbach's Alpha), and also the lowest correlations with self assessed dysfunction and illness (0.48 and 0.38 correspondingly). Interviewer delivered/ self-administered produced the greatest correlations with self assessment dysfunction and illness scores, these being 0.74 and 0.67 correspondingly. Interviewer administered were in the middle, with correlations of 0.64 and 0.55.

7 Interpretation of Test Scores

With health status measurement instruments possessing potentially great policy and clinical application, the implications of differences in test score need to be judged carefully. Independently of the issue of scale-type (see Chapter C), there is the general matter of how to interpret the size of a difference. Testing for the statistical significance of the effect does not resolve the issue, as eg. a large F can result from a numerically large treatment effect, or large sample size, or both.

In considering this issue Kazis et al. (1989) suggested that differences in mean scores (from baseline to test) should be related to the standard deviation of the scores at the baseline (see also Guyat et al. 1987). Deyo and Patrick (1989) suggest that such an index could be used to calculate a `responsiveness coefficient', which would indicate the smallest clinically important score change for an instrument (eg. questionnaire) which is unclear for many health status instruments (although this could vary among diseases, populations, interventions etc. and ultimately requires subjective judgements).

Keppel (1982) may be referred to for further techniques for assessing relative treatment magnitude. In particular, Omega Squared is proposed, which provides a measure of the proportional amount of the total population variance accounted for by the experimental treatment. Anastasi (1990) may also be consulted regarding the matter of how to interpret differences in test scores.

G SOME CURRENT TOOLS

The following briefly reviews the development of several of the more prominent health status assessment tools. Some of the information provided has already been mentioned in preceding Chapters. Where this occurs it has been presented again so as to more clearly understand the tool under discussion.

1 Bergner's Sickness Impact Profile (SIP

The Sickness Impact Profile (SIP) was originally developed by Bergner and coworkers (eg. see Bergner et al. 1976a, 1976b) as part of funding from the U.S. National Centre for Health Service Research to develop general measures of health status. The SIP, QWB, and the General Health Rating Index later developed by the RAND Corporation to become the RAND Health Status Measure, were all results of this initiative. Kaplan et al. (1989) has noted that each of these measures was `guided by the WHO definition of health status: "Health is a complete state of physical, mental and social well-being and not merely absence of disease"' (Kaplan et al. 1989, p. S31).

The SIP was developed with the specific aim of providing information on the efficacy of health programs to assist decision regarding the appropriate allocation of the government's resources. It was aimed to provide a `fiscally and logistically practically measure of health status'. (Bergner et al. 1976a, p. 393).

In developing their measure, Bergner et al. referred to three broad conceptions of health on which individuals base their appraisal of their own health status (citing Barmann, 1961). These were a feeling state conception, a clinical conception, and a performance conception. Feeling state refers to statements of the type, `I feel good', or `I don't feel well today'. Clinical conception refers to specific symptoms, and performance conception refers to activities that can or cannot be performed.

Bergner et al. decided that only the last of these, the performance conception, was suitable. It could be based on respondent report, but could also be easily observed and reported by an untrained observer, and also allowed easy comparison between different diseases and dysfunctions. The feeling state conception was ruled out as it seen to be inaccessible to external validation, and the clinical conception was seen as unsuitable as it required medical interpretation and hence was reliant on the definitions of physicians and not the person concerned.

The SIP was hence conceptualised as *`an instrument which would provide a descriptive profile of the responses of a given individual in terms of the specific behavioural impacts of sickness'* (Bergner et al. 1976, p. 401), with the impacts capable of being summarised within specific areas of living as well as in some form of overall assessment. While the respondent can be someone other than a subject, the usual respondent if the subject himself. The statements were also first person statements reported in simple terms. As stated by Bergner, the items conformed to the criteria for attitude scales.

1.1 Initial development

(a) Item preparation

The instrument was developed by first preparing a list of health-related dysfunctions that would cover a range from minimal to maximal dysfunction. These were obtained from patients, health care professionals, individuals caring for patients, and the apparently healthy. From this process was first obtained 1100 statements describing a dysfunction and the corresponding behaviour. Through a process of refinement the SIP instrument was reduced to 312 items in 14 categories.

The items were present tense, simple, first person statements of the type that appear in attitude scales. The categories included areas such as social interacting eg. `I am going out less to visit people', mobility and confinement eg. `I stay within my room', emotion, feelings, and sensations eg. `I isolate myself as much as I can from the rest of my family'. The items were to be administered by a trained interviewer, and the respondent was to respond only to items that he/she was sure described him/herself on the day, and were related to his/her health

(b) Single item scaling

The items were scaled by asking judges to allocate each item within a category to an 11 point scale, ranging from minimally to severely dysfunctional. The most and least dysfunctional items from each category were then allocated to a 15 point scale, with the intra-category items then re-scaled mathematically to ensure consistently across categories.

(c) Item and judge reliability-testing

The judges were 7 nursing students, 8 medical students, 4 physicians, and 6 health administration students. T-tests showed no significant differences between the clinically sophisticated and unsophisticated types of judges in terms of mean scale values measured per item.

Items shown to be scaled by different judges such that their 95% confidence intervals were more than 2 scale points apart were deleted (29 of the initial 312 items).

(d) Item combinations

The development steps described so far were concerned to scale individual items. However patients would normally possess numbers of items (item profiles) and to adjust for item interactions 246 patients were assessed for the applicability of each item. Further groups of judges were then used to make global ratings of dysfunction of each subject's protocol of responses in each category of the SIP, and then to rate each subject's complete SIP on a

15 point scale.

Four methods of scoring the SIP profiles were considered, including a mean of the scale values used, a mean of the squared values (ie. weights high scoring items), percent of total possible dysfunction (total of scored items/total possible score X 100), and a profile, the number of items checked in each of 4 groupings, where groupings were determined by the values of the items. Thus items of value 15-11 were in group 1, 10-7 in group 2, 6-5 in group 3, 4-1 in group 4. A profile of 1430 would mean 0 scores between 1-4, 3 scores between 5-6, 4 of between 7-10, and 1 between 11-15. These profile scores were regarded as integers.

Correlations between scoring methods and global ratings were conducted to determine which were most suitable. The results were that the profile and percent methods gave the highest correlations with the global assessments, and Bergner et al. suggested that the protocol judges appeared to attend to both the number of items checked and the items checked with the highest value (although they did not know the item scale values).

Further work was carried out to validate the instrument, eg. using multiple regression analyses to look at interrelationships between items when used to describe subjects. Regression analyses were conducted separately for groups of subjects whose SIPs were rated as `high dysfunction' and those rated as `low dysfunction'. These analyses were made to ensure that items which were useful predictors for several or minimally dysfunctional subjects, but which were not useful predictors for the sample as a whole, were retained in the revised SIP.

In conclusion Bergner et al. noted that the SIP was designed to assess health-related dysfunction, as opposed to QOL and was not intended to be used as a sole criterion for either evaluating health programs or assessing population health levels.

1.2 1974 field validation

Bergner et al. (1976b) reported the field validation of a 1974 version of the SIP, using a 235 item version of the SIP.

This and later versions employed a scaling system where a SIP per cent score was calculated by summing the scale values of items checked and dividing by the sum of the scale values for all items and multiplying by 100. Such scores may be calculated for the whole SIP (all items) and individual categories.

To validate the SIP against a field population, 278 subjects were gathered in four sub samples: rehabilitation medicine patients, speech pathology patients, outpatients with chronic problems, and group practice attenders. The SIP scores on these patients were compared to subjects self-assessments of both sickness and general dysfunction, clinicians assessment for the rehabilitation and speech pathology patients, and scores on the functional assessment instruments of ADL (activities of daily living scale) and NHIS (a

gross estimate of functional restriction in the previous two weeks).

It should be noted that the self-assessment instructions differed by more than concerning sickness/injury or poor functioning: the function instructions were much longer, more detailed, described the way in which functioning could be diminished, and as a product were also less ambiguous.

(a) Discriminant capacity

The four sub sample groups were analysed using ANOVA, revealing a significant main effect of group, ie. patient type influenced SIP score as would be expected given the differences between the groups.

(b) Self-assessment of health status

Although for the whole sample the correlation between self assessment scores and SIP scores was moderately high and significant (sickness r=0.54; dysfunction =0.52), their was a trend for correlations to be higher for some sub-samples. Thus group practice patients correlation coefficients of r=0.74 and 0.45, chronic patients r=0.37 and 0.42, and speech pathology patients r=0.21 and -0.1 (non significant). Those SIP scores that best predicted the self assessments were determined using multiple regression, revealing that **four of the** categories (ambulation, mobility and confinement, body movement, leisure past times) were most highly related to self assessment scores.

(c) Clinical assessments

The SIP score for chronic patients correlated 0.49 ($p<.001$) with clinical rating, although it was noteworthy that the strength of the correlation declined with the length of time in practice. For these physicians ambulation, mobility and confinement, and sleep and rest behaviour were the factors most related to their dysfunction rating. The SIP score for speech pathology patients correlated non-significantly with the ratings of their physicians

(d) Relationships between self-assessment and clinician assessment

The correlation between chronic patients self-assessments and their clinicians assessment was moderately high and significant ($r=0.52$, $p<.001$). However the corresponding correlation for speech pathology patients was low and non significant.

This data is of interest as it shows again how an overall moderate or high correlation may mask shortcomings in sensitivity for particular sub groups. For example the correlation for the total sample between self-assessment of dysfunction and SIP score was 0.52. This was more than each of the subgroups alone, and with the correlations for speech pathology patients being -.01.

(e) Assessment based on ADL and NHIS

SIP correlated with the ADL score for rehabilitation patients 0.46 (p<.01), while the NHIS correlated with SIP for all patients 0.61 (p<.001). However, again, the aggregate score masked major lack of sensitivity for some groups. Thus rehabilitation patients correlated only 0.17 (nonsig.) and speech pathology patients 0.30 (nonsig.), versus chronic patients 0.52 (p<.001) and group practice patients 0.58 (p<.001).

This data demonstrates the importance for the validation of health status instruments to include examination of different patient populations, and to use a range of convergent and discriminant measures.

1.3 1976 field testing

As described by Bergner et al. (1981), the final form of the SIP followed a 1976 field testing, entailing study of a large structured random sample from a prepaid group practice, plus a quota of self-described sick people from a family medicine clinic. The validation resulted in the SIP being reduced to 136 statements in 12 areas of activity, with the final form of the instrument capable of being administered in 20 to 30 minutes, or of being self administered. It covers physical dysfunction, psychosocial dysfunction, and a number of independent areas such as eating, work sleep and rest etc.

(a) Administration mode

As part of the 1976 field validation, Bergner et al. investigated the reliability of three types of administration: interviewer administered; interviewer delivered and explained and then self-administered; and mail delivered and self administered. It was found that mail delivered, self administered SIPs had the lowest internal consistency reliability (measured by Cronbach's Alpha), and also the lowest correlations with self assessed dysfunction and illness. In fact the interviewer delivered and explained and then self-administered mode produced the greatest correlations with self assessment dysfunction and illness scores, these being 0.74 and 0.67 correspondingly, vs correlations of 0.64 and 0.55 for interviewer administered, and 0.48 and 0.38 for mail delivered.

(b) Convergent/discriminant validity

The validity of the SIP was again assessed, with tests of its ability to correlate highly with overall self assessment, and correlate less highly with overall clinicians assessment and other instruments.

SIP scores were found to be most correlated with self-assessment of dysfunction (0.69, followed by self-assessment of sickness (0.63), then NHIS index (0.55), and least correlated with clinicians assessment of illness (0.40). This pattern of diminishing correlations was seen to be consistent with the construct as planned.

Correlations were further examined within a model similar to the multitrait-multimethod

technique of Campbell and Fiske, and by multiple regression analyses. The latter analysis was found to demonstrate that SIP scores explained more of the variance in measures of dysfunction (eg. ADL) than measures of sickness (eg. self-assessment of sickness).

(c) Clinical validity

The SIP places considerable emphasis upon the item categories that make it up, which at the broadest level correspond to a dimension of physical functioning (Dimension 1: physical categories) versus a dimension of psychosocial functioning (Dimension 2: psychosocial categories). This aspect is seen to allow a demonstration of clinical validity by examining whether clinical conditions load as expected on the internal elements of the SIP.

An example is Hyperthyroidism. Bergner et al. (1981) hypothesised that for hyperthyroid patients, scores on dimension 2 (psychosocial categories) would be more highly correlated with thyroid hormone levels than would Dimension 1 scores (physical categories) or overall SIP scores. This was found to be partially supported, in that Dimension 2 scores correlated 0.35 with thyroid hormone levels versus 0.21 for Dimension 1 scores. However overall SIP levels were more highly correlated again, r=0.41. This was found largely to reflect the action of the independent category `sleep and rest', which was found to be highly correlated with thyroid level.

While the SIP could not be seen to have passed a test of clinical validity here this test did indicate its potential diagnostic value, ie. the capacity to pick up what may be unexpected effects of a disease. Temkin et al. (1989) have recently attested to the value of the SIP as a measure of specific disease conditions.

(d) Descriptive validity

Bergner et al. (1981) refer to the ability of the SIP to characterise the pattern of dysfunction specific to a given condition. They also state that it can be difficult or impossible to predict a priori which categories will be most important for a particular sample, and categories should be retained in the instrument even if they make no difference to accounted for variance on the grounds that they contribute to the descriptive capacity of the SIP.

In conclusion, Bergner et al. have subjected the categories and items of the SIP to considerable analysis to examine the relationships among items, between items and category scores, between items and criterion variables, the frequency of checking of items, and the reliability of items. In this pursuit a variety of multivariate techniques have been used, including stepwise multiple regression, interaction detection analysis, and cluster analysis (resulting in a reduction of the SIP to 12 categories and 136 items).

The SIP illustrates the process whereby the validation of a functional status instrument entails application of a range of convergent and discriminant measures, and the examination of different patient populations. The SIP has been closely examined, and its `descriptive validity' aspect may give it considerable value as a diagnostic instrument. On

the other hand it cannot be assumed as valid for all conditions (as eg. with speech pathology), and its method of combining items from different domains into an aggregate SIP score seems to rely upon the assumption that all domains have been sampled comprehensively and equally: content validity seems to have been assumed, not specifically tested.

2 Quality of Well-being (QWB) Scale

The Quality of Well-Being scale (QWB; the term Index of Well-Being or IWB may also be used) was originally developed by Bush and co-workers, later largely by Kaplan.

Kaplan et al. (1989) characterised the QWB as within the conceptual approach of the SIP and RAND scales, ie. focussed on the impact of disease and disability on function and observable behaviours, such as performance of social role, ability to move around the community, and physical functioning. While developed from a psychometric perspective, Kaplan saw the QWB as providing measures of utility, ie. the Quality of Well-Being (QWB) provided a numerical point-in-time expression of well-being that ranges from 0 for death, to 1.0 for optimal functioning.

2.1 Nature of the QWB

The instrument was developed by gathering global assessments of scenarios represented on paper cards, where each card contained an age level or AGE, a functional level or LEV, and a symptom-problem complex (CPX).

There were 4 levels of *age,* eg. `older adult (65 years and over)', while *functional level* was determined by a combination of three characteristics: mobility (5 levels, eg. drove car and used bus or train without help), physical activity (four levels, eg. walked with physical limitations), and social activity (five levels, eg. had help with self-care activities). While there were 100 possible function levels (5x4x5), only 42 of these were considered feasible and ultimately used. There were originally *36 Symptom-Problem Complexes* (CPXs), later grouped to 21, an example of one being `had pain, burning, bleeding itching, or other difficulties with rectum, bowel movements, or urination (passing water).'

Kaplan and Anderson (1988) stated that 5 distinct steps were involved in building the Health Decision Model.

1 Defining a functional status clarification - first defining a set of mutually exclusive and collectively exhaustive levels of functioning.

2 Classifying symptoms and problems. Kaplan and Anderson say that in addition to Function level clarifications, an exhaustive list of symptoms and problems was generated as `subjective complaints are an important component of a general health measure because they relate dysfunction to a specific problem', and that 21 CPX complexes represented `all the possible symptomatic complaints that might

inhibit function'.

3 Using preference weights to integrate the 3 functional sub-scales and the Symptom/Problem complexes into a single numerical expression. Human judgement studies were used to determine weights for the different states, and random samples of citizens from the community were used to evaluate the relative desirability of a good number of health conditions. A mathematical model was developed to describe the consumer decision process, these weights then being used to describe the relative desirability of all the function states on a scale from zero (for death) to 1.0 (for asymptomatic optimum function).

Kaplan (1988a) has further applied the QWB to develop a *General Health Policy Model.* The aim of the Model is to express benefits and side-effects of different health programs in terms of equivalents of completely well-years of life. This entails the calculation of the expected duration of stay in each QWB levels over time, calculating the product and then adding them, to produce what Kaplan terms the Well-Life expectancy, expressed in well-years. The model is further described in Kaplan et al. (1989) and Kaplan and Anderson (1988).

Kaplan and Anderson (1988) note that standardised questionnaires are used to clarify individuals into each of the function scale steps, while individuals are classified into the CPX by the one symptom or problem that bothers them most. It is claimed that with structured questions an interviewer can obtain a classification on the QWB in 11 to 16 minutes.

The QWB has been used to evaluate outcomes in conditions such as AIDS, cystic fibrosis, and arthritis (see Kaplan et al. 1989).

2.2 Validation

Akin to but seemingly more than the SIP, the QWB appears to rely upon the appropriateness of its initial process of item development and selection. Kaplan et al. (1976) explain how an extensive specialty-by-specialty review of medical reference works was conducted to list all the ways diseases and injuries could affect behaviour and role performance. This lead to items expressed in terms of behaviours (as opposed to capacities, as such phrasing was reported to cause under-reporting of dysfunction), embodied in three functional scales, with survey instruments developed to classify a person into one only of the steps of each of the three scales.

The construct validity of the IWB was seen to be established through convergence between its scores and scores of numbers of chronic conditions (r=0.96), number of physician contacts (r=-0.55), etc. and discriminant evidence that eg. the correlation between IWB on as given day and self-rated well being on days successively distant decreased systematically (from 0.46 when both relate to the same day, to 0.36 when the well-being measure related to 8 days previously).

A further interesting issue reported by Kaplan et al. was that self-rating of overall health status, where the future as well as the present was taken into account, correlated negligibly with IWB on a given day. That is, IWB (which is specific to a day), gave almost no information about expected future well-being as perceived by individual respondents.

A checking of the QWB model

Balaban et al. (1986) re-developed weights for the QWB using a population of sufferers from rheumatoid arthritis. A comparison was made between the QWB score obtained for the 132 scenarios obtained from this population with the scores that would be relevant if Bush's original weights were used. 132 scenarios from combinations of Age, LEV, and CPX were formed, including `anchor' scenarios with optimal functioning, and duplicate scenarios to assess respondent consistency.

The task required subjects were asked to sort each scenario into one of 11 sorting slots, numbered 0 to 10. Zero (0) was marked `as bad as dying', and 10 as `completely well'. The data was then analysed by regressing the global rating given by the subjects against the scenario variables. The model for the regression was QWB score = constant+ CPX+ PAC+ SAC+ MAB. The independent variables were represented by dummy variables, with 0 = no deficit, there being 36 dummy variables in all, ie. Rating = constant + CPX1.. .CPX21+ ... + PAC1.. 3 + SAC1. .4 + MAB1..4

The computer regression model assumed no interaction between variables, but analysis of variance was performed upon a test subset of 24 scenarios which was made up of 4 randomly selected LEVs, 3 randomly selected CPXs, and 2 age groups.

Results were analysed by taking an individual rater's assignment of a score to a particular scenario (category 0-10) and dividing these by 10 to give a range of 01. The average value given by raters to perfect health scenarios was found to be 0.972. The mean ratings for the 2 pairs of duplicate scenarios were 4.35/4.36, and 5.94/5.93.

The output of the regression analysis gave MOB values a range of from .000 to .147, PAC .000 to .073, SAC .000 to .126, and CPX .000 (no CPX) to .654. From the article it would seem that the higher the value, the lower the quality of wellbeing, so should be treated as negative values.

In order of potential effect, they were CPX (.654), MOB (.147), SAC (.126), then PAC (.073). In fact the smallest CPX (of 21) was .130, which was larger than the highest SAC, the highest PAC, and the second highest MOB - which suggests that the CPX tended to swamp the other aspects.

The test for interaction between the variables in the 24 item subset was reported to show no significant interactions.

A comparison was made between the QWB score obtained for the 132 scenarios obtained from this population (rheumatoid arthritis sufferers) with the scores that would be relevant if Bush's weights were used, and the two data groups were reported to have a correlation of r=0.937. However results for scenarios which compared arthritis-type scenarios only were not given. This leaves open the question as to the accuracy of estimates made by the general public vis a vis patients of specific conditions. It would not be surprising if, as was suggested here, arthritics did not differ much from the general population when considering non-arthritis type condition which both are equally uninformed about.

2.3 Problems for the QWB

Despite Kaplan et al's efforts to ensure the validity of the QWB (and much effort was invested by Kaplan and co-workers to determine that a rating task could produce interval data, as discussed earlier in Chapter C), the values that emerge from the scale seem quite unexpected.

For example, the highest negative weight associated with Physical Activity is 0.077, which corresponds to `in wheelchair, did not move or control the movement of wheelchair without from someone else, or in bed, chair, or couch for most or all of the day (health related).' At the same time, CPX-11 `cough, wheezing, or shortness of breath, with or without fever, chills or aching all over' has a weight of -0.257. Because each of the elements of the QWB scale are simply added to 1 (ie. W = 1- CPX weight - MOB weight - PAC weight - SAC weight), CPX-11 is estimated to have over 3 *times* the importance of well-being as does the most severe physical activity limitation. This means that a person totally confined to a wheelchair with no other limitation or symptom would be determined to have a score of .923 of total health functioning, while a person with `cough, wheezing, or shortage of breath' but otherwise totally healthy would have a score of .743, which seems almost absurd.

Even quite minor CPX complexes can have a major effect on QWB. For example, wearing eyeglasses (weight = - 0.101), having a runny nose (- 0.170), and breathing polluted air (- 0.101). Such symptom/problem weights are stated by Kaplan and Anderson (1988b) to allow the health status index to become `very sensitive to minor top-end variations in health status'. However the `minor' adjustments seem to have the capacity to overwhelm the function scales. Thus the need to wear glasses (-0.101) causes greater reduction in wellness than being in a wheelchair (-0.060 or -0.077). This is not necessarily a problem if other aspects of the function scales always also contribute in such cases, but the fact remains that the scales are deemed to be independent and linearly added in the QWB model.

The QWB may also be criticised on the ground that it was developed on the basis of general community rather than patient or care-giver weights. Kaplan and Anderson (1988b) defended this aspect by stating that Balaban et al. (1986) found weights obtained with arthritis sufferers to be `remarkably similar to those we obtained from members of the general population' (Kaplan & Anderson 1988, p.213), while also noting that community

weights means are general and do not bias policy analysis towards any interest group. On the other hand a mix of representative health care receivers may be an alternative way to avoid bias that also allows the QWB to be scaled based on knowledge/experience as opposed to hypothetical estimates.

A further criticism of the QWB in that it does not include measures of social health or mental health. Kaplan has responded to this by stating that social support is not an outcome that can serve as a target for health care, but that on the other hand social functioning is included in the model in the Social Activity Scale. Kaplan sees social function as `social contacts' (eg. participation in work, attendance at school), while social support refers to `social resources' (social life, friendships, family relationships). Kaplan sees the latter as not being in the realm of health policy (asking rhetorically`would we want to develop a health policy that requires people to have friends?'). However Hall et al. (1989) have shown that the loss of social support as a product of a health state may have a major effect on well-being, and for this reason is worthy of consideration for assessment. Regarding mental health functioning, this is seen to be an intervening variable in affecting physical functioning, and Kaplan claims, not all that convincingly, that the QWB picks up such effects in its current formulation.

A final question concerns the content validity of the QWB. This is reliant upon the scenarios used to develop the QWB containing accurate representations of the range of possible conditions, or at least a reasonable sub-set. The value of the global ratings elicited by the scenarios is limited by the comprehensiveness/accuracy of scenario descriptions as well as the knowledge of raters. The gaining of a correlation between arthritis sufferers and the general public on the same scenarios seems more a test of reliability than validity, and it cannot be assumed that content validity has been satisfied.

3 Torrance's Utility Model

Torrance (1972) proposed the development of a health utility index for measuring health improvement that was disease and program independent. An adequate scale was seen to be one that might range from 0 for death (with negative values assigned to fates worse than death) to 1 for good health (defined as the absence of physical, mental, and social disabilities and symptoms). With such an index every individual could be assigned a score to represent level of health, and the effects of a health program could be measured in health improvement calculated in health days. The index value assigned to a particular health state was the *utility of that state as perceived by society,* and the model would allocate resources in the health service system so as to maximise their total health utility to society (Torrance 1972, p. 100). The health index for the state was intended to represent the utility of that state unaffected by future states that might or might not follow, with prognoses incorporated into the model at a later stage of the analysis.

The techniques proposed to measure the utility of specific health states on a linear scale were the Von Neumann-Morgenstern standard gamble, and a new technique the `time trade-off' method. These were claimed to produce equivalent and reliable results, but with

the time trade-off easier to administer. Further research was advocated to develop a general health utility scale that would eliminate the need to measure specific utilities for each application of the model.

As set out by Torrance (1987), utility refers to the desirability or preference that individuals exhibit for a condition, and a definition of utility is that it is a cardinal measure of the strength of one's preference. The notion as used by Torrance refers to modern utility theory, citing Von Neumann and Morgenstern (1953), which describes a method of decision making under uncertainty, based on a set of axioms of rational behaviour. The theory is seen to have face validity in terms of how a problem *should* be approached if a decision is to be made rationally as defined by the basic axioms of the theory (see Torrance & Feeny 1989), ie. it is not concerned with modelling how decisions are *actually* carried out.

3.1 Testing of instruments

Torrance (1976) noted that a number of health status index models had already been proposed, each defining a set of health states and all requiring a set of numerical weights for these states; these weights were seen to be the weakest aspects of the models. Torrance proposed that a measuring instrument was needed that could reliably and validly quantify preferences for health states, and examined the standard gamble, time trade-off, and category scaling methods.

The subjects used varied with the task, such that only a small university educated group (43 subjects) received all three tasks, the standard gamble being considered too complex for the general public. The university educated group also were re-tested a year later to measure the test retest reliability of the three techniques. The major subject group was a general public sample (246 subjects), who received the time trade-off and category rating tasks, as also did a small group of patients in a home dialysis program (29 subjects).

Ten (10) health states were presented, each described by a scenario in narrative form that outlined the physical, emotional and social characteristics of the state.

The standard gamble task consisted of the subject being offered two alternatives, one consisting of an outcome health state with certainty, the other being a gamble with specified probabilities of two possible outcome health states. This task was presented with the assistance of props of a probability wheel and associated colour coded cards.

The time trade-off task also entailed a choice between two alternatives, but neither was a gamble. Each was a different health state but for differing periods of time.

The category rating task was a complex one, requiring simultaneous assessments of the effects of both the illness state per se, and the affect of being in that state for varying periods of time until death. The task entailed presenting the subject with a 100 millimetre

line marked `Death, least desirable' at one end, and `Healthy, most desirable' at the other, and told it represented 101 equal interval categories. For each health state scenario presented, the subject was asked to mark three lines on the `desirability line', one corresponding to the relative desirability of being in that state for three months then death, one for being in that state for eight years then death, and a final line indicating the relative desirability of being in that state for the remainder of his/her normal life expectancy.

The results were that subjects found the time trade-off task the easiest, the standard gamble slightly more difficult (but probably impossible without the props), and the direct scaling task the most difficult. Only the time trade-off task was considered to be capable of being executed without a well trained interviewer

Measures of internal consistency (coefficient of reliability) provided correlations of .77 and .79 for the standard gamble and time trade-off correspondingly, with standard errors of estimate of .125 and .135 correspondingly, and test re-test reliabilities of .53 and .62 correspondingly. The category scaling task test re-test reliability was found to be .49.

3.2 Approach to validity

The validity of the tasks were assessed on the assumption that the standard gamble was a criterion measure as it was valid by definition, and it was not necessary or appropriate to validate the measures obtained from this task.

The other two tasks were compared to the standard gamble, obtaining uncorrected correlations of .65 and .36 for the time trade-off and category rating tasks correspondingly. In essence it was concluded that the time trade-off was the method of choice, being simpler to apply, of relatively high validity, and if anything more reliable. The category rating task was condemned as the poorest technique, its only redeeming virtue being its potential lower cost.

Sackett and Torrance (1978) reported further analysis of the general population interviews using the time trade-off task, looking at variables that affected utility estimates within the group. They found effects of:

age - with older subjects providing lower utility scores for dialysis and transplantation, and higher for hospital confinement for an unnamed infectious disease.

socio-economic class - hospital confinement tended to have higher utility for those living in lower income neighbourhoods.

duration of illness state - the mean daily utility was less for lengthy illness states than short duration ones.

label given to the illness state - eg. use of the label `tuberculosis' gave higher health state

utilities than `unnamed contagious disease'.

experience with the condition - home dialysis patients gave a higher utility to home dialysis than did the general public.

It should be clear that Torrance (eg. Torrance & Feeny 1989; see also Feeny & Torrance 1989) sees the standard gamble as being basically different to, say, the rating scale approach. The standard gamble is seen to provide numerical values that may be termed utilities, while the rating scale provides numbers that are more properly termed values, the reason being that the standard gamble values health states under uncertainty while the rating scale values health states under certainty. Only the standard gamble is seen to directly measure Von Neumann-Morgenstern utilities, all other instruments are at best approximations. As noted by Mulley (1989), the measurement technique is seen to be validated based on axioms of rational choice. This approach does not seem to allow for the limitations of human cognitive ability. It assumes, for example, that probabilities and values for states can be manipulated according to the axioms of the model. To the extent to which, say, the standard gamble task is not computed as expected by the model, then the parameters derived from an analysis using such a task are likely to be in error. The value placed on a given health state cannot necessarily be derived from a decomposition of the results from a complex decision under uncertainty. The approach is quite different to that used by those concerned with measuring a psychological construct. If quality of life is interpreted as psychological well-being, then the aim is to develop a tool which can provide consistent and valid estimates of well-being depending upon health state, not normatively modelling a decision process.

3.3 Development of a Multi-attribute Utility (MAU) Scale

Torrance (Torrance et al. 1982; Boyle et al. 1983) reported the application of multi-attribute utility (MAU) theory to measure social preferences for health states. It should be noted that multi-attribute utility theory is not seen to introduce any new methods of measuring preference, but rather introduces a way of selecting specific preferences to be measured and combining them into a mathematical models of the subjects utility structure (Torrance 1986).

Torrance (1982) described a measure of health status that categorised states using four attributes. These were mobility and physical activity (6 levels), self care and role activity (5 levels), emotional well-being and social activity (4 levels), and health problem (8 levels). Each individual was to be classified using one level only for each of the four attributes. The standard gamble task was not used for the purpose of developing the MAU, due to its complexity and difficulty of administration. Instead a simple category rating task was used to make single attribute measurements, respondents asked to rate the desirability or undesirability of a health state relative to other health states and relative to the reference states `health' and `dead'. Time trade-off was also used to make multi-attribute measurements. The relative value that members of society attach to the various possible health states was determined through measuring aggregated social preference (utility) for

each of the possible 960 states in the classification system) using a random sample of parents of school aged children from the general population.

This general approach was described by Torrance (1986, see also Torrance 1987) as representing the idea of a multi-attribute health state classification system, based on the concept that health state can be defined in terms of a number of attributes (the QWB system of Bush/Kaplan is given as another example of this).

Regarding who should provide utility values, Torrance seems quite flexible. He suggested that informed members of the public are the appropriate subjects for gathering utilities regarding public policy decisions, although he noted that obtaining informed subjects, in the sense that the subject truly understands what the health state is like, is a very difficult issue. He also stated that the simplest way of obtaining utility values was to use judgement to estimate it, either by the analyst or by a few physicians or other experts (however the literature or other forms of measurement was advocated if the analysis showed the results were sensitive to utility values). Torrance also considers that in some applications the relevant health states for which utilities are required are simply the health states of the patients in the study, and there the patients are available to give those utilities - this is seen to avoid the need to prepare a description of the health state for use by subjects.

Torrance's approach to validity appears to have been modified since the earlier papers. Torrance (1986) noted that an alternative approach to validity (other than assuming the necessary validity of the standard gamble) is that health state utilities measure the overall quality of life associated with the health state, and a test of validity is to determine if the utility measure is correlated significantly with other trusted measures of health-related quality of life.

Torrance's MAU scale seems to have undergone little validation that would be accepted by those with a background in psychological measurement: similar in this respect is the work of Rosser.

4 Rosser's Classification of Illness State

Rosser (eg. Rosser & Watts 1972; Rosser & Kind 1978; Kind & Rosser 1988) developed the Classification of Illness State, often referred to as the Rosser Index. This index has been used in a number of cost utility studies, most notably by Williams (eg. Studies of Coronary Artery Surgery; Williams 1985, 1987; see also Loomes & McKenzie 1989).

As described by Gudex and Kind (1988), the format for describing states of illness was derived from material initially generated by a small group of doctors who were asked to describe the criteria they used to decide on severity of illness of their patients, considering only the present state of the patient.

Diagnosis was rejected as being too complex for descriptive purposes, and two major elements were determined: disability (loss of function and mobility), and subjective distress

(pain). From this were developed 8 levels of disability, and 4 levels of distress. This produced 29 states, one of the levels of disability being unconsciousness. Exploration of the same issues with economists and health administrators was described as supporting a similar classification system.

The reliability, accuracy and practicability of the system was trialled by Rosser and Watts (1972), using 50 specialist doctors who rated 2,120 patients in admission and discharge at a Teaching hospital. The descriptive system was found to be reliable and easily used (taking 10 seconds to assess a patient). Weighting of the 29 states was carried out using 70 subjects from 6 different groups, ie. 10 patients in medical wards, 10 psychiatric inpatients, 10 general nurses, 10 psychiatric nurses, 20 healthy volunteers, 10 specialist doctors. In brief, the subjects were required to place six selected states in order of severity, and then to consider the states in that order in pairs, judging how many times more ill a person in the more ill state would be compared to the less ill state. This was seen to be a form of magnitude estimation. The provisional scores of these marker states were then used to provide a framework within which the remaining states were ranked and scored by the same subjects.

Goodinson and Singleton (1989) criticise the QALY concept, noting that it `is an extremely impoverished one', that QOL tests developed for treatment selection include many more dimensions than physical mobility and pain, and actual quality of life for patients may be very different from that implied from calculations based on physical disability and pain alone. They also criticise the notion that healthy respondents yield information that is universally applicable.

Gudex (1986) made similar criticisms of the Rosser scale, noting that:

1 It leaves out important factors, eg. marriage satisfaction, sexual functioning, reproductive ability. This was demonstrated when applying the scale to the treatment of illness states such as cystic fibrosis and scoliosis. In cystic fibrosis female patients are often advised not to become pregnant because of potential complications in pregnancy. In scoliosis a high proportion do not marry, and where marriage occurs there is a higher than average divorce rate and fewer children.

2 There appears to be an undue emphasis placed on the ability to undertake paid employment, and a corresponding lack of sensitivity to incapacities experienced by groups such as the aged, children etc.

3 The values placed on the illness states were questionable, given that they were developed solely upon the small, unrepresentative, mixed sample of 70 people noted earlier.

5 Concluding Observations

None of the instruments reviewed appeared to represent a generic instrument suitable for

general application to resolve policy or program issues. Much more research and development is needed. In the meantime as suggested by Skeel (1985), one needs to use parallel methods when employing QOL instruments: to use instruments the psychometric properties of which are known, and to concurrently develop improved instruments that will be further refined and validated over time.

To aid policy making, clinical research and patient care, there is also a role for a `reference laboratory' which maintains an inventory of instruments which it has characterised, and could recommend instruments, interviewers, and analyse the data (as suggested by Patrick & Deyo 1989). Such a laboratory would need to conduct head-to-head comparisons of competing scales, to test instruments in the same settings and populations, so they can be ranked in reliability, validity, sensitivity, and ease of application.

BIBLIOGRAPHY

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

Anastasi, A. 1990, *Psychological Testing*, MacMillan Publishing Company, New York.

Anderson, N.H. 1976, `How functional measurement can yield validated interval scales of mental attitudes', *Journal of Applied Psychology*, vol. 61, no. 6, pp. 677-692.

Anderson, J.P., Bush, J.W. & Berry, C.C. 1986, `Classifying function for health outcome and quality of life evaluation', *Medical Care*, vol. 24, no. 5, pp. 454-470.

Balaban, D.J. et al. 1986, `Weights for scoring the quality of well being instrument among rheumatoid arthritics', *Medical Care*, vol. 24, pp. 973-.

Bergner, M. et al. 1976a, `The sickness impact profile: Conceptual foundations and methodology for the development of a health status measure', *International Journal of Health Issues*, vol. 6, pp. 393-.

Bergner, M. et al. 1976b, `The sickness impact profile: Validation of a health status measure', *Medical Care*, vol. XIV, pp. 57-67.

Bergner, M. et al. 1981, `The sickness impact profile: Development and final revision of a health status measure', *Medical Care*, vol. XIX, pp. 787-.

Bergner, M. 1989, `Quality of life, health status, and clinical research', *Medical Care*, vol. 27, pp. S148-.

Bombardier, C., Ware, J., Russell, I.J. et al. 1986, `Auranofin therapy and quality of life in patients with rheumatoid arthritis: Results of a multi-centre trial', *American Journal of Medicine*, vol. 81, pp. 565-.

Boyle, M.H., Torrance, G.W., Sinclair, J.C. & Horwood, S.P. 1983, `Economic evaluation of neonatal intensive care of very-low-birth-weight infants', *New England Journal of Medicine*,

vol. 308, pp. 1330-1337.

Breetvelt, I.S., & Van Dam, F.S.A.M. 1991, `Under reporting by cancer patients: The case of response-shift', *Social Scientific Medicine*, vol. 32, no. 9, pp. 981-987.

Brooks, R.G. 1991, Health status and quality of life measurement: Issues and developments, The Swedish Institute for Health Economics, Lund.

Campbell, A. 1981, *The Sense of Well-Being in America*, McGraw-Hill, New York.

Campbell, D.T. & Fiske, D.W. 1959, `Convergent and discriminant validation by the multitrait - multimethod matrix', *Psychological Bulletin*, vol. 56, pp. 81-105.

Carter, W.B. et al. 1976, `Validation of an interval scaling: The sickness impact profile', *Health Services Review*, vol. 11, pp. 516-528.

Cattell, R. `Factor analysis', in *Encyclopaedia of Psychology*, eds H.J. Eysenck, W.J. Arnold & R. Meili, Fontana/Collins.

Cohen, J. 1968, `Multiple regression as a general data analysis system', *Psychology Bulletin*, vol. 70, pp. 426-443.

Crocker, L.M. & Algina, J. 1986, *Introduction to Classical and Modern Test Theory*, Holt, Rinehart & Winston.

Deyo, R.A. & Patrick, L.P. 1989, `Barriers to the use of health status measures in clinical investigation, patient care, and policy research', *Medical Care*, vol. 27.

Diener, E. 1984, `Subjective well-being', *Psychology Bulletin*, vol. 45, pp. 542-575.

Donovan, K. et al. 1989, `Measuring quality of life in cancer patients', *Journal of Clinical Oncology*, vol. 7, pp. 959-968.

Eisler, H. 1962, `On the problem of category scales in psychophysics', *Scandinavian Journal of Psychology*, vol. 3, pp. 81-84.

Epstein, A.M., Hall, J.A., Tognetti, J., Son, L.H. & Conant, Jr, L. 1989, `Using proxies to evaluate quality of life', *Medical Care*, vol. 27, no. 3, pp. S91-S98.

Eyfuth, K. 1972, `Scaling', in *Encyclopaedia of Psychology* eds H.J. Eysenck, W.J. Arnold & R. Meili, Fontana/Collins.

Feeny, D.H. & Torrance, G.W. 1989, `Incorporating utility-based quality-of-life asessment measures in clinical trials', *Medical Care*, vol. 27, no. 3, pp. S190-S204.

Ferguson. 1986, *Statistical Analysis in Psychology and Education*, McGraw Hill, New York.

Gescheider, G.A. 1988, `Psychophysical scaling', *Annual Review of Psychology*, vol. 39.

Goodinson, S.M. & Singleton, J. 1989, `Quality of life: A critical review of current concepts, measures, and their clinical implications', *International Journal of Nursing Studies*, vol. 6, no. 4, pp. 327-341.

Greer, S. et al. 1979, `Psychological response to breast cancer: Psychological outcome', *Lancet*, vol. ii, p. 785.

Gudex, C. 1986, QALYs, and their use by the health service, Discussion paper 20, Centre for Health Economics Consortium, University of York, York, United Kingdom.

Gudex, C. & Kind, P. 1988, The QALY toolkit, Discussion paper 38, Centre for Health Economics Health Economics Consortium, University of York, York, United Kingdom.

Guyatt, G.H., Van Zantan, S.J.O.V., Feeny, D.H. & Patrick, D.L. 1989, `Measuring quality of life in clinical trials: A taxonomy and review', Special article, *Clinical Medicine of America Journal*, vol. 140, pp. 1441-1448.

Hall, J.A., Epstein, A.M. & McNeil, B.J. 1989, `Multidimensionality of health status in an elderly population', *Medical Care*, vol. 27, pp. S168.

Haig, T.H., Scott, D.A. & Wickett, L.I. 1986, `The traditional zero point for an illness index with ratio properties', *Medical Care*, vol. 24, no. 2, pp. 113-124.

Haig, T.H., Scott, D.A. & Stevens, G.B. 1989, `Measurement of the discomfort component of illness', *Medical Care*, vol. 27, no. 3, pp. 280-287.

Hays, W.L. 1981, *Statistics for Psychologists*, Holt, Rinehart & Wilson, New York.

Heady, B., Glowacki, T., Holmstrom, E., & Wearing, A. 1985, `Modelling change in PQOL', *Social Indicators Research*, vol. 17, pp. 267-298.

Heady, B. & Wearing, A.W. 1989, `Personality, life events, and subjective well-being: Towards a dynamic equilibrium model', *Journal of Personal and Social Psychology*, vol. 57, pp. 731-739.

Heady, B., Veenhoven, R. & Wearing, A. 1991, `Top-down versus bottom-up theories of subjective well-being', *Social Indicators Research*, vol. 24, pp. 81-100.

Hughes, J.E. 1985, `Depressive illness and lung cancer: Depression before diagnosis', *European Journal of Surgical Oncology*, vol. 11, pp. 15-20.

Kaplan, R.M. 1988a, `Models of health outcome for policy analysis', Paper prepared for special issue of *Health Psychology*, August.

Kaplan, R.M. 1986, `Health related quality of life in cardiovascular disease', *Journal of Consultation and Clinical Psychology*, vol. 56, pp. 382-392.

Kaplan, R.M. & Anderson, J.P. 1988, `A general health policy model: Update and applications', *Health Services Research*, vol. 23, pp. 203-.

Kaplan, R.M., Anderson, J.P., Wu, A.W., Mathews, W.C., Kozin, F. & Orenstein, D. 1989, `The quality of well-being scale', *Medical Care*, vol. 27, no. 3, pp. S27-S43.

Kaplan, R.M., Bush, J.W. & Berry, C.C. 1976, `Health status: Types of validity and the index of well being', *Health Services Research*.

Kaplan, R.M., Bush, J.W. & Berry, C.C. 1979, Health status index: Category rating versus magnitude estimation for measuring levels of well-being', *Medical Care*, vol. XVII, no. 5, pp. 501-521.

Kaplan, R.M. & Ernst, J.A. 1983, `Do category rating scales produce biased preference weights for a health index?', *Medical Care*, vol. XXI, no.2, pp. 193-207.

Kazis, L.E., Anderson, J.J. & Meena, R.F. 1989, `Effect sizes for interpreting changes in health status', *Medical Care*, vol. 27, no. 3, pp. S178-S189.

Kind, D.L. & Rosser, R. 1988, `The quantification of health', *European Journal of Social Psychology*, vol. 18, pp. 63-77.

Lipscomb, J. 1989, `Time preference for health in cost-effectiveness analysis', *Medical Care*, vol. 27, no. 3, pp. 233-253.

Llewellyn-Thomas, H., Sutherland, H.J., Tibshirani, R., Ciampi, A., Till, J.E. & Boyd, N.F. 1984, `Describing health states', *Medical Care*, vol. 22, no. 6, pp. 543-552.

Llewellyn-Thomas, H.A., Thiel, E.C. & McGreal, M.J. 1992, `Cancer patients; evaluations of their current health states', *Medical Decision Making*, vol. 12, pp. 115-122.

Loomes. G. & McKenzie, L. 1989, `The use of QALYs in health care decision making', *Social Scientific Medicine*, vol. 28, no. 4, pp. 299-308.

Martin, D.P., Gilson, B.S., Bergner M. et al. 1976, `The sickness impact profile: Potential use of a health status instrument for physician training', *Journal of Medical Education*, vol. 51, pp. 942-.

McCauley, C. & Bremer, B.A. 1991, `Subjective quality of life measures for evaluating

medical intervention', *Evaluation and the Health Professions*, vol. 14, no.4, pp. 371-387.

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

McNeil, B.J., Pauker, S.G., Sox, H.C., & Tversky, A. 1982, `On the elicitation of preferences for alternative therapies', *New England Journal of Medicine*, vol. 306, pp. 1259-1262.

Moser, C. & Kalton, G. 1979, *Survey Methods in Social Investigation*, 2nd edn, Heinemann Educational Books, London.

Mulley, Jr, A.G. 1989, `Assessing patients; utilities', *Medical Care*, vol. 27, pp. S269-281.

Murphy, D.J. & Cluff, L.E. 1990, `Introduction: The SUPPORT study', *Journal of Clinical Epidemiology*, vol. 43, pp. S v-x.

O'Connor, R.E. 1991, Impairment assessment review, Report prepared for the Impairment Assessment Review Committee of the Accident Compensation Commission of Victoria (ACC) and the Victorian Accident Rehabilitation Commission (VARC), June.

O'Connor, R.E. 1992a, The development of a second edition of the disability and handicap severity scales, Prepared for the Disability and Handicap Working Party of the Accident Compensation Commission of Victoria (ACC) and the Victorian Accident Rehabilitation Commission (VARC), June.

O'Connor, R.E. 1992b, `Health-related quality of life measures need content validity, *Australian Health Review*, vol. 15, no. 2, pp. 155-163.

Parducci, A. 1968, `The relativism of absolute judgements', *Scientific American*, vol. 219, pp. 85-90.

Patrick, D.L. & Deyo, R.A. 1989, `Generic and disease-specific measures in assessing health status and quality of life', *Medical Care*, vol. 27, no. 3, pp. S217-232.

Patrick, D.L., Bush, J.W. & Chen, M.M. 1973, `Methods for measuring levels of well-being for a health status index', *Health Services Research*, pp. 228-245.

Pettingale, K.W. 1984, `Coping and cancer prognosis', *Journal of Psychosoma Research*, vol. 28, pp. 363-364.

Read, J.L., Quinn, R.J., Berwick, D.M., Fineberg, H.V. & Weinstein, M.C. 1984, `Preferences for health outcomes: Comparison of assessment methods', *Medical Decision Making*, vol. 4, no. 3, pp. 315-329.

Revicki, D.A. 1989, `Health related quality of life in the evaluation of medical therapy for

chronic illness', *The Journal of Family Practice*, vol. 29, pp. 377-.

Rosser, R.M. & Kind, D.P. 1978, `A scale and valuation of states of illness: Is there a social consensus?', *International Journal of Epidemiology*, vol. 7, pp. 347-.

Rosser, R.M. & Watts, V.C. 1972, `The measurement of hospital output', *International Journal of Epidemiology*, vol. 1, pp. 361-368.

Sackett, D.L. & Torrance, G.W. 1978, `The utility of different health states as perceived by the general public', *Journal of Chronic Disability,* vol. 31, pp. 697-704.

Shavelson, R.J. 1988, *Statistical Reasoning for the Behavioural Sciences*, 2nd edn, Allyn & Bacon Incorporated.

Skeel, R.T. 1989, `Quality of life assessment in cancer and clinical trials: Its time to check up', *Journal of the National Cancer Institute*, vol. 81, pp. 72-73.

Stevens, S.S. 1972, `Psychophysics' in *Encyclopaedia of Psychology*, eds H.J. Eysenck, W.J. Arnold, & R. Meih, Fontana/Collins.

Stevens, S.S. & Galanter, E.H. 1957, `Ratio scales and category scales for a dozen perceptual continua', *Journal of Experimental Psychology*, vol. 54, no. 6, pp. 377-411.

Sutherland, H.J., Dunn, V. & Boyd, N.F. 1983, `Measurement of values for states of health with linear analog scales', *Medical Decision Making*, vol. 3, pp. 479-487.

Temkin, N.R., Dikmen, S., Machamer, J. & McLean, A. 1989, `General versus disease-specific measures', *Medical Care*, vol. 27, no. 3, pp. S44-53.

Torrance, G.W. 1976, `Social preference for health status', *Socio-Economic Planning Science*, vol. 10, pp. 129-136.

Torrance, G.W. 1986, `Measurement of health state utilities for economic appraisal', *Journal of Health Economics*, vol. 5, pp. 1-30.

Torrance, G.W., Boyle, M.H. & Horwood, S.P. 1982, `Application of multi attribute utility theory to measure social preference for health states', *Operations Research*, vol. 30, pp. 1043-1069.

Torrance, G.W. & Feeny, D. 1989, Utilities and quality-adjusted life years, *International Journal of Technology Assessment in Health Care*, vol. 5, pp. 559-575.

Torrance, G.W., Thomas, W.H. & Sackett, D.L. 1972, `A utility maximisation model for evaluation of health care programs', *Health Services Research*, vol. 7, pp. 118-133.

Torrance, G.W. 1987, `Utility approach to measuring health-related quality of life', *Journal of Chronic Disability*, vol. 40, pp. 593-600.

Tsevat, J. et al. 1990, `Assessing quality of life and preference in the seriously ill using utility theory', *Journal of Clinical Epidemiology*, vol. 43, pp. 735-775.

Tversky, A., & Kahneman, D. 1981, `The framing of decisions and the psychology of choice', *Science*, vol. 211, pp. 453-458.

Williams, A. 1985, `Economics of coronary artery bypass grafting', *British Medical Journal*, vol. 291, pp. 326-329.

Williams, A. 1987, `The cost-effectiveness approach to the treatment of Angina', in *The Management of Angina Pectoris*, Castle House Publications Ltd.

APPENDIX A

Criteria for developing a satisfactory expert-referenced test of work-related disability.

For a reliable and valid measure to be formed, the following conditions need to be satisfied:

During initial test development:

1 The team of judges used to provide the proxy criterion measure need to possess the mix and range of skills necessary to estimate the effect of a disability on capacity to work.

2 The judges must have a clear and agreed conception regarding the judgements they are to make (ie. need to clearly define `ability to work')

3 The team must be presented with characterisations that as far as possible embody and reflect the nature of actual cases.

4 The cases presented must be representative of the population of cases to be assessed in practice.

5 The cases must be able to be represented by sets of items or descriptors (sub-scales), that indicate disability in each of the disability sub-areas found in the cases.

6 The items/descriptors used must be clear, unambiguous, and weighted so as to predict the contribution to overall disability arising from the disability area. This requires that descriptor within sub-scales be independent.

7 All disability areas relevant to accurately assessing disability must be represented.

8 The disability areas making up the overall scale must be weighted to allow valid representation of overall work-related disability.

On being applied in practice:

9 The test must be shown to be exhibit inter-rater reliability, ie. different clinicians should make the same assessment if faced with a common patient (in terms of both descriptor selection, and overall score).

10 The test score should correlate highly with unbiased clinical estimates of work-related disability (convergent validity).

11 The output of the test should correlate significantly with other measures of disability (convergent validity).

12 The test should be capable of clearly distinguishing between sub-populations of patients (discriminant validity)

13 The test should be capable of clearly distinguishing within sub-populations of patients (sensitivity)

14 Ultimately, the test might also be examined for agreement with other indicators of validity. For example, of clients not granted disability status, did those with higher disability ratings show a greater tendency to not return to work/ appear for reassessment/ make an appeal, etc.