CENTRE FOR HEALTH
PROGRAM EVALUATION

WORKING PAPER 26

How Should We Measure 'Change' in Utility Measures of Health Staus – or Should We?

Kaye Brown
Research Fellow, Centre for Health Program Evaluation

Colin Burrows
Senior Research Fellow, Centre for Health Program Evaluation

October, 1992
ISSN 1038-9547
ISBN 1 875677 04 6

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 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
Web Address http://chpe.buseco.monash.edu.au/

ACKNOWLEDGMENTS

The National Centre for Health Program is supported by a grant from the National Health and Medical Research Council's Public Health Research and Development Committee, and by a matching grant from the Victorian Health Promotion Foundation. The research reported in this paper was made possible as a result of this support.

The research was also supported directly by a seeding grant from the Public Health Research and Development Committee. The authors wish to express their gratitude to these funding bodies for their support.

Table 2 Reprinted with permission from Journal of Chronic Disease, Vol 38, B Kirshner, G Guyatt, A methodological framework for assessing health indices, 1985, Pergamon Press Ltd.

Table 3 Reprinted with permission from Journal of Chronic Disease, Vol 40, G Guyatt, S Walter, G Norman, Measuring change over time: assessing the usefulness of evaluative instruments, 1987, Pergamon Press Ltd.

Table 5 Reprinted with permission from M Paterson, Assessment of treatment in rheumatoid arthritis in G Teeling Smith (Ed), *Measuring health: a practical approach*, 1988, John Wiley & Sons Ltd.

Figure 4 Reprinted with permission from Journal of Chronic Disease, Vol 40, W O Spitzer, State of science 1986: quality of life and functional status as target variables for research, 1987, Pergamon Press Ltd.

Figure 5 Reprinted with permission from M Paterson, Assessment of treatment in rheumatoid arthritis, in G Teeling Smith (Ed), *Measuring health: a practical approach*, 1988, John Wiley & Sons Ltd.

Abstract

The two standard requirements of instruments that measure cross-sectional differences in respect of abstract and subjective phenomena (such as health status or health related quality of life) are reliability and validity. However, for evaluative instruments (those designed to measure longitudinal change) another property is required: the instrument must be sensitive to clinically important changes over time, even if such changes are small. To the extent that economic evaluation is dependent on the equality of the underlying medical evidence; and the assessment of the efficiency becomes (more) "vertically integrated" with the assessment of the efficacy and effectiveness dimensions of health care, as prophesied by Drummond and Stoddart (1984), health economists, too, may have to become concerned with measuring responsiveness. The question of responsiveness also lies at the root of some of the disputation about whether to use disease-specific or generic health status measures and the clinical significance or usefulness of scores on health status instruments.

The literature on the measurement of change is often confusing and contradictory. There is controversy on at least two fronts: whether a separate concept in necessary and if it is, how should it be measured. Some researchers advocate the use of change scores as the best approach to the analysis of treatment effects and others maintain that they should be avoided like the plague.

We explore the issues surrounding the measurement of change scores and the methodological advantages and disadvantages surrounding the several change measures that have been suggested. We also examine the limited literature on the extent to which utility measures of health status are capable of measuring change. Implications for the practice of economic evaluation are drawn.

Introduction

The case for the "vertical integration: of economic evaluation and clinical trials is tenable, if not overwhelming. The rationale is multi-faceted but straightforward: there exists a necessary-but-not-sufficient relation between the quality of underlying medical evidence and the quality of an economic evaluation (Sulyer, 1982; Culyer & Maynard, 1981; Laupacis, Feeny, Detsky & Tugwell, 1992; Williams, 1983); the scope for systematic analysis of costs and benefits is limited and the possiblity of pre-emptive action next to non-existent, given the rate of diffusion of new technologies: 'economic evaluation can be done too soon or too late' (Williams, 1983) but 'it's always too early until, unfortunately, it's suddenly too late!' (Buxton, 1987); and, given the complexity and cost of many trials, the costs of assembling economic data are marginal (Banta, 1987).

As with many things, the strength of the case can be improved by retricting the focus somewhat. Drummond and Stoddart (1984) and Mugford and Drummond (1989) have proposed a number of indications for the appropriateness of including economic analysis alongside clinical trials (see also, Feeny, Labelle & Torrance, 1990). One guideline has to do with the nature of the trial itself and is related to the imminence of the therapy's diffusion into the health care system with attendant consequences for resource allocations. Economic evaluations should be restricted to management, rather than explanatory, trials.[1] A second indicator is concerned with whether resource considerations are prominent in the minds of those conducting or funding the research or those financing health care. Is there a shortage of resources that will force a choice between alternative options for care or health programs per se? A third criterion is when the resource consequences a re large: Is the new therapy currently (or potentially gazetted for use on a large patient population? Is there (likely to be) a significant cost difference between the new therapy and the alternative (which may be no therapy)? Is the new therapy radically different in terms of either the treatment (eg, surgery versus medical management) or setting (eg, hospitla versus home)? Are there mooted reductions in the use of other health service resources, due to the new therapy's greater effectiveness or fewer side effects? Although these criteria are in the nature of constraints, they cannot be viewed as overly restrictive. Mugford and Drummond indicate that, whereas not one of the reports in a random sample of 100 trials in the perinatal field included economic analyses, 48% of the studies should have done so, according to the above criteria. This accords with a prevalence study undertaken by Adams, McCall, Gray, et al (1992) which found that only 121 of over 50,000 randomised trials (0.02%) which were published between January 1966 and June 1988 included economic analyses. (On the other hand, their finding that, among the 51 of these 121 studies sampled, the mean quality of research score was 0.32 (SD=0.32) and the mean economic analysis completeness score was 0.52 (SD = 0.13) on scales of 0 to 1, is not propitious).

To all intents and purposes, these strictures specify the conditions under which it would be useful to gather cost data. They are largely concerned with delimiting the circumstances under which economic evaluation of alternatives is likely to pay off (Williams, 1974). More recently, Drummond, Teeling Smith and Wells (1988) have prescribed a role for economists on the benefits side of the calculus. They intimate that economic analysis is also indicated where 'the benefits from the new medicine are unlikely to be captured by normal clinical measures and will therefore require direct assessment of quality of life.' By this they no doubt have in mind the inclusion of utility measures of health-related quality of life which, when combined with costs, yield indexes that can be compared across alternative programs and

[1] Management trials are those which are directly concerned with assessing the effectiveness of a new therapy compared to existing practices (including the 'do nothing' alternative if the status quo is such that there is no effective therapy for the disease in question) when delivered to a defined population. They can be contrasted with explanatory trials where the new therapy's efficacy under ideal conditions, dosage levels, toxicity and pharmacological interactions are being assessed and compared with a placebo.

interventions to gauge the optimal use of scarce resources (Dowie 1991; Richardson & Cook, 1992; Williams, 1988, 1991; Torrance, 1986).

Hence, although the cost-related arguments rehearsed above may have provided the primary impetus for undertaking economic analysis alongside clinical trials, among most economists (and some others), the generation of an index that captures the benefits of interventions in a single number with ration scale[2] properties seems to be regarded as a necessary requirement for policy decisions (Bennett, Torrance & Tugwell, 1990; Glasziou & Sime, 1992; Kaplan, 1989; Kaplan & Bush, 1982; Patrick & Erickson, 1988; Torrance, 1986; Torrance & Feeny, 1989). Certainly, the ability to summarise the implications of clinical trials for resource allocation in a league table that expresses the costs of interventions in relation to the number of quality-adjusted life-years (QALYs) is something that economists regard as useful. As Drummond and Davies (1991, p.565) put it, 'Economists tend to favor utility measures similar to those pioneered by Kind et al {1982} and Torrance {1987], because these enable differences in the costs of interventions to be related to a common unit of output, the quality-adjusted life year (QALY). In turn, this enables economic analyses to have a direct input to discussions about health care priorities.'

The use of utility approach to measuring quality of life as an outcome in clinical trials has one further, and frequently overlooked, corollary if it is to be used for comparisons across interventions. Utility measures must serve two measurement goals: discrimination -- detecting differences among individuals in different health states or receiving different treatments – and evaluation – detecting changes within individuals over time. Very often the goal of discrimination is preeminent, as in the following quotations which all emphasise comparisons facilitated by having a common denominator or common unit of outcome:

> 'The appeal and power of the QALY approach come from its ability, at least in theory, to capture, in a single summary measure, QALYS gained, the health improvement created by any proposed or existing program or technology regardless of disease, type of patient, or type of program' (Torrance & Feeny, 1989, p.572).
>
> 'What we need is a measure that makes explicit and acceptable ethical assumptions, is sensitive in relevant wasy to the possible differences in outcome that there may be, and that enables cross-programme comparisons of outcome and ultimately efficiency to be made' (Culyer, 1991, p.7).
>
> 'Generic measures permit the comparisons of different populations and different programs, a most important objective for policy analysis and decision-making. Continued use of generic measures is necessary for comparing benefits of different health interventions and allocating resources. Cumulative knowledge of health and quality-of-life outcomes using generic measures, will establish the relative burden of different diseases and the relative merit of different alternatives' (Patrick, 1990, p.40).

[2] Not infrequently, health economists assert that the different techniques used to measure utility yield data with interval scale properties. For example, Richardson (1990, p.7) states that 'utility revealed by these techniques {category rating, standard gamble, time trade-off, magnitude estimation, and ratio scale} is taken as having an interval property, that is for example, the difference between utility values of 0.2 and 0.4 is quantitatively equivalent to the difference between 0.6 and 0.8' (see also, Capewell, 1988) In fact, cost-per-QALY comparisons presuppose a ratio scale i.e., one on which multiplication and division, as well as other mathematical operations, can be performed (Green & Lewis, 1986; Veney & Kaluzny, 1984). Utility measurements are manifestly taken to represent the actual amount of the property measured. They are routinely multiplied by remaining life expectancy and divided into treatment costs when cost-per-QALY quotients are calculated.

However, it is always implicit that the evaluative dimension(s) of efficacy or effectiveness are paramount because 'if the project doesn't work, efficiency is of no help (Evans, 1984, p.264). The evaluative dimension is also quite explicit where health-related quality of life profiles, with and without a given program or treatment, are plotted over time, and the effect of the intervention measured as the difference in areas beneath the two curves (Bryan, Parkin & Donaldson, 1991; Drummond, 1990; Fanshel & Bush, 1970; Kaplan & Anderson, 1988; Kaplan & Bush, 1982; Rosser, 1990; Torrance & Feeny, 1989).

Fairly obviously, there must be satisfactory discrimination for cross-comparisons between interventions and, of course, to the extent that such comparisons are made across a large number of interventions competing for resources as proposed in cost-per-QALY league tables (Kaplan & Bush, 1982; Kaplan & Anderson, 1990; Torrance & Zipursky, 1984; Williams, 1985), the demands on discriminability increase. Similarly, the requirements for measurement of change become more formidable as one moves from evaluation of the efficacy of an intervention in a single trial to direct comparison across interventions. Again, it should be emphasised that such comparisons imply ratior-scaled properties of the measures.

According to Kirshner and Guyatt (1985) the goals of discrimination and evaluation are different, may require different instruments, and 'the requirements for maximising one of the functions…may actually

In what follow, we seek to do four things: (1) explore the subtleties of differences between these measurement goals, particularly as they impinge on the psychometric properties of reliability, validity and responsiveness[3]; (2) consider whether change scores ought to be avoided like the plague; (3) discuss the methods for assessing responsiveness; and (4) survey the literature dealing with the application of utility measures in clinical trials and the extent to which responsiveness is manifest. In terms of the taxonomy of health-related quality of life measures put forward by Guyall, Veldhuyzen van Zanten, Feeny and Patrick (1989), much of tihis duscussion applies to health profiles and utility measures alike. In the final section, however, we specifically limit ourselves to examining the responsiveness of utiltiy measures because it is these that impose the greatest demands on both validity and measurement, and it is these that are proposed as vehicles for theoretically sound (and simplified) solutions to complex resource-allocation decisions. However, even there our conclusions are, in part, based on the behavior of the other disease-specific and generic instruments that constitute the primary outcome measures in the studies discussed.

GOALS OF MEASUREMENT: DISCRIMINATIVE AND EVALUATIVE INDEXES

Two basic types of instruments are available to measure health-related quality of life in clinical trials: generic instruments and specific instruments. As the label suggest generic instruments cover the complete spectrum of function, disability and distress that is relevant to health-related quality of life and are therefore applicable in a wide variety of populations. Specific instruments are very often related to specific diagnostic groups or patient populations but they may be specific to given conditions or problems (eg, pain or shortness of breath), certain functions (eg, emotional function, foot function) or special populations (eg, elderly individuals or children). Specific instruments often have the goal of

[3] Our inclusion of responsiveness here anticipates the discussion that follows since there is some controversy as to whether a separate concept of responsiveness in necessary. A psychometrican may reasonably object to the inclusion of the notion of responsiveness along with the standard concepts of reliability and validity. Psychometrics is defined as the area of psychological measurement concerned with individual differences and is therefore usually identified with discriminative indexes (also discussed below) and the concepts of reliability and (cross-sectional) validity.

measuring clinically important changes and tend to focus on areas that physicians routinely examined. Generic instruments can be further categorised into profiles of health dimensions, utility measures and single item self-rating health scales (Guyatt et al, 1989). As the name implies, health profiles yield diaggregated profiles of scores on different health dimensions, that may or may not be able to be aggregated into a single score (index). Utility measures of health status provide a quantitative measure of the value or preference an individual attaches to a particular heath state traditionally relative to perfect health (score of 1) and death (score of 0).

Another classification of health-related quality of life measures with which health economists appear to be less familiar, but which is of paramount importance, is concerned with means-ends relationships. Why are we measuring health-related quality of life? What is the purpose behind the measurement process? Kirshner and Guyatt (1985) identify three broad purposes prediction, discrimination and evaluation. We consider only the last two purposes here.

Kirshner and Guyatt (1985) define a discriminative index as one that is 'used to distinguish between individuals and groups on an underlying dimension when no external criterion or gold standard is available for validating these measures.' By contrast, an evaluative index is 'used to measure the magnitude of longitudinal change in an individual or group on the dimension of interest.' In the present case, the underlying dimension is, of course, health status or health-related quality of life, and measurement goals implicit in each of these definitions are recognisable in the application of instruments designed to discriminate among individuals along a continuum of health, illness or disability and to evaluate within-person changes in health over time, respectively.

Instruments that measure cross-sectional differences must satisfy the two familiar requirements of reliability – does the instrument consistently yield more or less the same result when administered on several occasions to stable subjects?; and validity – does the instrument measure what it is supposed to measure? Very often, however, exercises in health status or health-related quality of life measurement undertaken with the avowed goal of developing an instrument for use in clinical trials begin by demonstrating reliability and validity, and end with the conclusion that the instrument is ready for use in experimental studies. In fact, responsiveness – sensitivity to clinically important changes over time, even if such changes are small – is the *sine qua non* of usefulness for evaluative indexes. A discriminative index that is reliable and cross-sectionally valid may be responsive or unresponsive. An evaluative index that is unreliable may be responsive and longitudinally valid.

RELIABILITY: DEFINITION, MEASUREMENT AND APPLICATION

Reliability is a generic term that is used by psychometricians to indicate both the internal consistency of a scale (are the scores on each of several items that address the same underlying dimension correlated?) and reproducibility (do measurements of individuals on different occasions, or by different observers, or by similar or parallel tests, produce the same or similar results?). We focus here only on reproducibility, more specifically on test-retest reliability (does the same measuring procedure yield the same results on independent repeated trials under the same conditions?) and the statistics used to quantify it.

The most common measure of reproducibility for continuous data in the development of health status instruments, is the product-moment correlation (Buxton, Ashby & O'Hanlon, 1987; Churchill, Morgan & Torrance, 1984; Richardson, Hall & Salkeld, 1990; Torrance, 1976). Pearson's *r* is based on regression analysis and is a measure of the extent to which the relationship between two variables (observations on a group of subjects) can be described as linear. Its disadvantage is that it fails to take into account

variability attributable to systematic, as opposed to random, differeneces in test scores across successive applications. Replicate measurements may be systematically different (perhaps due to learning effects), and yet highly (orperfectly) correlated. For this reason, the intraclass correlation coefficient (ICC) is now widely accepted as a preferable measure of reliability (Deyo, Diehr & Patrick, 1991; Feinstein & Kramer, 1980; Fleiss, 19870. Unlike Pearson's r, the intraclass correlation coefficient reflects both random and systematic differences in test scores, and is applicable for $n \geq 2$ measurements per subject.

The ICC can be calculated as the ratio of the between-subject variance to total variance (variance attributable to both between- and within-subject differences over multiple repetitions of the test). The reliability coefficient can be calculated using standard analysis of variance (ANOVA) routines by recognising the relationship between means square terms and the equivalent variance components. Table 1 indicates the relevant formulae for an ANOVA with one repeated measure and one grouping factor. The reliability coefficient, R, qua intraclass correlation, is:

$$R = \sigma^2(\text{subjects}) / (\sigma^2(\text{subjects}) + \sigma^2(\text{error}) \qquad (1)$$

Rather than measuring the degree of association between two sets of scores, the ICC assesses concordance, the extent to which repetitions of a test yield the same values under the same conditions in the same individuals. It assesses not only the strength of the correlation, but also whether the slope and intercept vary from those expected with replicate measures. The ICC will yield a value of 1.0 only if all the observations on each subject are identical, which dictates a slope of 1.0 and an intercept of 0.0.

Different aspects of reproducibility are relevant for discriminative and evaluative indexes. To discriminate among subjects, the differences observed among individuals must be stable over time. The reproducibility of a discriminative instrument is inversely proportional to the variability within subjects, and directly proportional to the magnitude of variance between subjects. For an evaluative instrument, on the other hand, the only requirement is that replicate measurements on each individual remain stable over time, that is, the magnitude of the within-person variance is small.

However, small changes in within-person variation may be associated with low reliability coefficients. To illustrate the point, consider Table 2 which presents hypothetical data, drawn from Guyatt, Walter and Norman (1987), which pertains to the repeated administration (at Times 1 and 2) to stable subjects of three quality of life instruments (Instruments A, B and C) and subsequent administration of the same instruments after intervention Time 3.) Each questionnaire yields a single score on a scale form 0 to 20. Suppose we are interested only in measuring with subject change according to underlying health status, and that the data in Table 2 were, in fact, generated from two separate studies, as Jaeschke and Guyatt (1990) prescribe: one examining the degree of variability in stable subjects and the other examining whether the questionnaire score changes when real change has occurred. In the latter study, unbeknownst to us (until the code is broken), subjects 1-4 were randomized to the treatment group and subjects 5-8 to the control group. What should we make of these data? What conclusions can we draw about the usefulness of Instruments A, B and C as evaluative indexes?

In summary, it can be said that Instrument A is unreliable and responsive, Instrument B is reliable and unresponsive, and Instrument C is reliable and responsive. Instrument A discriminates poorly between subjects on the basis of their health status, as is reflected in an intraclass correlation coefficient of 0, but has a small within-subject variance. Instrument B, on the other hand, has an intraclass correlation coefficient of 1. Following conventional wisdom, Instrument a would be jettisoned forthwith and Instrument B retained. In fact, however, Instrument A yields a statistically significant difference between treatment and control groups which is consistent with the results of exercise testing (the gold standard

for assessing efficacy in this example). Because we are able to use change scores to measure outcomes, the size of between-subject differences is not relevant to Instrument A's usefulness as an evaluative index.

Instrument C parallels Instrument A in terms of within-subject variance between Times 1 and 2 and with respect to the change scores between pre- and post-tests. Unlike Instrument A, however, Instrument C a high proportion of the total variance in subject scores for Times 1 and 2 is accounted for by between-subject variance, resulting in an intraclass correlation coefficient of 0.976. Hence, Instrument C is both valid, reliable (in the conventional sense) and responsive.

VALIDITY AND RESPONSIVENESS

We know as if by rote that an instrument is valid if it actually measures what it is intended to measure. What is implicit in this definition is that the accoutrements of validity differ according to the purpose for which an instrument is used. This is already well established in the case of discriminative versus predictive indexes (where individuals are classified into a set of predefined measurement categories and a gold standard or criterion is available, either concurrently or in the future, to determine whether individuals have been classified correctly) but it applies to discriminative versus evaluative indexes, too. Thus, whereas cross-sectional construct validity is central in the case of discriminative indexes, with evaluative indexes we seek longitudinal construct validity. The first interpretation focuses on the question, "Do cross-sectional or between-subject differences in index measurements taken at a single point in time bear the expected relationship to external measures?" The second interpretation is concerned with the question, "Do longitudinal or with-subject changes in index scores associated with an intervention bear the expected relation to changes in other variables measured?" Validity has different nuances according to the purpose for which an instrument is developed: just as reliability is a necessary but not sufficient for cross-sectional validity, so responsiveness is necessary but not sufficient for longitudinal validity. Figure 1 summarises these relationships. It treats discriminability" as a synonym for reliability to underscore the fact that discriminative indexes measure position on a spectrum and therefore require high levels of between-subject variance relative to total variance. The reliability coefficient is generally accepted as a measure of the ability of an instruments to discriminate among individuals.

It does no follow that an instrument that is able to measure satisfactorily between-subject differences in health status should necessarily be able to measure satisfactorily changes in health-related quality of life within individuals (Churchill, Wallace, Ludwin, et al, 1991; MacKenzie, Charlson, DiGioia, Kelley, 1986a, 1986b; Wiklund & Karlberg, 1991), or vice versa. As pointed out earlier, instruments that measure quality adjusted life years for cost-utility analysis must satisfy both requirements. It an instrument is to address satisfactorily questions about the efficacy and/or effectiveness of an intervention, its longitudinal validity should be established. If the same instrument is to be used to examine questions of efficiency – to make comparisons across programs – it also needs to have an acceptable level of cross-sectional validity.

Guyatt, Deyo, Charlson et al (1989) provide a series of examples that clarity the relationship between responsiveness and validity of instruments designed to measure health status in clinical trials. The discoursive nature of the examples is helpful where intuition is not. Collectively, they make clear that what counts is whether changes can be validated.

The examples can be categorised in terms of two 2-way classifications according to whether or not an instrument/item is accepted as cross-sectionaly valid (yes/no) and proves to be responsive (yes/no),

and whether or not an instrument/item is responsive (yes/no) and cross-classification to underscore the fact that an instrument/item that is valid for the purpose of discrimination may not be valid for the purpose of measureing change between towo or more points in time.

The relationship between responsiveness and validity can bee illustrated by a means of a number of single items from the Sickness Impact Profile (SIP) (Bergner, Bobbitt, Carter & Gilson, 1981) in the context of a controlled trial of different strategies for managing patients with back pain (Deyo, Diehl & Rosenthal, 1986). Clearly, although items like "I'm eating no food at all, nutrition is taken through tubes and intravenous fluid" or " I do not feed myself at all, and must be fed" may be reproducible and valid in discriminating among individuals in the general population in terms of their health status, they will not be responsive in setting limited to patients with chronic lower back pain. By contrast, other items such as "I keep rubbing or holding areas of my body that hurt or are uncomfortable" or "I am not doing any of the house cleaning that I would usually do" may prove useful in locating individuals in general along a continuum according to their health status and in detecting changes over time in patients with back pain. The first set of items lies in cell C of Figure 2, and the second set is located in cell a within cell A. (Needless to say, this example invites consideration of the trade-offs between generic versus disease-specific measures in terms of the role of hypothesis-driven item selection and the (perceived) effect on responsiveness, as discussed below).

A second example involves a randomized controlled trial of adjuvant chemotherapy in 418 women with breast cancer which involved two treatment arms, a short (12 weeks) arm and a long (36 weeks) arm. A number of process and outcome measures were used including: the Breast Cancer Chemotherapy Questionnaire (BCQ), an instrument designed specifically to measure aspects of physical, emotional and social functions during adjuvant treatment for breast cancer; the Rand physical function and emotional function instruments; a toxicity questionnaire, the Eastern Co-operative Oncology Group Criteria (ECOG); and patient and physician ratings of emotional and physical change at specified intervals.

The BCQ was able to detect statistically significant differences in the 24-week minus 12-week change scores of patients in the short (12 weeks) versus long (36 weeks) arms of the trial whereas the Rand instruments, which have previously been shown to be valid discriminative measures (Ware, Brook, Davies-Avery, et al, 1980), did not demonstrate responsiveness. Accordingly, the Rand instruments are shown in cell C of Figure 2 and the BCQ can be located in cells A and a or cells B and a (depending on whether it is does o does not discriminate among women undergoing adjuvant chemotherapy for breast cancer according to their health status).[4]

By contrast, the behaviour of patients' ECOG change scores indicates that this index in responsive but not (longitudinally) valid as a measure of overall health status. ECOG includes items such as white blood cell and platelet counts, laboratory tests of liver function, evidence of allergic reaction and physician assessment of patients' problems with nausea, vomiting, stomatitis and diarrhoea. Thus, ECOG includes some subjective health components, and may correlate to some extent with the change in subjective health status a patient experiences during the course of chemotherapy. However, it was designed to measure drug toxicity rather than subjective health status. Not surprisingly, it fails the test of clinical sensibility (Feinstein, Josephy & Wells, 1986; Feinstein, 1987a, 1987b): inter alia, it lacks face validity ad content validity as a measure of health status. Accordingly, the ECOG instrument shows up in cells B and b of Figure 2.

[4] To the extent that the BCQ was designed as an evaluative index for this clinical trial, it would be reasonable to tilt in favour of its being responsive and longitudinally valid but not cross-sectionally valid.

A third (and conceptual) example located in cells B and b of Figure 2 is described by the case of a rehabilitation program wherein subjects are asked if their health status has improved as a result of participating in the program. They may well indicate that they feel better in which case the approach is responsive. But the change reflected in subjects' responses may indicate satisfaction with the program or a courtesy bias, rather than a change in their subjective health status, in which case the item/instrument would be responsive but not valid.

To illustrate the coincidence of longitudinal validity and responsiveness, Guyatt et al (1989) discuss the assessment of disease-specific aspects of physical and emotional function in patients with Crohn's disease or ulcerative colitis, using the Inflammatory Bowel Disease Questionnaire (IBDQ) (Guyatt, Mitchell, Irvine et al, 1989b). Two interviews were conducted, one month apart. The IBDQ showed only small intrasubject variability over time in patients who reported their health status as stable. Global ratings of change showed moderate to high correlations with changes in IBDQ score. Moreover, the largest changes in IBDQ scores were registered by patients who reported overall improvement or deterioration, suggesting that the instrument is responsive. Coupled with the fact that these changes were greater than the differences in scores for patients whose global rating of their disease activity suggested they were stable, this result provides evidence for the instrument's longitudinal validity. Hence, the IBDQ example is located either in cells A and a or B and a (again, depending on whether it has a demonstrated ability to discriminate among individuals with this disease condition).

This last example shows clearly that, in assessing the validity of an evaluative index, we must more or less adopt a gold standard. It is absolutely necessary that we have a way of determining whether or not true change has or has not occurred as a direct result of an intervention. Here, there are two preferred alternatives: use an intervention that we know works or use another health status measure that is regarded as longitudinally valid, too.[5]

A corollarry of the need for a gold standard is the fact that simultaneously trying to validate an instrument and assessing the efficacy of a treatment or program involves a problem of circularity. Like many glimpses of the obvious it is an often overlooked trap. Examining score changes following an intervention of known efficacy is one means of validating an evaluative instrument. Likewise determining the efficacy of an intervention requires, inter alia, an evaluative instrument of know validity and, of course, depending on the nature of he intervention and the nature of the disease condition, accomplishing either task may require research findings that involve measurements made at more that two points in time.
]
Figure 3 illustrates the relationship among efficacy and validity and demonstrates two simple truths: if an improvement in self-reported health status is recorded using a instrument of unknown validity following an experimental intervention of unknown efficacy, there is no way of deciding whether the instrument is valid and the treatment is efficacious or it the instrument is not valid and the treatment is not efficacious. Similarly, if no change in self-reported health status is recorded there is no way of deciding whether the instrument is unresponsive and therefore not valid, given the treatment is really efficacious; or whether the instrument is or is not valid, given the treatment is not efficacious.

[5] A common approach is to use a transition index (eg. MacKenzie et al, 1986a, 1986b). Indexes constructed to reflect a single state ("How tired have you been? Very tired, somewhat tired or full f energy?) which are used to collect before- and after – measures of health status are more efficient than instruments that focus on transitions (eg. "How has your tiredness been? Better, the same or worse?") because changes are calculated rather than explicitly assessed. On the other hand, transition indexes avoid the potential problem of floor and/or ceiling effects which may occur with indexes that reflect a single important performance characteristic; changes for the worse in severely ill patients with low baseline scores should not go unnoticed (see Bindman, Keane & Lurie, 1990, for an example of this instrument bias).

CHANGE SCORES: TO BE AVOIDED LIKE THE PLAGUE?

Although a colleague did comment that many of her best (psychologist) friends have done PhDs on change scores, over the years psychologists have been warned repeatedly of the hazards of change scores. The upshot is hat many researchers now believe that the use of use of change scores is universally misleading and therefore ought to be avoided like the plague. For example, Cronbach and Furby (1970) are often quoted as stating that 'investigators who ask questions regarding gain scores would ordinarily be better advised to frame their questions in other ways' (p. 80). Needless to say, such advice is rather disconcerting from the perspective of researchers engaged in health program evaluation who conceive of their role as one of determining whether the intervention can, or does, work. On philosophical grounds, gain scores appear to be the natural measure to be obtained. Afterall, the goal of clinical research and a good many health care interventions is to induce changes in health status of patients or clients. On methodological grounds, too, change scores seem appropriate. Instruments that are responsive to changes in health status can reasonably be expected to be more sensitive measures of the effects of clinical interventions (Guyatt, Walter & Norman, 1989; MacKenzie, Charlson, DiGioia & Kelley, 1986). What's the story?

In order to understand the nature of the controversy surrounding the use of change scores, it is necessary to recognise that the measurement of change can be directed at different goals (Carver, 1974, Lin & Slade, 1977; Norman, 1989; Streiner & Norman, 1989). They can be described as follows:

1. To measure differences between individuals in terms of the amount of change experience. Although seemingly similar to the notion of discrimination, the aim is to identify individuals who gain unusually large (or small) amounts – perhaps so that these individuals may be given special treatment. Much of the literature in psychology (and educational psychology, in particular) that addresses the measurement of change accepts this as the goal of change measurement (however, see Carver, 1974). In health care contexts, comparisons of individual differences in change scores may be undertaken as part of a secondary analysis of a trial to identify who did and did not respond to treatment. However, this may well be only a means to an end. Very often the focus in measuring changes is not on individual difference scores per se, but on their correlates. Measures of change may be computed for individuals as a means to the end of finding variables that predict the amount of change. Ultimately, what's of interest is why some individuals or sub-groups changed a lot while others changed only a little.

2..To infer treatment effects from group differences. This is probably the rationale of most randomised controlled trials. This randomised controlled trial (RCT) win which patients are randomly assigned to one of at least two groups, is considered the most definitive method for evaluating the efficacy of health care interventions and technologies. By randomly assigning individuals to experimental versus control groups, measuring health status before and after treatment, and then comparing the average change in the helath status in each group, we can determine whether individuals in the experimental group change more, on average, than those in the control group.

These two goals regarding the measurement of change are antagonistic..[67] To the extent that there are individual differences in response to treatment, it will be more difficult to detect an overall treatment

[6] Carver (1974) offers a resolution o these positions for psychologists, in differentiating between psychometric and edumetric dimensions of tests. Thus, a test may be evaluated 'interms of its psychometric properties, that is, the extent to which it reflects the stable between-individual differences that traditionally have been of interest to psychological testing'; or 'interms of its edumetric properties, that is, the extent to which it reflects the within-individual growth that traditionally has been of primary nearest to educational testing' (p.512).

The problems that psychologists have with measuring change or gain always involve correlational, that is, psychometric models. For example, Cronbach and Furby (1970) define reliability of gain or difference scores as 'the correlation of the score with an independently

s Carver elaborates,

effect. The differences in change between individuals that are a precondition for identifying correlate of change (if there are no individual differences, the search is pointless) reduce the chance of demonstrating overall treatment effects (there's a lot of noise in the signal).

Note, however, that this does not imply a conflict between the goals of discriminating between individuals and evaluating change within individuals. The presence of large differences between individuals does not preclude the detection of small treatment effects. Demonstrating that an instrument is able to discriminate between individuals does not necessarily ensure or preclude its ability to detect changes due to treatment.

An important distinction between studies aimed at inferring treatment effects concerns whether or not the experimental and control groups are formed on the basis of random assignment. In what follows we assume that experimental groups are formed by randomly assigning individuals to experimental conditions.

There are basically four choices available for ananlysis of this design:

(1) Use a paired t-test to analyse the difference between pre- and post-test scores fo rthe treatment group only i.e.,

$$t = \bar{D}_{exp} / \sigma_{\bar{D}exp} \qquad (2)$$

Where D_{exp} is the mean difference over all pairs of measurements for the s individuals in the experimental condition, $(\sum(X_{i\cdot post\cdot exp} - X_{i\cdot pre\cdot exp})/s$ and σ_{Dexp} is the sampling variance of D, $\sum(D - D)^2 / s(s - 1)$. Although the 'difference method' takes into account the correlation between paired measure ments, it is not recommended because it uses a pre-experimental research design that is vulnerable to numerous threats to validity (Cook & Campbell, 1979).

(2) Use an unpaired t-test to compare the treamen tand control groups on the basis of their post-test scores only. The treatment effect is estimated by $(X_{post.exp} - X_{post.control})$. The error associated with this estimate is equal to the sum of the between-subject variance, σ^2_{error}, divided by the sample size.

'The gain score so a perfectly reliable edumetric test that everyone failed prior to treatment and everyone passed subsequent to treatment would have to correlate zero with any other variable, thereby rendering gains as perfectly unreliable in Cronbach and Furby's psychometric model. It seems reasonable to agree with Cronbach and Furby that gain scores on a psychometrically developed test would rarely be useful no matter how they may be adjusted or refined. However, it is absurd to suggest that gain scores on an edumetrically developed test are rarely useful.

'On reflection, it is hardly surprising that psychometricians have had so much difficulty with gain scores on tests. If a test is designed according to the best psychometric principles, then it would be somewhat serendipitous if the test scores reflected gain very well. If test scores that measure gain well are desired, then the test should be designed, developed, and evaluated with a focus on edumetric principles. Furthermore, the psychometrician who desires to measure and study gain must recognize hat psychometric statistics, such as variances and correlations, are likely to be inappropriate…….

'The distinction between the psychometric dimension and the edumetric dimension has always been understood intuitively by experimentalists. … The main reason the psychometrician has had problems with measuring gain or change is that gain or change means that some treatment condition is involved and therefore the psychometrician is entering the domain of the experimentalist. The psychometrician has tried to bring his psychometric test principles with him when he entered the experimentalist's domain, but these

[7] Note that Cronbach (1957) first identified the paradox as the distinguishing feasture of two strams of psychology – one attempting to explain differences between individuals (corelational psychology) and the other attempting to advance theory via experimentation (experimental psychology).

$$t = (X_{post.exp} - X_{post.control}) / \sqrt{(2/s(\sigma^2_{subjects} + \sigma^2_{error}))} \quad (3)$$

(3) Compare the difference scores to the treatment groups with those for the control group using an unpaired t-test on difference scores (or repeated measures analysis of variance). i.e.,

$$t = [(X_{post.exp} - X_{pre.exp}) - (X_{post.control} - X_{pre.exp})] / \sqrt{(2/s(\sigma^2_{error} + \sigma^2_{error}))} \quad (4)$$

Assuming the population values of the pre-test means are equal, the effect of the experimental treatment is equal to the difference in post-test scores as in equation above.

(3) If the variances between pre- and post-treatment administration of the instrument are equal the error in the estimate is twice the within-subject variance. Hence, the t-test can be written as:

$$t = (X_{post.exp} - X_{post.control}) / \sqrt{(2/s(\sigma^2_{error} + \sigma^2_{error}))} \quad (5)$$

(4) Use an analysis of covariance (ANCOVA). The analysis of covariance consists of a statistical adjustment of the scores on the dependent variable (here, subjects' post-test scores), based on knowledge of the corresponding scores on the control variable (subjects' pre-test scores). There are two basic adjustments: an adjustment for the chance differences in pre-test scores for subjects within each treatment groups and an adjustment for chance differences for the treatment groups (which is of relatively minor importance in randomised designs). The first adjustment is individual and is based on the deviation of each subject form the relevant group mean. It effectively addresses the phenomenon known as "regression to the mean," whereby individual scores that are very high or very low on pre-test move closer to the mean value on post-test. The second adjustment is constant for all subjects within a given treatment group and is based on the deviation of each group mean from the overall mean. The relationship used in the adjustment comes form the linear relationship between pre- and post-test scores, as described by a best-fitting regression line. A different regression line is estimated for each group and the adjustment is made on the basis of an average of the regression lines for the experimental and control groups. In effect, the analysis removes from the variability of post-test scores any variability that is predictable from a knowledge of the linear relationship between the pre- and post-test scores. Thus, the analysis of covariance focuses on residual deviations (i.e., variability not associated with the control variable), rather than deviation scores per se.

Having eliminated from consideration use of the paired t-test on change scores for the experimental group only, is there anything to choose between the remaining analyses? The short answer is yes. Use of change scores will result in an increase in statistical power vis a vis an analysis based on post-test scores only when the denominator of equation (5) is less than the denominator of equation (3). This is tantamount to saying that the use of change scores results in an increase in statistical power when the between-subjects variance exceeds the within-subjects variance exceeds the within-subjects variance i.e., $\sigma^2_{subjects} \geq \sigma^2_{error}$. In fact, we are specifying that

$$2\sigma^2_{error} \geq \sigma^2_{error} + \sigma^2_{subjects} \quad (5)$$

Or equivalently,

$$R = \sigma 2_{error} / (\sigma^2_{error} + \sigma^2_{subjects}) \geq 0.5 \quad (6)$$

Hence, the use of change scores to assess treatment effects is associated with greater statistical power whenever the reliability of the instrument exceeds 0.5. However, it is also true that the analysis of covariance provides unbiased parameter estimates and tests and is, in general, the most powerful of the three approaches. The extent of the gain in power with ANCOVA is a function of the magnitude of the correlation between pre- and post-test scores, which is, in turn, related to the magnitude of the variance between- versus within-subjects.

Is there ever a time when it is preferable to use difference scores? Again, the short answer is yes. There are two exceptions to the rule that the analysis of covariance ought to be preferred as the method of analysis due to its greater power (Maxwell & Howard, 1981). The first situation where analysis of change scores may prove useful occurs with a multivariate pretest-postest design. To the extent that the number of variables is large relative to the sample size, the use of multivariate analysis of variance (MANOVA) on the change scores may result in more power that the multivariate analysis of covariance (MANCOVA) because of the larger number of degrees of freedom associated with the former analysis.

The second case where analysis of change scores is indicated arises where response-shift bias occurs. Response-shift bias occurs. Response-shift bias refers to an instrumentation-related threat to validity. It occurs when the subject's conceptualization of the variable measured changes as an effect of the experimental manipulation (Golembiewsky, Billingsley & Yaeger, 1976; Hoogstraten, 1985; Howard & Dailey, 1979; Howard, Ralph, Gulanick, et al, 1979; Howard, Schmeck & Bray, 1979; Terborg, Howard & Maxwell, 1980). The internal standard the subject uses as a basis for pre-test and post-test responses changes. The upshot is that it's no clear what is being measured, a changed self-report or self-reported change. Moreover, because the change in conceptualisation of the variable begin measured (e.g., health-related quality of life, utility), and in the internalised standard on which the self-report score is based – the response shift – occurs only in subjects in the experimental condition, the bias is not eliminated with the use of a control group. Although the phenomenon of response shift has been explored primarily in organisational development contexts, there is every reason to think that it may apply to the measurement of health status (Breetvelt & Van Dam, 1991; Brow & Burrows, 1992) and have significant ramifications for the use of self-report measures in clinical trials and in cost-utility analysis.

To attenuate this response-shift effect, Howard and associates (1979b, 1979c) recommend that, at the time of post-testing, subjects rate themselves retrospectively with respect to their pre-intervention level of functioning. This notionally yields a set of ratings made with respect to the same evaluation standard. Howard et al (1979b) demonstrate that analysis of covariance using either the pre-test or retrospective pre-test as the control variable will produce misleading results, as would an analysis based on post-test scores only.

RESPONSIVENESS: DEFINITION AND MEASUREMENT

In general terms, as we have indicated, responsiveness is to an evaluative index what discrimination[8] is to a discriminative index. As index "dimensions," they are each concerned with whether "meaningful"

[8] As indicated in an earlier section, the reliability coefficient (i.e., the ration o variance between subjects to total variance) is a generally accepted overall measrure of the ability of an instrument to discriminate among individuals. We refer instea to a property we have labelled 'discriminability' to underscore its significance with a discriminative index, and to divert attention from the overturned dictum that 'reliability

differences are, in fact, measured. They lie at the root of the debate about the relative merits of generic versus disease-specific measures.

Discrimination refers to the spread of scores or the number of categories into which individuals can be placed and the degree to which it is present is central to the validation of instruments that aim to define cross-sectional differences among individuals (Bergner & Rothman, 1987; Kirshner & Guyatt, 1985). Discrimination is very much a product of item selection, item reduction and item scaling processes, as indicated in Table 3. A crude scale will allow individuals to be placed in only a small number of categories whereas a scale with finer discrimination will allow a greater range of scores. In general, the fineness of the discrimination achievable will depend on the number of items per health dimension and/or the number of categories within a dimension. Crude scales will be adequate to detect differences among groups of individuals when the expected difference is large and this points to an interaction between the desired level of scale discrimination and the intended sample. Noncongruence between the level of health status assessed and the target population will result in a skewed distribution of health status scores i.e., a large proportion of individuals will receive the same or similar scores such that the detection of meaningful differences is problematic (Bergner & Rothman, 1987).

As with discrimination, responsiveness is very much a product of item selection, item reduction and item scaling processes – except that the accent is on measuring within-person change over time rather than on measuring between-person differences. Table 3 juxtaposes the requirements of an evaluative index with those for a discriminative index. Changeability will depend on the number of response options per item, the number of items per health dimension and the number of categories within a dimension. The criterion governing the selection and reduction of items is the likelihood that an individual's health status will change as a result of the application of an intervention. Clearly, items that are unresponsive – either unreliable in a test-retest sense with stable subjects or not sensitive to change in individuals whose health status is not stable – should be deleted. Beyond this, deciding how many items to retain or delete is less straightforward than with discriminative indexes. The usual procedures for measuring internal consistency, KR20 and Cronbach's alpha (Carmines & Zeller, 1979; Cronbach, 1951; Kuder & Richardson, 1937), assume that the precision of the index will increase incrementally with the covariance of the items and the number of items included. It is not clear that these assumptions are appropriate for evaluative indexes. In the first place, items in an evaluative index need not be correlated at a single point in time. Rather, they should be consistent in the way that they measure change in health status over two points in time. Secondly, regardless of whether they are correlated, the greater the number of items included in an index, t he greater the probabi8lity of including items that may prove insensitive to efficacious treatment and, as noted above, any variability in item scores that is not related to the intervention may obscure any treatments effects.

The issue of responsiveness is fundamental to the debate about whether to use so-called specific instruments, rather than generic instruments, in clinical research, Up to a point the relationship between the two approaches is a matter of the level of operationalization, as Spitzer (1987) indicates in Figure 4. This shows three levels of operationalization of health-related quality of life and functional status, subsumed by a general health status concept, which give rise to a hierarchy of possible data gathering instruments. It is significant that specific symptoms associated with specific diseases are regarded as targets for hypothesis-determined measures.

The rationale for disease-specific measures is the increased responsiveness to disease-specific interventions that may result from the inclusion of those aspects of quality of life that are of particular

is a necessary but not a sufficient condition for validity' which sometimes seems to get in the way of appreciating the different emphases and requirements of discriminative and evaluative indexes.

concern or relevance to patients being studied. In its simplest form, the idea is that the more focused the instrument, the more responsive it is likely to be. However, this assumption cannot be made on the basis of ipso facto reasoning and the issue has to be resolved empirically. Generic measures may be just as useful and responsive in particular settings as specific measures but, to date, there have been very few head-to-head comparisons. This is not altogether surprising when we consider that we do not yet know exactly how to quantify responsiveness.

Apart from practical barriers to the use of generic measures (including respondent burden and the resource requirements of interviewer-administered instruments), which are no doubt real, there are other reasons why specific measures seem to be preferred by physicians. These relate to the clinical salience of specific measures, and more particularly, to the content validity of measures and the meaningfulness or interpretability of change scores.

Specific measures have the advantage that they have high face validity and content validity for physicians; they appear to be clinically sensible as they relate closely to signs and symptoms and areas of functioning that are routinely explored by physicians. In contrast, generic measures sometimes have low content validity for patients and physicians alike. They may contain items that are of little or no relevance to the study population and that add to respondent burden without contributing to responsiveness (e.g, if only the upper limbs are affected by arthritis, responses to items pertaining to say, mobility, cannot register improvement because lower limb function is not impaired). On the other hand, it is not always possible to specify in advance the impact that a treatment or intervention will have on the patient's health status: there may be unanticipated positive or negative side-effects and physicians may make incorrect assumptions about the nature and severity of clinical symptoms that have an adverse impact on patient well-being.

This said, we should also say that validity is first and foremost concerned with the inferences that can be made on the basis of test scores. Content validity focuses on instruments rather than measurements; it is concerned with the imputs to the measurement process rather than the outputs of the measurement process. To the extent that content validity is not concerned with score-based inferences, it should not be allowed to loom large in the calculus when choosing between selecting specific versus generic health status measures.

MEASURES OF RESPONSIVENESS

How should the magnitude and meaning of changes in health status be assessed (and reported) in such a manner that the performance of different instruments can be compared? How should responsiveness be quantified? There is no shortage of candidates[9] (although as it happens most of the methods advocated are variants of effect size).

(1) paired t-statistic (for within-subject changes): A number of researchers have proposed examining score changes following an intervention of know efficacy. Score changes between "pre-test" and "post-test" measurements would provide evidence of responsiveness, and the instrument with the largest paired t-statistic (or F-ratio) for within-subject changes would be judged the most "sensitive". As an example, Liang et al (1985) compared several instruments in patients undergoing total hip replacement.

[9] We should not parenthetically that the reliability of the change score, $R(D) = s^2_D/s^2_D + s^2_{error}(D)$, is not included among this list of contenders. It is not an appropriate index for assessing the ability of an instrument to measure treatment effects. (However, R(D) can be used to assess the ability of an instrument to detect individual differences in change scores). The reason is straightforward. In the event of a uniform response to treatment, the ideal situation for the use of change sores to measure treatment effects, the between-subjects variance of change scores is zero (since every patient's post-test score is equal to his or her pre-test score plus a constant) and R(D) is zero.

Unforturnately, this method does not account for the score variability that may occur in apparently stable subjects (e.g., systematic score changes due to learning effects) and may vary from instrument to instrument. Moreover, the comparison of statistical tests is, in general, not advised because the magnitude of the test statistic is influenced by sample size.

(2) correlations between changes: Correlating change scores for measures of health status with changes in physiologic measures is another method that has been suggested to assess responsiveness (Meenan, Anderson, Kazis, et al, 1984). More generally, it entails correlating change scores for alternative patient outcome measures. A limiation of this method is its vulnerability to any artifactual change inherent in pre- to post-test scores that accrues due to regression to the mean and errors inherent in the process of deriving differences between any two measures (Buetler & Crago, 1981). Moreover, it is neither concise nor parsimonious as an approach to measuring responsiveness since it requires a multitude of pair-wise comparisons.

(3) effect size: Effect size has been suggested as another method for comparing the responsiveness of competing instruments. As operationalised by Kazis, Anderson and Meenan (1989), the effect size relates changes in mean score (from baseline to follow-up) to standard deviation of baseline scores (i.e., baseline scores are used as a proxy for control group scores). The largest possible effect size index for a non-negative measure is given by the reciprocal of the coefficient of variation at baseline i.e., mean score at baseline divided by standard deviation at baseline.

Although a number of different approaches can be used for calculating effect size, the Kazis et al approach allows effect size to be generated separately for experimental and control groups. These estimates can then be used as separate benchmarks for more important and less important changes, respectively. Kazis et al argue that the effect sizes can be used: to provide general and instrument-specific benchmarks for interpreting therapeutic change, to relate health status changes to changes in more familiar clinical measures; and as a bridge to compare outcomes across studies (e.g., to calibrate the potency of treatments). For the social sciences, at least, where the calculation of effect sizes is commonplace, Cohen (1977) defines an effect size of 0.20 as small, one of 0.50 as moderate, and one of 0.80 or greater as large.

(4) Guyatt's responsiveness measure: Guyatt et al (1989) have proposed a variant of effect size viz. The minimum clinically important difference (MCID) divided by the standard deviation of score changes among stable subjects (or, if more than two replicate observations are available, the square root of two times the mean square of the error term), determined separately. The minimal clinically important difference is defined as that difference in score on a health-related quality of life instrument that corresponds to the smallest change in status that patients consider important. This measure therefore tries to maximise the ratio of signal (the real change that has occurred) to the noise (the variability in scores that is not associated with true change in status).

An advantage of quantifying responsiveness in this way is that it incorporates the two parameters that determine the sample size required for any experiment in which the endpoint can be specified as change over time in test scores and for which pre- and post-test intervention scores are available. If the MCID is known, this approach would also yield a responsiveness coefficient characteristic of an instrument, analogous to a reliability coefficient. For now, the practicality of determining the MCID for each class of patients and instrument combination looms large as a limitation of "Guyatt's responsiveness measure" (however, see Jaeschke, Singer & Guyatt, 1989). Norman (1989) has also expressed reservations about the assumption that the variance in stable subjects in the experimental condition who will change as a result of intervention.

(5) standardised mean response: the standardised response mean (SMR) (Liang, Fossel & Larson, 1990) is a variant of effect size, defined as (mean response)/(standard deviation of responses), that equals the paired-sample t statistic (except for a sample size factor). The SMR has the same numerator as Kazis et al's "effect size index" but uses the standard deviation of responses in the denominator. An advantage of the SMR over the effect size index is that incorporates information about response variance and can therefore be used to test the statistical significance of response means.

(6) responsiveness coefficient: The responsiveness coefficient (Norman, 1989; Norman & Streiner, 1989) has been advocated as a way of facilitating comparison between responsiveness to change, and reliability, expressed as a ratio of variances. It is defined as follows:

$$\text{Responsiveness} = \sigma^2_{\text{change}} / \sigma^2_{\text{change}} + \sigma^2_{\text{error}} \quad (7)$$

Estimates o f the variance components of equation (7) depend on the analytical strategy used (ie., ANOVA versus ANCOVA). For example, in a standard two-factor experiment (e.g., treatment versus control groups) with repeated measurements on one factor (e.g., occasions)(Keppel, 1973) the error variance is the same as the error variance calculated for reliability. The equivalent change variance, σ^2_{change}, is calculated as $(E(MS)_{\text{treatment x occasion interaction}}) - MS_{(\text{error})}) / s$, where s is the number of subjects per treatment condition (Glass & Stanley, 1970).

(7) ROC curve analysis: Deyo and Centor (1986) argue that assessing the responsiveness of health-related quality of life instruments is analogous to assessing the discriminating properties of diagnostic tests. They propose that an instrument's responsiveness be described in terms of its sensitivity and specificity in detecting improvement or deterioration inpatients, as established by an external standard. For example, any given score change (e.g., say, one, five or ten points on a 100-point utility scale or health profile) will include some true positives (indicating a real change in patient status) and some false positives (nonspecific effects in the face of no real clinical change). Information on sensitivity and specificity can be synthesized in a ROC curve which plots sensitivity against (1 – specificity) for each of a number of possible "cut points" in change score. Thus, sensitivity and specificity are calculated for a change score of one point, two points, and so on. A change score of one point is likely to be highly sensitive (most patients who have improved would have a score change of this magnitude or greater) and quite nonspecific (many unimproved patients may register a change score of this size). By comparison, a change score of five points would be less sensitive (many patients who have improved according to the external criterion may not have changed this much), but more specific (not many unimproved patients would have a change score of this size). In this setting, the area under the ROC curve can be interpreted as the probability of identifying the improved patient from randomly selected pairs of improved and nonimproved patients. Calculation of this area provides a means of comparing alternative instruments for assessing health-related quality of life which is formally equivalent to the effect size.

At present, the only point on which there is a clear consensus is that standardised indexes of responsiveness ought to be reported so that meaningful comparisons can be made among evaluative instruments. However, the resolve to do so is weakened by the number of alternatives proposed and, perhaps, by their relative unfamiliarity. On grounds of number-crunching practicality and communicability, we are inclined towards the standardised response mean and the responsiveness coefficient. They seem to have fewer drawbacks than the other measures on offer and are rooted in analyses that are undertaken to infer treatment effects from group differences.

UTILITY MEASUREMENT AND CLINICAL TRIALS

Within Guyatt et al's (1989) taxonomy of health-related quality of life measures, there are two approaches to utility measurement. The first approach is to classify patients into categories based on their responses to a number of questions about their level of functioning. For example, the Quality of Well-Being Index classifies patients as belonging to 1 to 43 combinations of levels on the basis of what, for health reasons, the patient did or did not do within 3 areas (mobility, physical activity, and social activity) each of which has 4 or 5 levels of performance. Each state of health is then valued using preference weights obtained from individuals in the general population, modified by the presence or absence of a standard list of similarly valued symptoms. The overall QWB places patients on a utility scale that ranges from 0 (dead) to 1 (healthy). The Rosser Index, as operationalised using the self-administered questionnaire developed by Gudex and Kind (1987), uses a similar approach except that individuals are classified according to levels of disability and distress, and negative utility scores are permissible (to allow for states worse-than-death).

In the second approach to utility measurement, as it applies to clinical trial settings, a single rating of all aspects of health-related quality of life is elicited on the basis of the patient's actual health state using one or more of the following scaling methods: rating scale, magnitude estimation, equivalence or person trade-off, standard gamble, time trade-off or willingness to pay (Froberg & Kane, 1989, Haig, Scott & Wickett, 1986; Nord, 1992; Rosser & Kind, 1978 Thompson, 1986; Torrance, 1986, 1987). Although a number of standardised, hypothetical ("marker") health state descriptions may be used with a view to helping to orient the patient to the task (Bennett et al, 1991; Drummond, Mohide, Tew, et al, 1991; Mohide, Torrance, Streiner, et al, 1988; Torrance & Feeny, 1989), it is only the utility value of the self-state that is of real interest (Mohide et al, 1991). The "marker" health states serve as a training device and a framework within which the subject can define his/her health state and become acquainted with a range of levels of well-being.[10]

Health-state descriptions are supposed to be comprehensive, though, in practice, the level of detail varies greatly from one study to the next (Capewell, 1988; Froberg & Kane, 1989b). They may include information on levels of functioning – sensory, physical, socio-emotional, cognitive and self-care – as well as on the condition or health problem itself, any side-effects of treatment, and the level of pain (Feeny & Torrance, 1989).

The approach aims to capture, in one utility-based measure of overall well-being, the patient's tradeoff between the benefits of treatment and any associated side-effects. The patient is invited, as it were, to do his/her own preference weighting by integrating all the relevant characteristics into a single utility value. As Bennett et al (1991,p. 122S) put it: 'The measurement process aims to answer the question: Does the patient prefer his or her current health status to the one before therapy, and if so, by how much?'

Analysis of the published literature indicates that so far here have not been many clinical trials which have included utility measures among their outcomes.[11] This is perhaps not surprising given that health –related quality of life measurement in the broadest sense of the term is still not widely done (Guyatt et al, 1989; Veldhuyzen Van Zanten, 1991). Because trials including utility measurement are few and far

[10] If a health state description is (appropriately) regarded as a construct, there is also reason to provide a health state description to facilitate the interpretation of results. This does not seem to be widely understood by health economists. Capewell (1988) is an exception. She notes, 'when the relevant health states for utility measurement are simply those of the patients themselves involved in the study, the individuals can provide a utility measurement for their own health state. At first sight it would seem unnecessary in this case to provide a health state description; however, to enable others to interpret the results health state descriptions may still be required' (p.50).

[11] Drummond and Davies (1991) and Feeny and Torrance (1989) refer to a number of trials that are either underway or for which formal reports are pending.

between we broadened our search to include studies using quasi-experimental research designs. Relevant aspects of the trials for which details are available and of other studies that include pre- and post-test measures of utility are summarised in Table 4 under a number of headings: disease condition/ patient category; treatment/intervention; utility measure(s) used; and the kinds of evidence presented and conclusions drawn regarding responsiveness.

What conclusions emerge from this exercise? The first thing that should be said about Table 4 is that it stands as testament to the parlous state of reporting of responsiveness. The summary table we would have hoped to present would have catalogued the responsiveness of the Quality of Well-Being Index (QWB), time trade-off (TTO), standard gamble (SG) and willingness to pay (WTP) approaches to utility measurement, using one or more of the measures discussed in the previous section. However, with a few notable exceptions, no data on the responsiveness of the specific and generic instruments used in the studies cited, are presented. This is rather disappointing given that the debate about the comparative usefulness of the two classes of instruments an only be resolved empirically. Moreover, for the most part, it is not possible to calculate even the most pragmatic of the measures of responsiveness described in the previous section from the data reported. Indeed, the paucity of evaluative studies that have used utiltiy measurement and the lack of comparative data, across measures and clinical settings, makes it difficult to see how one can conclude, as some have, that utility measures are responsive. At best, we were able to calculate the effect size index from the reported mean difference and baseline standard deviation data.

However, a number of points to ponder do emerge from a review of the studies cited. One relates to the interpretability or meaningfulness of changes measured. This is very much a question of validity: making score-based inferences presupposes that the scores are interpretable. The issues is real. Feeny and Torrance (1989) recognise it when, in cataloguing the advantages and disadvantages of utility measures, they note that 'utility scores may not be readily understood or easily interpreted' (p. S192). Bennett, Torrance and Tugwell (1991) are also aware of this problem, commenting that 'the usefulness of utility measures will be greatly enhanced by strategies to make utility scores meaningful' (p. 126S0. But how?

There is no widely accepted notion of what represents a clinically meaningful change in health status or utility. Once upon a time, the same could have been said in respect of changes in many traditional clinical measures, like blood pressure and forced expiratory volume (FEV1), that physicians now regard as self-explanatory. The difference is that such measures are now familiar and can be interpreted in terms of well-established or agreed cutoff points that have become linked to concrete performance-oriented outcomes through accumulated experience. Paterson's (1988) observations in respect of the findings of the trial of auranofin therapy for the treatment of rheumatoid arthritis (Bombardier, Ware, Russell, et al, 1986) emphasise the contribution of such linkages. To quote him at length:

'to be meaningful a quality-of-life instrument should further a judgement as to the practical importance of the score observed. The traditional measures do not do this, since their units of measure, such as millimeters of mercury or seconds of walk time, have little meaning in the context of daily life. For example, it is probably only those rheumatologists experienced in the use of traditional measures and the literature about them who would know how the grip strength of healthy 16 year-old boys, say, compares with that of 60 year-old women, let alone how grip strength may be expected to change with different therapies. The global measures' units have no bridge whatever to concrete experience, so that the therapeutic importance of a score change is unknowable by itself. Only through repeated correlation of the results with other concrete experience, so that the therapeutic importance of a score change is unknowable by itself. Only through repeated correlation of the results with other concrete results can the global measures take on meaning. Of the simple, scalar type measures perhaps the 10-

centimetre Pain line comes closest to having built up a meaningful framework of such correlations. This argues for quality-of-life instruments with component items based on performance, since these have meaning in terms of common experience. Indeed, the score changes on the HAQ (Health Assessment Questionnaire) and QWB (Quality of Well-Being Index) can be expressed in terms of change in performance of a single daily act, and the imporatance of the ability to performance of a single daily act, and the importance of the ability to perform that act or not can be reasonably judged. While the PUMS (Patient Utiltiy Measurement Set) does not have performance items, its score appears amenable to the same kind of translation into experientially meaningful units – for example chances of death or years of life.

'Unless concrete equivalents of the QWB or PUMS scores are eked out, the practical importance of a score change requires the same kind of framework of prior correlational experience that any other unfamiliar measure requires. That the two measures each employ a 100-point continuum and are each anchored by death and full health helps orient the lay evaluator but does not communicate the practical importance of, say, a + 16 point change – even if expressed as percentage of full health (pp. 186-187).

As Deyo and Patrick (1989, p. S259) observe, without the necessary bridges to concrete experience, 'many will wonder if a given absolute change is equally important at different points of the health status

In their write-up of the Auranofin trial's findings, Thompson et al (1988) considered it necessary to explain what it meant, in terms of patient functioning, to secure an overall gain across all components of the QWB scale of 0.020 points. They judged that this level of improvement to be 'about equivalent to all AF patients improving, identically, on the subscale of physical activity form moving one's own wheelchair without help to walking without physical limitations (a gain of 0.017 points), if no other components of the QWB scale had been affected' (p.40). Although we may quibble about the usefulness of this "translation" (for individual patients who are likely to have different priorities and for physicians who may not agree on what constitute clinical important changes (Kirwan, Chaput De Saintonge, Joyce & Currey, 1984)), the problems are more than palpable where global utility scores are elicited. This is especially so if, as Revicki's (1992) findings indicate, only 25-27% of the variance in utility scores is explained by health status measures.

Another interpretability question that crops up in evaluative studies is whether "a change is a change is a change." It is generally recognised that statistically significant change over time is not necessarilly synonymous with clinically significant change (Feinstein, 1987b; Kazis et al, 1989; Tugwell, Bombardier, Buchanan, et al, 198). What's clinically significant may not be statistically significance, and vice versa. As Liang and Robb-Nicholson explain,

'Function is relative. Psychometrically sound instruments assume that function can be measured with the same standard instrument. However, a patient's function is relative to their age, sex, motivational social supports, priorities, goals and his or her needs for daily living, work, and recreation. …Because function is relative, a small change in an individual's function may make a lot of difference. The small change may be totally adequate for the person's needs, and yet not meet statistical norms' (pp. 580-1).

Changes in disease-specific may be easier to interpret because they are more familiar, more specific, more precise, or more closely linked to changes in clinical measures of disease activity (Deyo & Partrick, 1989; Guyatt & Jaeschke, 1990; Patrick & Deyo, 1989). As a result, there may b a greater degree of consensus among physicians about how to interpret and communicate the prognostic implications or practical importance of observed change scores for disease-specific instruments. Determining what it is we can or should conclude about the effects of treatment on the basis of changes

in utility scores is far from straightforward. Indeed, it seems an order of magnitude harder. Bennett et al (1991, p. 123S, emphasis added) acknowledge as much in commenting that, 'although an individual utility has meaning in that it is defined on a scale form perfect health to death, the interpretation of differences between utility scores is controversial…. The current approach is arbitrary with the minimum difference being defined as 0.1 on the utility scale of 0 to 1.'

This issue arises in the trial of a support program for caregivers of the demented elderly (Drummond et al, 1991; Mohide et al, 1990) where caregivers' depression and anxiety scores (as measured by the Center for Epidemiologic Studies Depression Scale (CES-D) and the state-anxiety portion of the State-Trait Anxiety Inventory (STAI)) are more or less constant over the courseof the 6-month trial but their quality of life, as measured by the Caregiver Quality of Life Instrument (a modified time-tradeoff technique with marker states), improved in the experimental group, but decreased in the control group. The 11-point end-of-trail difference from baseline scores in CQLI-measured utility for the experimental group is judged to be clinically important but not statistically significant. In fact, as the researchers indicate, the actual number of subjects per group (30) falls well short of the improbably large sample size of 178 subjects per group which is estimated to be necessary to achieve statistical significance for this clinically significant difference.

Thus, the findings are difficult to interpret. Why did the caregivers' CQLI scores change when their CES-D and STAI scores did not? Are the two latter instruments useful as discriminative indexes but unresponsive in a n evaluative context? In evaluating the effectiveness of the caregiver support program, should our conclusion be based only on the treatment effect measured by the CQLI, given that the two principal outcome measures did not register score differences from baseline? Is it reasonable to calculate the cost-per-QALY gained for this intervention on the basis of a treatment effect that is clinically important but not statistically significant? Is it reasonable to do so when, if the clinically significant change in the CQLI that did occur had been declared impractical? Because we know so little about the responsiveness of specific and generic instruments alike at this juncture there are lots of questions and not many answers.

A different set of issues is raised in the context of the auranofin trial (Bombardier et al, 1986, 1991; Paterson, 1988; Thompson et al, 1988) wherein the favourable effect of this agent on patients' quality of life was confirmed for a broad range of clinical, health status, functional, pain and utility measures. They concern QALY calculations. In effect, Paterson (1988) poses the following question: "If $300 produces a gain of 2.4 per cent of full health on the QWB by month 6, as in the case of auranofin, can we say that a half-QALY has been added, at a cost of $12,000 (ie, $300/0.024), and that a full QALY can be cost at $25,000?"

He answers it in the negative. The reason is reflected in the plot of QWB scores over the course of the trial (see Figure 5). The impact of auranofin is delayed: there was no significant QWB gain until after month 2, and the gain of 0.024 is not achieved until month 4. To generate 2.4 percent of a QALY the 0.024 gain on the QWB for months 4 to 6 would have to be experienced for a full year. The average QWB score for the first 6 months can, of course, be calculated from the data obtained during the trial, but it cannot be extrapolated to the second 6 months.[12] The delayed onset of activity would not repeat itself, and there may be either further improvement or a worsening of QWB scores in months 7 through 12. And what of the second and subsequent years of auranofin treatment?

[12] Here we are at odds with Drummond et al (1990). They reason thus: 'the improvement in the quality of life of caregivers, as measured by the CQLI, was 0.11 over the 6-month intervention period. Over 1 year, this would be equivalent to 0.11 QALYs gained. Therefore, the implied incremental cost per QALY gained is $2,204 divided by 0.11, or $20, 036' (p. 216).

The studies by toevs et al (1984) and Liang et al (1990) also suggest that the period over which response to treatment is measured is important when calculating QALYs. For example, Liang et al's analyses of early (re-operative to 3 months), late (3 months to 1 year) and net (pre-operative to 1 year) responses to knee or hip arthroscopy indicate that patients who register a large early improvement may plateau or even regress, and that patients who have not improved initially may still do so later. And what of patients who drop out during the course of the trial? Depending on the distribution of early and late improvement across patients, quite different cost-per-QALY values could emerge. For many major procedures and treatments, this question becomes more important if only vecause the further away in time from the intervention, the greater would be the probability that other confounding factors would intervene.

The question of what period cost-per-QALY comparisons are relative: a given ratio can be deemed to be high or low only in comparison with the rations for other treatments. The issue is most transparent in the case of particular clinical trials involving repeated measures at specified intervals but can be raised more generally. For example, should a patient in a given health state with say, 20 years of remaining life expectancy be required to specify how may years in total he or she would trade-off for perfect health, or how many years he or she would trade-off for each successive increment of 5 years? Would it make any difference to the cost-per-QALY value? And if so, would it make any difference to the resource allocation decisions that might be made on the basis of cost-per-QALY comparisons?

Paterson (1988) raises another problem in connection with the calculation of cost-per-QALY ratios from clinical trials. To the extent that effect of the treatment is based on the difference in response between the experimental and control groups the cost-per-QALY ratio is likely to be a function of the control agent. Thus, the incremental cost and the incremental QALYs may well depend on whether the alternative treatment is a placebo, an active agent that is only marginally less effective than the treatment in question, or an alternative treatment modality (eg., surgery). Needless to say, the dilemma is made more complex when treatments are considered as complements rather than substitutes. Table 5, which is based on a hypothetical example presented by Paterson, indicates the range of cost-per-QALY ratios that could emerge from this exercise. Drummond et al (1991) finesse the problem by using as their measure of effect the increase in QALys from baseline in the experimental group. Is this appropriate? The issue is open for debate.

Finally, the trial of zivoduvine (ZDV) (Wu et al, 1990) raises the issue of whether the differential mortality of experimental- versus placebo-condition subjects should be reflected in the treatment effect on which the calculation of cost-per-QALY ratios may be based. One of the advantages that is often claimed for utility measures is that they allow us to integrate morbidity and mortality effects (eg., Feeny & Torrance, 1989). But should the zeros entered for patients who died during the blinded trial count when we compute the cost-per-QALY ratio? On the one hand, doing so would seem to capture the average experience of the cohort of patients who entered the trial in utility terms. On the other hand, assuming that ZDV would become the treatment of choice once its efficacy is established (absent any6 alternatives), ZDV may be handicapped unreasonably in cost-per-QALY league-table comparisons if higher mortality in the control group is reflected in the denominator of its cost-per-QALY ratio. Afterall, the implications for resource allocation would stem from its acceptance as (for argument's sake) the treatment of choice. The dilemma seems more pronounced with a placebo-control trial but, at least sometimes, a placebo-control trial may be a management trial rather than an explanatory one. Again, we are back ot he issue of how to measure effect.

DISCUSSION

Up to a point w have some sympathy with the reader of redoubtable stamina who reaches this point and enquires, “But what has all this to do with health economics?’ Our candid response would be that health economists (or at least, those who engaged in economic evaluation, in general, and cost-per-QALY comparisons as a basis for resource allocation, in particular), whether they realise it or not, have brought the issues we have raised upon themselves. The need for utility measurement methods to perform a discriminative function is rooted in the resource allocation motivation for cost-utility analysis which equires comparisons to be made across health care interventions and programs. Activity to this end is reflected in the construction of league tables form published cost-utility studies (Kaplan &Bush, 1982; Kaplan & Anderson, 1990; Torrance & Zipursky, 1984; Williams, 1985). Incongruously, however, the evaluative function has received much less attention. Nonetheless, it is quite explicit when the benefit associated with a program is identified with the difference between the area under two curves which plot the utility a patient derives from particular health states over time, with and without a given health program (eg. Fanshel & Bush, 1970; Rosser, 1990; Torrance & Feeny, 1989). Many, probably most studies, do nothing more than calculate the utility of a given health state cross-sectionally and multiply it by remaining life expectancy.

The comparative advantage that health economists bring and the circumstances under which it is reasonable for helath economists to have a foot in the door with respect to the cost assessment side of clinical trials (Drummond & Stoddart, 1984; Mugford & Drummond, 1990) are quite clear cut in comparison to that for the benefits assessment side. Certainly, economics can contribute to a clinical trial whenever the focus of the trial is explicitly about economics of patient management (or sometimes its synonym, cost-shifting), and prospective economic evaluations of therapies that are not yet incorporated into generally accepted treatment regimens may allow policy-makers to make proactive decisions about whether they are worth buying.

Whether health economists can contribute much on the benefits side is not yet proven. In the main it will depend on whether utility measures of health status are useful evaluative and discriminative indexes. This will require a considerable amount of validation work and much better documentation of responsiveness, in particular. The same can be said of other approaches to the assessment of health-related quality of life. Thus, in many respects, we regard
Drummond et al’s (1989) assertion that health economists ought to be involved in the assessment of benefits which are not well captured by normal clinical measures quality of life and require direct assessment of quality of life as more in the nature of an ambit claim.

Indeed, in some respects, utility measures are at a comparative disadvantage in application to clinical trials. To begin with, measurements of utility are not highly precise (Drummond & Davies, 1991; Feeny & Torrance, 1989). The standard deviation of a single measurement of utility is in the order of 0.2 to 0.4 for patients experiencing a given health state (Drummond et al, 1991; Torrance, 1986). This means that clinical trials with concurrent cost-utility comparisons will require larger sample sizes, with consequent resource implications for the conduct of the trial. The incremental costs of including utility measurement may therefore limit the scope of economic evaluations to piggy-back on clinical trials. Further, where a major objective of the trial is to assess how particular patients should be treated, utility measures are disadvantaged to the extent that they do not have clinical meaning. This limitation need not be fatal, of course, because other specific and generic measures of health-related quality of life may be included but, in general, if utilities are to be acceptable measures of outcomes in trials, there must be a defensible mapping of scores onto clinically meaningful indicators. The fact that the clinical significance of utiltiy differences is not easily interpretable does diminish their marketability to trial organisers (especially when there is no regulatory coercion to consider cost-effectiveness). Use of direct utility

measurement techniques like the time trade-off are further impeded to the extent that clinical investigators consider that the nature of the measurement process may cause patients unnecessary psychological distress and would therefore be unethical.

Beyond this, it is not clear what health economists gain by eliciting the views of patients enrolled in a trial rather than patients suffereing from the condition, a cohort of potential patients, or members of the general public. In principle, data collected in clinical trials at baseline and at intervals using health profiles and disease-specific instruments might usefully inform the construction of health state descriptions or scenarios.[13] They would at least permit the development of health state descriptions that have some standing as constructs in their own right and as objects of validation. It would also provide a basis for refining "marker states" presented to orient subjects to the utility measurement task. Setting all health dimensions to low, medium and high levels in marker states, as happens at present, may indicate more about transitivity of subjects' preferences than the reliability of utility measures. Health state descriptions that vary the levels of different dimensions would provide a much stronger test.

Ultimately, then, this paper raises more issues than it resolves. However, we would argue that they are issues about which health economists engaged in cost-utility studies should be concerned. In the final analysis, even if utility measurements can be shown to be cross-sectionally and longitudinally valid, whether they should be included in clinical trials will depend on whether the extra cost of obtaining and using utility values is judged to be cost-effective in and of itself (Drummond, Stoddart & Torrance, 1987). At pragmatic level, this criterion may resolve into a question of how the allocation of resource would be affected if policy-makers were to base their decisions on cost-effectiveness rather than cost-utility analyses. This translates into a question of whether what's in the denominator makes a difference to the relative ranking of health care programs and interventions. Whether it does is, of course, an empirical question and the answer may depend on which approach to utility measurement is used. Preliminary work, focusing on one widely accepted approach to QALY measurement, the Rosser Index, and based on published data on the distrubution of pre- and post-intervention health states, suggests that we may be playing with "decimal dust" (Brown, Burrows & Macarounas-Kirchmann, in preparation).

[13] To attempt to map condition-specific outcome measures onto directly onto utility scales a s Caairns, Johnston and McKenzie (1991) propose seems altogether more problematic. Their assumption that 'individuals would be indifferent between all possible scenarios that share a common aggregate score' is, as they admit, a very strong one. It does not sit well with one of the advantages claimed for utility attaches to his or her overall health status. Afterall, one of the major shortcomings of currently available health-status instruments is that they include items included may not accord with patient outcome priorities; some of the items included are of great importance to the patient and others are inconsequential (Tugwell, Bombardier, Bell, et al, 1991).

References

Adams ME, McCall NT, Gray DT, Orza MJ, Chaimers TC (1992). Economic analysis in randomized control trials. *Medical Care* 30:231-243.

Banta HD (1 987). Clinical and epidemiological research on new technologies: a role for economic appraisal? In MF Drummond (Ed), *Economic appraisal of health technology in the European community.* Oxford: Oxford University Press.

Bennett K, Torrance G, Tugwell P (1 991). Methodologic challenges in the development of utility measures of health-related quality of life in rheumatoid arthritis. *Controlled Clinical Trials* 12(Supplement):1 18S-1 28S.

Bergner M, Bobbitt RA, Carter WB, Gilson BS (1 981). The Sickness Impact Profile: development and final revision of a health status measure. *Medical Care 19:787-* 805.

Bergner M, Rothman ML (1 987). Health status measures: an overview and a guide for selection. *Annual Review of Public Health* 8:191-210.

Beutier LE, Crago M (1 981). Self-report measures of psychotherapy outcome. In Lambert MJ, Christensen ER, DeJulio S (Eds), *The assessment of psychotherapy outcome,* 453-497. New York: Wiley.

Bindman AB, Keane D, Lurie N (1 990). Measuring health changes among severely ill patients. *Medical Care* 28:1142-1152.

Bombardier C, Raboud J, and the Auranofin Cooperating Group (1 991). A comparison of health-related quality-of-life measures for rheumatoid arthritis research. *Controlled Clinical Trials* 12(Supplement):243S-256S.

Bombardier C, Ware J, Russell IJ, Larson M, Chaimers A, Read JL, and the Auranofin Cooperating Group (1986). Auranofin therapy and quality of life in patients with rheumatoid arthritis: results of a multicenter trial. *American Journal of Medicine* 81:565-578.

Breetvelt IS, Van Dam FSAM (1 991). Underreporting by cancer patients: the case of response shift. *Social Science & Medicine 32:981-987.*

Brown K, Burrows C. Self-reported change or changed self-reports? Biases in measuring the effects of health care interventions. Paper presented to the 1 992 Annual Conference of the Public Health Association of Australia, Canberra.

Brown K, Burrows C, Macarounas-Kirchmann K. Cost-utility analysis versus cost- effectiveness analysis: a matter of denominators and decimal dust? (in preparation).

Bryan S, Parkin D, Donaldson C (1 991). Chiropody and the QALY: a case study in assigning categories of disability and distress to patients. *Health Policy 18:1* 69-185.

Buxton MJ (1 987). Problems in the economic appraisal of new health technology: the evaluation of heart transplants in the UK. In MF Drummond (Ed), *Economic appraisal of health technology in the European community.* Oxford: Oxford University Press.

Buxton M, Ashby J, O'Hanion M (1 *987). Alternative methods of valuing health states: a comparative analysis based on an empirical study using the time trade-off approach in relation to health states one year after treatment for breast cancer.* Health Economic Research Group Discussion Paper, Brunei University.

Cairns JA, Johnston KM, McKenzie L *(1991). Developing QALYS from condition- specific outcome measures.* Working paper. Health Economics Research Unit, University of Aberdeen.

Capewell G (1 988). Techniques of health status measurement using a health index. In Teeling Smith G (Ed), *Measuring health: a practical approach,* 45-67. Chichester: John Wiley.

Carmines EG, Zeller RA (1 *979). Reliability and validity assessment.* Beverly Hills: Sage.

Carver RP (1974). Two dimensions of tests: psychometric and edumetric. *American Psychologist 29:512-518.*

Churchill DN, Morgan J, Torrance GW (1 984). Quality of life in end-stage renal disease. *Peritoneal Dialysis Bulletin 4:21-23.*

Churchill DN, Wallace JE, Ludwin D, Beecroft ML, Taylor DW (1991). A comparison of evaluative indices of quality of life and cognitive function in hemodialysis patients. *Controlled Clinical Trials* 12:159S-1 67S.

Cohen J (1977). Statistical power analysis for the behavioural sciences. New York: Academic Press.

Cronbach LJ (1 951). Coefficient alpha and the internal structure of tests. *Psychometrika 16:297-334.*

Cronbach LJ (1 957). The two disciplines of scientific psychology. *American Psychologist 12:671-684.*

Cronbach LJ, Furby L (1 970). How should we measure "change" -- or should we? *Psychological Bulletin* 74:68-80.

Culyer AJ (1 982). Assessing cost-effectiveness. In D Banta (Ed), *Resources for health.* New York: Praeger.

Culyer AJ (1 *991). Ethics and efficiency in health care. some Plain economic truths.* Centre for Health Economics and Policy Analysis, Department of Clinical Epidemiology and Biostatistics, McMaster University. CHEPA Commentary Series #C91 -1.

Culyer AJ, Maynard A (1 981). Cost-effectiveness of duodenal ulcer treatment. *Social Science and Medicine* 15C, 3-1 1.

Deyo RA, Centor RM (1 986). Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. *Journal of Chronic Diseases* 39:897-906.

Deyo RA, Diehi AK, Rosenthal M (1 986). How many days of bed rest for acute low back pain? A randomized clinical trial. *New England Journal of Medicine 315:1064-* 1070.

Deyo RA, Diehr P, Patrick DL (1 991). Reproducibility and responsiveness of health status measures: statistics and strategies for evaluation. *Controlled Clinical Trials* 1 2:1 42 S- 1 58S.

Deyo RA, Patrick DL (1 989). Barriers to the use of health status measures in clinical investigation, patient care, and policy research. *Medical Care* 27(Supplement):S254-S268.

Dowie J (1 *991). Health outcomes and evaluation.* Paper presented to the Australian Institute of Health Forurn, Priorities for National Health Statistics, Canberra.

Drummond MF (1 989). Output measurement for resource allocation decisions in health care. *Oxford Review of Economic Policy* 5:59-74.

Drummond M, Davies L (1 991). Economic analysis alongside clinical trials. Revisiting the methodological issues. *International Journal of Technology Assessment in Health Care.* 7:561-573.

Drummond MF, Mohide EA, Tew M, Streiner DL, Pringle DM, Gilbert JR (1 991). Economic evaluation of a support program for caregivers of demented elderly. *International Journal of Technology Assessment in Health Care 7:209-21* 9.

Drummond MF, Stoddart GL, Torrance GW (1 *987). Methods for the evaluation of health care programmes.* Oxford: Oxford University Press.

Drummond MF, Stoddart GL (1 984). Economic analysis of clinical trials. *Controlled Clinical Trials* 5:115-128.

Drummond M, Teeling Smith G, Wells N (1 *988). Economic evaluation in the development of medicines.* London: Office of Health Economics.

Fanshel S, Bush JW (1 970). A health status index and its application to health service outcomes. *Operations Research 18:1021-1066.*

Feeny D, Labelle R, Torrance GW (1 990). Integrating economic evaluations and quality of life assessments. In BF Spilker (Ed), *Quality of life assessments in clinical trials,* pp. 71-79. New York: Raven Press.

Feeny D, Torrance G (1 989). Incorporating utility-based quality-of-life assessment measures in clinical trials: two examples. *Medical Care* 27(Supplement):S190- S204.

Feinstein AR (1 987a). Clinimetric perspectives. *Journal of Chronic Diseases* 40:635-640.

Feinstein AR (1 *987b). Clinimetrics.* New Haven: Yale University Press.

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

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

Fleiss JL (1 986). Reliability of measurement. In JL Fleiss (Ed), *The design and analysis of clinical experiments.* New York: Wiley.

Froberg DG, Kane RL (1 989a). Methodology for measuring health-state preferences--11: scaling methods. *Joumal of Clinical Epidemiology 42:459-471.*

Froberg DG, Kane RL (1 989a). Methodology for measuring health-state preferences--IV: progress and a research agenda. *Journal of Clinical Epidemiology* 42:675-685.

Glasziou P, Simes J(1992). Using QALYs as the endpoint in randomised trials. Paper presented to 24th Annual Conference of the Public Health Association of Australia, Canberra.

Golembiewsky RT, Billingsiey K, Yeager S (1 976). Measuring change and persistence in human affairs: types of change generated by OD designs. *Journal of Applied Behavioural Science 12:133-157.*

Green LW, Lewis FM (1 *986). Measurement and evaluation in health education and health promotion.* Palo Alto, California: Mayfield.

Gudex C, Kind P (1 987). *The QALY toolkit.* Centre for Health Economics/Health Economics Consortium Discussion Paper 38. University of York.

Guyatt GH, Berman LB, Townsend M, Taylor DW (1 985). Should study subjects see their previous responses? *Journal of Chronic Diseases 38:1003-1007.*

Guyatt GH, Deyo RA, Charlson M, Levine MN, Mitchell A (1989). Responsiveness and validity in health status measurement: a clarification. *Journal of Clinical Epidemiology 42:403-408.*

Guyatt GH, Mitchell A,]rving EJ, Singer J, Williams N, Goodacre R, Tompkins CA (1 989). A new measure of health status for clinical trials in inflammatory bowel *disease. Gastroenterology* 96:804-81 0.

Guyatt GH, Veldhuyzen Van Zanten SJC), Feeny D, Patrick DL (1 989). Measuring quality of life in clinical trials: a taxonomy and a review. *Canadian Medical Association Journal 140:1441-1448.*

Guyatt G, Waiter S, Norman G (1 987). Measuring change over time: assessing the usefulness of evaluative instruments. *Journal of Chronic Diseases 40:171-178.*

Haig JHB, Scott D, Wickett L] (1 986). The rational zero point for an illness index with ratio properties. *Health Services Research 8:228-245.*

Hoogstraten J (1 985). Influence of objective measures on self-reports in a retrospective pre-test-protest design. *Journal of Experimental Education 53:207-* 210.

Howard GS, Dailey PR (1 979). Response-shift bias: a source of contamination of self-report measures. *Journal of Applied Psychology 64:144-150.*

Howard GS, Ralph KM, Gulanick NA, Maxwell SE, Nance SW, Gerber SK (1 979). internal invalidity in pre-test - protest self-report evaluations and a re-evaluation of retrospective pre-tests. *Applied Psychological Measurement 3:1-23.*

Howard GS, Schmeck RR, Bray JH (1 979). Internal invalidity in studies employing self-report instruments: a suggested remedy. *Journal of Educational Measurement* 16:129-135.

Kirwan JR, Chaput De Saintonge DM, Joyce CRB, Currey HLF (1984). Clinical judgement in rheumatoid arthritis. Ill. British rheumatologists' judgements of' change in response to therapy.' *Annals of Rheumatic Diseases* 43:686-694.

Jaeschke R, Guyatt GH (1 990). How to develop and validate a new quality of life instrument. In BF Spilker (Ed), *Quality of life assessments in clinical trials,* 47-57. New York: Raven Press.

Jaeschke R, Singer J, Guyatt GH (1 989). Measurement of health status: ascertaining the mimimal clinically important difference. *Controlled Clinical Trials* 10:407-415.

Kapian RM (1985). Quality-of-life measurement. In P Karoly (Ed), *Measurement strategies in health psychology, 11* 5-145. New York: Wiley.

Kapian RM (1 989). Health outcome models for policy analysis. *Health Psychology* 8:723-735.

Kaplan RM, Anderson JP (1 988). The Quality of Well-Being Scale: rationale for a single quality of life index. In SR Walker, RM Rosser (Eds), *Quality of life: assessment and application,* pp. 51-77. Lancaster: MTP Press.

Kapian RM, Anderson JP (1 990). The general health policy model: an integrated approach. In BF Spilker (Ed), *Quality of life assessments in clinical trials, 131-149.* New York: Raven Press.

Kaplan RM, Anderson JP, Wu A, Mathews WC, Kozin F, Orenstein D. (1 989). The Quality of Well-Being Scale: applications in AIDS, cystic fibrosis, and arthritis. *Medical Care* 27(Supplement): S27-S43.

Kapian RM, Atkins CJ, Wilson DK (1 988). The cost-utility of diet and exercise interventions in non-insulin dependent diabetes mellitus. *Health Promotion 2:331-* 340.

Kapian RM, Bush JW (1 982). Health-related quality of life measurement for evaluation research and policy analysis. *Health Psychology 1:61-80.*

Kazis L, Anderson JJ, Meenan RF (1 989). Effect sizes for interpreting changes in health status. *Medical Care* 27(Supplement): S110-S127.

Keppel G (1 *973). Design and analysis: a researcher's handbook.* New jersey: Prentice-Hall.

Kind P, Rosser RM, Williams A (1 982). Valuation of quality of life: some psychometric evidence. In Jones-Lee MW (Ed), *The value of life and safety, 159-* 170. Amsterdam: North-Holland.

Kirshner B, Guyatt G (1 985). A methodological framework for assessing health indices. *Journal of Chronic Diseases.* 38:27-36.

Kuder GF, Richardson MW (1 937). The theory of the estimation of test reliability. *Psychornetrika 2:151-160.*

Laupacis A, Feeny D, Detsky AS, Tugwell PX (1 992). How attractive does a new technology have to be to warrant adoption and utilization? Tentative guidelines for using clinical and economic evaluations. *Canadian Medical Association Journal* 146:473-481.

Laupacis A, Wong C, Churchill D, and the Canadian Erythropoietin Study Group (1991). The use of generic and specific quality-of-life measures in hemodialysis patients treated with Erythropoietin. *Controlled Clinical Trials* 12(Supplement): 1 68S- 179S.

Levine MN, Guyatt GH, Gent M, DePaue S, Goodyear MD, Hryniuk WM, Arnold A, Findlay B, Skillings JR, Bramwell VH, Levin L, Bush H, Abu-Zahra H, Kotalik J. Quality of life in stage 11 breast cancer: an instrument for clinical trials. *Journal of Clinical Oncology* 6:1798-181 0.

Liang MH, Larson MG, Cullen KE, Schwartz JA (1 985). Comparative measurement efficiency and sensitivity of five health status instruments for arthritis research. *Arthritis and Rheumatism 28:542-547.*

Liang MH, Robb-Nicholson C (1 987). Health status and utility measurement viewed from the right brain: experience from the rheumatic diseases. *Journal of Chronic Diseases* 40:579-583.

Linn PL, Slade JA (1 977). Determination of the significance of change between pre- and post testing periods. *Review of Educational Research 47:121-150.*

MacKenzie CR, Charisson ME, DiGioia D, Kelley K (1986). Can the Sickness Impact Profile measure change? An example of scale assessment. *Journal of Chronic Diseases 39:429-438.*

MacKenzie CR, Charisson ME, DiGioia D, Kelley K (1 986). A patient-specific measure of change in maximal function. *Archives of Internal Medicine 146:1325-* 1329.

Maxwell SE, Howard GS. Change scores--necessarily anathema? (1981). *Educational and Psychological Measurement 41:747-756.*

Meenan RF, Anderson JJ, Kaziz LE, Egger MJ, Altz-Smith M, Samuelson CO, Wilikens RF, Soisky MA, Hayes SP, Blocka KL, Weinstein A, Guttadauria M, Kapian SB, Klippel J (1 984). Outcome assessment in clinical trials: evidence for the sensitivity of a health status measure. *Arthritis and Rheumatism 27:1344-1352.*

Mohide EA, Pringle DM, Streiner DL, Gilbert JR, Muir G, Tew M (1 990). A randomized trial of family caregiver support in the home management of dementia. *Journal of the American Geriatrics Society* 38:446-454.

Mohide EA, Torrance GW, Streiner DL, Pringle DM, Gilbert R (1 988). Measuring the wellbeing of family caregivers using the time trade-off technique. *Journal of Clinical Epidemiology 41:475-482.*

Mugford M, Drummond MF (1 989). The role of economics in the evaluation of care. In Chalmers I, Enkin M, Keirse MJNC (Ed), *Effective care in pregnancy and child birth.* Oxford: Oxford University Press.

Nord E (1 992). Methods of quality adjustment of life years *. Social Science and Medicine* 34:599-569.

Norman GR (1989). Issues in the use of change scores in randomized trials. *Journal of Clinical Epidemiology* 42:1097-1105.

Nunnally J. The study of change in evaluation research: principles concerning measurement, experimental design, and analysis, Volume 1. In Streuning EL, Guttentag M (Ed), *Handbook of evaluation research, 101* -1 38. Beverly Hills: Sage, 1975.

Oken MM, Creech RH, Tormey DC, Horton J (1 982). Toxicity and response criteria of the Eastern Cooperative Oncology Group. *American Journal of Clinical Oncology* 5:649-655.

Orenstein DM, Pattishall EN, Nixon PA, Ross EA, Kapian RM (1 990). Quality of well-being before and after antibiotic treatment of pulmonary exacerbation in patients with cystic fibrosis. *Chest* 98:1081-84.

Paterson M (1 988). Assessment of treatment in rheumatoid arthritis. In G Teeling Smith (Ed), *Measuring health: a practical guide,* 157-189. New York: Wiley.

Patrick DL (1 990). Disease-specific versus general health state measures. In M Drummond (Ed), *Measuring the quality of life of people with visual impairment: proceedings of a workshop.* U.S. Department of Health and Human Services, National Institutes of Health National Eye Institute, NIH Publication No. 90-3078.

Patrick DL, Deyo RA (1989). Generic and disease-specific measures in assessing health status and quality of life. *Medical Care* 27(Supplement):S217-SS232.

Patrick DL, Erickson P (1 988). Assessing health-related quality of life for clinical decision making. In SR Walker, Rosser RM (Ed), *Quality of life: assessment and application,* 9-49. Lancaster: MTP Press.

Revicki D (1 992). Relationship between health utility and psychometric health status measures (1 *992). Medical Care* 30(Supplement):MS274-MS282.

Richardson J (1 991). Economic assessment of health care: theory and practice. *Australian Economic Review* 1:14-21.

Richardson J, Cook J (1 992). Cost utility analysis: 'new directions' in setting health care priorities. *Australian Health Review* 15:145-154.

Richardson J, Hall J, Salkeld G (1 990). Cost utility analysis: the comparability of measurement techniques and the measurement of utility through time. In C Selby Smith (Ed), *Economics and health: 1989.* Proceedings of the Eleventh Australian Conference of Health Economists. Melbourne: Public Sector Management Institute, Monash University.

Rosser RM (1 990). From health indicators to quality adjusted life years: technical and ethical issues. In Hopkins A, Costain D (Eds), *Measuring the outcomes of medical care, 1-1* 7. London: Royal College of Physicians of London.

Rosser RM, Kind P (1 978). A scale of evaluations of states of illness: is there a social *consensus? International Journal of Epidemiology 7:347-358.*

Spitzer WO (1 987). State of science 1986: quality of life and functional status as target variables for research. *Journal of Chronic Diseases 40:465-471.*

Streiner DL, Norman GR (1 *989). Health measurement scales: a practical guide to their development and use.* Oxford: Oxford University Press.

Terborg JR, Howard GS, Maxwell SE (1980). Evaluating planned organisational change: a method of assessing alpha, beta, and gamma change. *Academy of Management Review 5:109-121.*

Thompson MS (1 986). Willingness to pay and accept risks to cure chronic disease. *American Journal of Public Health* 76:392-396.

Thompson MS, Read JL, Hutchings C, Paterson M, Harris ED (1 988). The cost- effectiveness of auranofin: results of a randomized clinical trial. *Journal of Rheumatology 15:35-42.*

Toevs CD, Kapian RM, Atkins CJ (1 984). The costs and effects of behavioural programs in chronic obstructive pulmonary disease. *Medical Care* 22:1088-1 1 00.

Torrance GW (1 976). Social preferences for health states. An empirical evaluation

of three measurement techniques. *Socio-economic Planning Sciences 10:1* 29-136.

Torrance GW (1 986). Measurement of health state utilities for economic appraisal: a review. *Journal of Health Economics 5:1-30.*

Torrance GW (1 987). Utility approach to measuring health-related quality of life. *Journal of Chronic Diseases 40:593-600.*

Torrance GW, Feeny D (1989). Utilities and quality-adjusted life years. *International Journal of Technology Assessment in Health Care.* 5:559-575.

Torrance GW, Zipursky A (1 984). Cost-effectiveness of anteartum prevention of Rh immunisation. *Clinics in Perinatology 11:267-281.*

Tugwell P, Bombardier C, Bell M, Bennett K, Bensen W, Grace E, Hart L, Goldsmith C (1 991). Current quality-of-life research challenges in arthritis relevant to the issue of clinical significance. *Controlled Clinical Trials* 12:217S-225S.

Tugwell P, Bombardier C, Buchanan WW, Goldsmith C, Grace E (1 987). The MACTAR Patient Preference Questionnaire--An individualised functional priority approach for assessing improvement in physical disability in clinical trials in rheumatoid arthritis. *Journal of Rheumatology 14:446-561.*

Veldhuyzen Van Zanten SJO (1991). Quality of life as outcome measures in randomized controlled trials: an overview of three medical journals. *Controlled Clinical Trials* 12(Supplement):234S-242S.

Veney JE, Kaiuzny AD (1 *984). Evaluation and decision making for health services programs.* Engelwood Cliffs, New Jersey: Prentice-Hall.

Ware JE, Brook RH, Davies-Avery A, Williams KN, Stewart AL, Rogers WH, Donald CA, Johnston SA. (1 980). Conceptualisation and measurement of health for adults in the health insurance study. 1 *n: Model of health and methodology, Volume 1.* Santa Monica, CA: Rand.

Wikiund 1, Kariberg J (1 991). Evaluation of quality of life in clinical trials: selecting quality-of-life measures. *Controlled Clinical Trials* 12(Supplement):204S-216S.

Wikiund 1, Dimenas E, Wahl M (1990). Factors of importance when evaluating quality of life in clinical trials. *Controlled Clinical Trials 1* 1:1 69-179.

Williams A (1 974). The cost benefit approach. *British Medical Bulletin* 30:252-256.

Williams A (1 983). The role of economics in the evaluation of health care technology. In Culyer AJ, Horisberger B (Eds), *Economic and medical evaluation of health care technologies.* Berlin: Springer-Veriag.

Williams A (1 985). The economics of coronary artery bypass grafting. *British Medical Journal 291:325-329.*

Williams A (1 988). The importance of quality of life in policy decisions, 279-290. In SR Walker, RM Rosser, (Eds), *Quality of life: assessment and application.* Lancaster: MTP Press.

Williams A (1991). Is the QALY a technical solution to a political problem? Of course not! International Journal of Helath Services 21:365-369.

Wu a, Mathews W, Brysk L, Atkinson JH, Grant I, Abramson I, Kennedy CJ, McCutchan JA, Spector SA, Richman DD (1990). Quality of life in a placebo-controlled trial of zidovudine in patients with AIDS and AIDS-related complex. Journal of Acquired Immune Deficiency Syndromes 3:683-690.

TABLE 1
SINGLE-FACTOR REPEATED MEASURES DESIGN

NOTATIONAL SYSTEM: (A x S) DESIGN

	Levels of Factor A				
Subjects	a_1	a_2	a_3	a_4	Sum
s_1	AS_{11}	AS_{21}	AS_{31}	AS_{41}	S_1
s_2	AS_{12}	AS_{22}	AS_{32}	AS_{42}	S_2
s_3	AS_{13}	AS_{23}	AS_{33}	AS_{43}	S_3
s_4	AS_{14}	AS_{24}	AS_{34}	AS_{44}	S_4
Sum	A_1	A_2	A_3	A_4	T

COMPUTATIONAL FORMULAS: (A x S) DESIGN

Source of variance	Sum of Squares (SS)[a]	Degrees of freedom (df)	Mean Square (MS)
$A_{(occasions)}$	$\frac{\Sigma(A)^2}{s} - \frac{(T)^2}{as}$	(a - 1)	$\frac{SS_A}{df_A}$
$S_{(subjects)}$	$\frac{\Sigma(S)^2}{a} - [T]$	(s - 1)	$\frac{SS_S}{df_S}$
A x S (residual error)	$\Sigma(AS)^2 - [A] - [S] - [T]$	(a - 1)(s - 1)	$\frac{SS_{AxS}}{df_{AxS}}$
Total	[AS] - [T]	(as - 1)	

a. Bracketed letters represent complete terms in the computational formulas; a particular terms is identified by the letter(s) appearing in the numerator.

Equations relating means squares (MS) to variances:

$$MS_{occasions} = \sigma^2_{error} + a\sigma^2_{subjects}$$

$$MS_{subjects} = \sigma^2_{error} + s\sigma^2_{occasions}$$

$$MS_{error} = \sigma^2_{error}$$

TABLE 2
MEASURING CHANGE OVER TIME

Instrument A	Time 1	Time 2	Intervention	Time 3	Difference Score	Exercise test result
Subject 1	8	9		15	+6	Much improved
Subject 2	9	8		15	+7	Much improved
Subject 3	8	9		15	+6	Much improved
Subject 4	9	8		15	+7	Much improved
Subject 5	8	9		8	--1	Unchanged
Subject 6	9	8		8	+1	Unchanged
Subject 7	8	9		8	--1	Unchanged
Subject 8	9	8		8	+1	Unchanged
Instrument B	**Time 1**	**Time 2**	**Intervention**	**Time 3**	**Difference Score**	**Exercise test result**
Subject 1	5	5		5	0	Much improved
Subject 2	9	9		9	0	Much improved
Subject 3	13	13		13	0	Much improved
Subject 4	17	17		17	0	Much improved
Subject 5	5	5		5	0	Unchanged
Subject 6	9	9		9	0	Unchanged
Subject 7	13	13		13	0	Unchanged
Subject 8	17	17		17	0	Unchanged
Instrument C	**Time 1**	**Time 2**	**Intervention**	**Time 3**	**Difference Score**	**Exercise test result**
Subject 1	4	5		7	+2	Much improved
Subject 2	6	7		11	+4	Much improved
Subject 3	12	13		15	+2	Much improved
Subject 4	16	15		19	+4	Much improved
Subject 5	4	3		4	+1	Unchanged
Subject 6	8	9		8	--1	Unchanged
Subject 7	12	11		12	+1	Unchanged
Subject 8	16	17		16	--1	Unchanged

Source: Adapted from Guyatt, Walter & Norman, 1987 (reproduced with permission)

TABLE 3
MAJOR ISSUES IN INDEX CONSTRUCTION AND VALIDATION

	DISCRIMINATIVE INDEX:	EVALUATIVE INDEX:
PURPOSE	used to distinguish between individuals or groups on an underlying dimension when no external criterion or gold standard is available for validating these measures.	used to measure the magnitude of longitudinal change in an individual or group on the dimension of interest.
CONSTRUCTION		
item selection	- tap components of the domain on which underlying disease has a substantial impact - universal applicability to respondents - stability over (at least short periods) of time	- tap areas related to change in health status - responsiveness to clinically significant change
item reduction	- short response sets that facilitate uniform interpretation	- response sets with sufficient gradations to register change
item scaling	- internal scaling or consistency - comprehensiveness and reduction of random error versus respondent burden	- responsiveness versus respondent burden
VALIDATION		
reliability	- large and stable between-subject variation: correlation between replicate measures	- stable within-subject variation: insignificant variation between replicate measures
validity	- cross-sectional construct validity: relationship between index and external measures at a single point in time	- longitudinal construct validity: relationship between changes in index and external measures over time
responsiveness	- not relevant	- power of the test to detect a clinically important difference

Source: Adapted from Guyatt et al, 1989 (reproduced with permission).

TABLE 4
RESPONSIVENESS AND UTILITY MEASURES OF HEALTH STATUS

sample	intervention(s)	utility measure(s) (months)	measurement interval	evidence presented re: treatment effects and responsiveness	reference(s)
CLINICAL TRIALS:					
n = 303 arthritis patients	. Auranofin v. placebo	. QWB .modified TTO (PUMS) .standard gamble (SG) .willingness to pay (WTP)	-1/2, 0, 1, 2, 4, 6 5 5 5	Time-by treatment interaction (ANCOVA) significant for QWB (P=.005), modified TTO (P=.002), standard gamble (P=.05) but not WTP (P=.29). Modified TTO had largest relative treatment effect of all 28 outcome -- perhaps because patients directly assessed the difference between pre-trial and current health states. Standardised effect sizes reported as follows: QWB (0.23); modified TTO (-0.43), standard gamble (-0.27) and willingness to pay (-0.15).	Bombardier et al, 1986; Bombardier et al, 1991 See also: Feeny & Torrance, 1989; Kaplan et al, 1989; Thompson et al, 1988.
n = 118 ESRD patients	. Erythropoietin (EPO) v. placebo	. TTO	0, 2, 4, 6	Time-by-treatment interaction (ANOVA) not significant (P<.05) for TTO; KDQ most responsive to change (P<.001 for fatigue and physical dimension). Aggregate global (P<.02) and physical SIP scores improved with EPO therapy; correlation of change scores between baseline and 6 months significant in most cases. Reported data insufficient to calculate effect size, SMR or responsiveness coefficient.	Laupacis et al, 1991.
n = 42 caregivers of demented elderly	caregiver support program v. community nursing	.TTO with "marker states" (CQLI)	0, 3, 6	Time-by-treatment interaction (ANCOVA) not significant for CES-D, TTO with marker states (CQLI) or STAI; 20% difference from baseline scores for CQLI in experimental group judged to be clinically important (post hoc). Reported data sufficient to calculate effect size: -0.73 (experimental); .027 (control).	Drummond et al, 1991; Mohide et al, 1990 See also: Mohide et al, 1988.

TABLE 4 (continued)
RESPONSIVENESS AND UTILITY MEASURES OF HEALTH STATUS

sample	intervention(s) ts measure(s)	utility interval (months)	measurement evidence presented re: treatment effects and responsiveness	reference(s)
n = 76 COPD patients	behavioural programs to enhance exercise compliance v. attention and no attention controls	. QWB	0, 3, 6, 12, 18	Time-by-treatment interaction (ANOVA) significant (P<.002) at 3-month follow-up and at each subsequent assessment period (but only marginally so at last follow-up for QWB and exercise tolerance). Data reported insufficient to calculate effect size, SMR, or responsiveness coefficient.	Toevs et al, 1984 Atkins et al, 1984
n = 31 AIDS and AIDS-related complex patients	. Zidovudine (ZDV) v. placebo	. QWB	0, 1, 2, 3, 4, 5, 6, 9, 12	Time-by-treatment interaction(ANOVA) significant at end of blinded trial for KPS (P<.004) and QWB (P<.028), and at end of 12 months -- KPS (P<.002) and QWB (P<.002) -- despite placebo patients crossing over to ZDV. However, because QWB incorporates death in its score, these results largely reflect differences in mortality. For survivors only time-by-treatment interaction (ANOVA) significant at end of trial for KPS (P<.016) but not QWB (P>.089). Reported data sufficient to calculate effect size: 0.463 (experimental condition, group), 0.019 (control group, survivors).	Wu et al, 1990 See also: Kaplan et al, 1989.
n = 76 non-insulin -dependent diabetic patients	education v. diet-plus-education	. QWB	0, 4, 8, 12, 18	By time of 18-month follow-up, diet/exercise group experienced 0.06 QWB units of improvement, compared with -0.04 for the control group i.e., a difference of 0.092 QWB units. Time-by-treatment interaction (ANOVA) was significant (P<.01). Data reported insufficient to calculate effect size, SMR, or responsiveness coefficient.	Kaplan, Atkins & Wilson, 1988

TABLE 4 (CONTINUED)
RESPONSIVENESS AND UTILITY MEASURES OF HEALTH STATUS

sample	intervention(s)	utility measure(s) (months)	measurement interval	evidence presented re: treatment effects and responsiveness	reference(s)
QUASI-EXPERIMENTAL DESIGNS:					
n = 47 chronic dialysis patients	dialysis "dose" to Kt/V > 1	.TTO	0, 1 1/2-2	No significant correlation between change in Kt/V values and change in TTO values. Expected failure of TTO to demonstrate responsiveness not specified as *a priori* hypothesis. Separate pre- and post-test means and SDs not reported for patient sub-groups whose dialysis dose was either maintained at KT/V > 1 or increased to Kt/V > 1.	Churchill et al, 1991
n = 21 patients with ulcerative colitis	surgery	. TTO . direct questioning of objectives (DQO)	0, 10-19 0, 10-19	Differences between pre- and post-operative utilities statistically significant (P<.05, paired t-test). Reported data allows calculation of effect size: 1.147 (TTO), 2.0 (DQO).	McCleod et al, 1991
n = 38 arthritis patients	hip or knee arthroplasty	. QWB	-1/2, 3, 12, 15	Response values on 4 scales (global health, pain, mobility and social function) on 4 instruments (QWB, FSI, HAQ and SIP) mostly significant (paired t-test) at 3- and 12-month follow-ups. SMR for QWB: 1.13 0.88(3-month), 0.88 (12-month). Reported data permits calculation of effect size: 1.88 (3-month), 1.71 (12-month).	Liang, Fossel & Larson, 1990 See also: Liang et al, 1985.
n = 28 cystic fibrosis patients	antibiotic treatment (oral ciprofloxacin)	. QWB	0, 1/2	QWB scores generally improved, with a mean change of 0.104+0.122, reflecting a responsiveness coefficient of 0.85. Significant correlation between change in QWB values and change in most pulmonary function tests: FVC, FEV_1 and SaO_2.	Orenstein, Pattishall, Nixon, Ross & Kaplan, 1990.

FIGURE 1
NUANCES OF VALIDITY: DISCRIMINATIVE VERSUS EVALUATIVE INSTRUMENTS

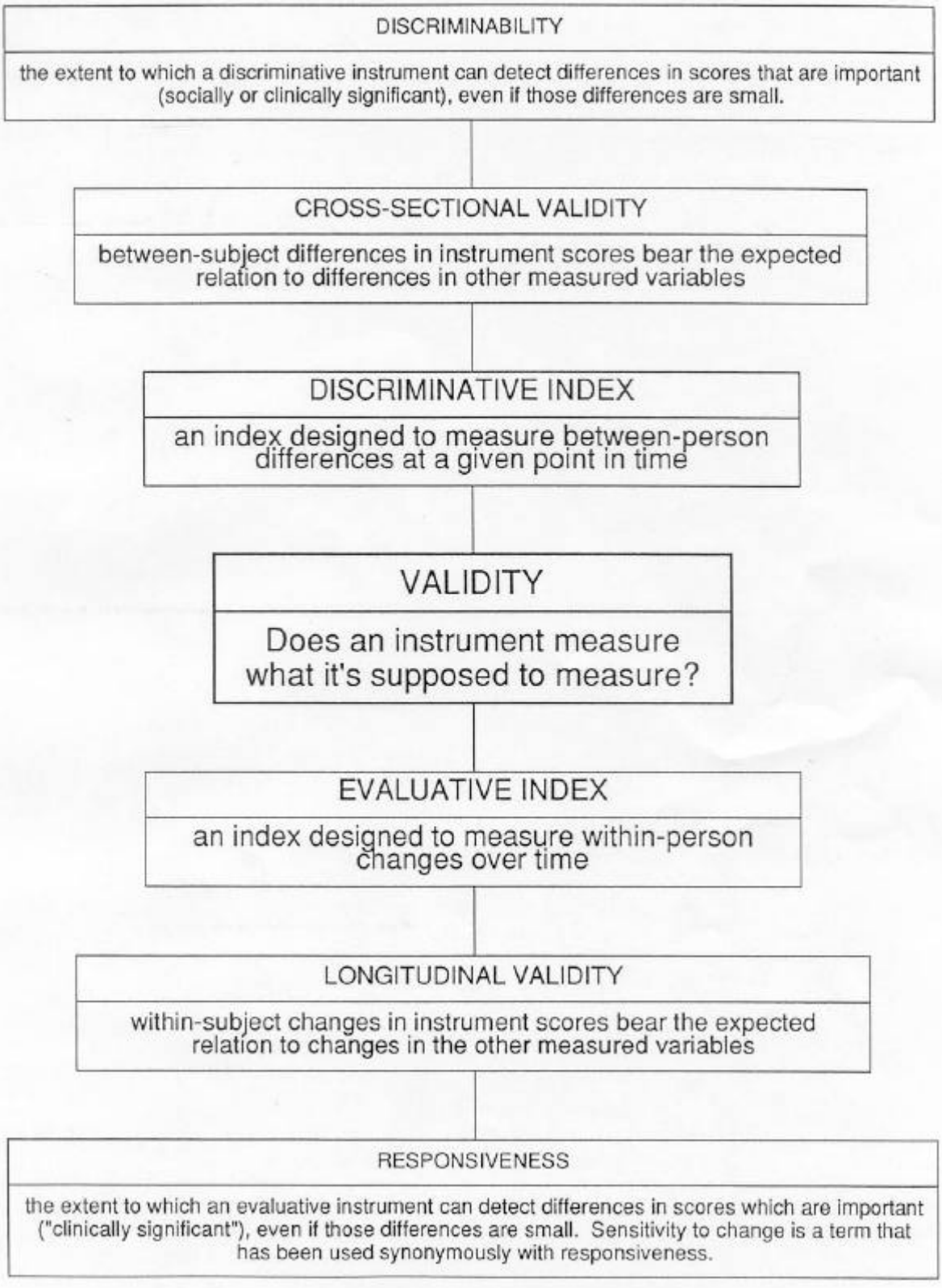


FIGURE 2
RELATIONSHIP BETWEEN RESPONSIVENESS AND VALIDITY: GUYATT ET AL'S EXAMPLES

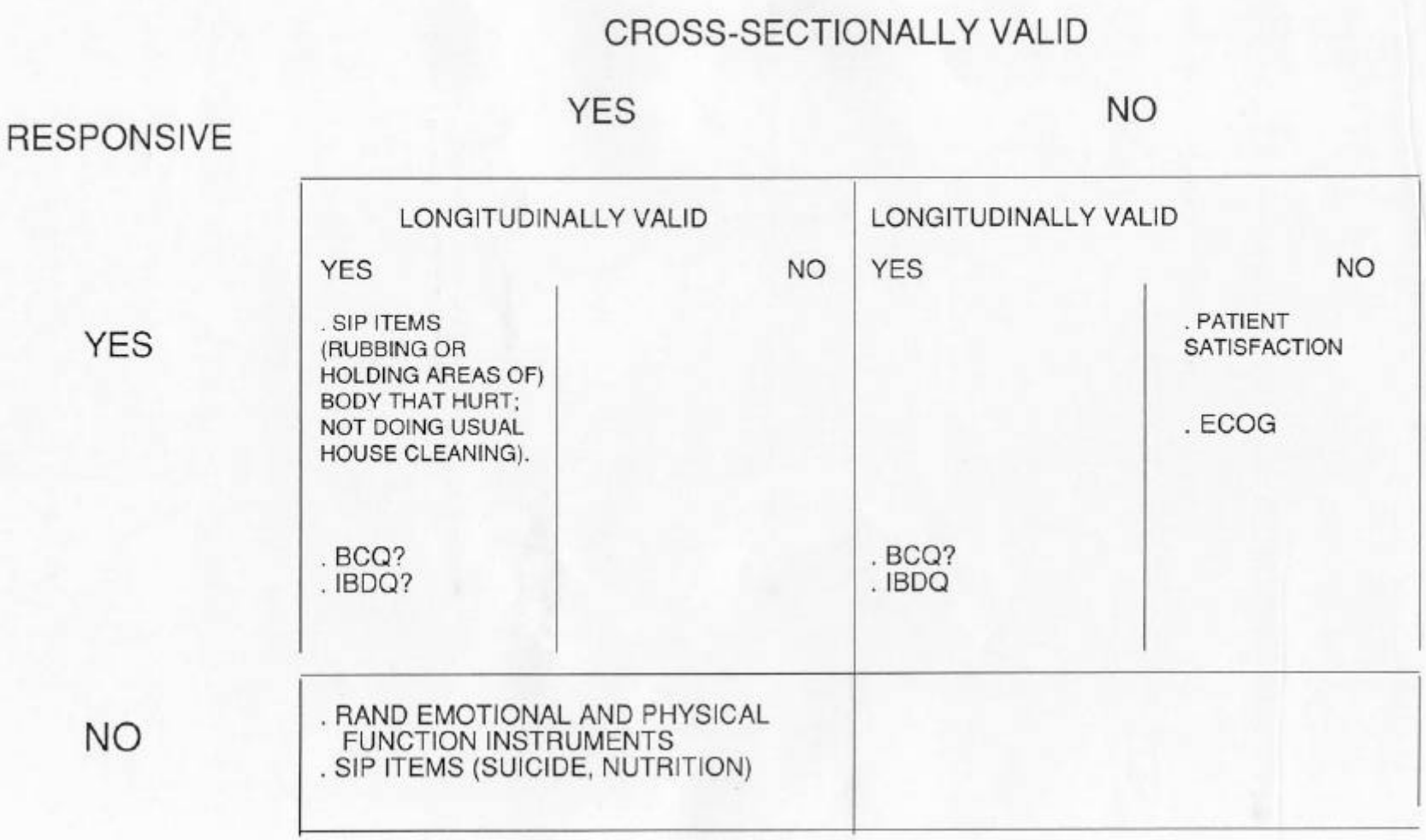


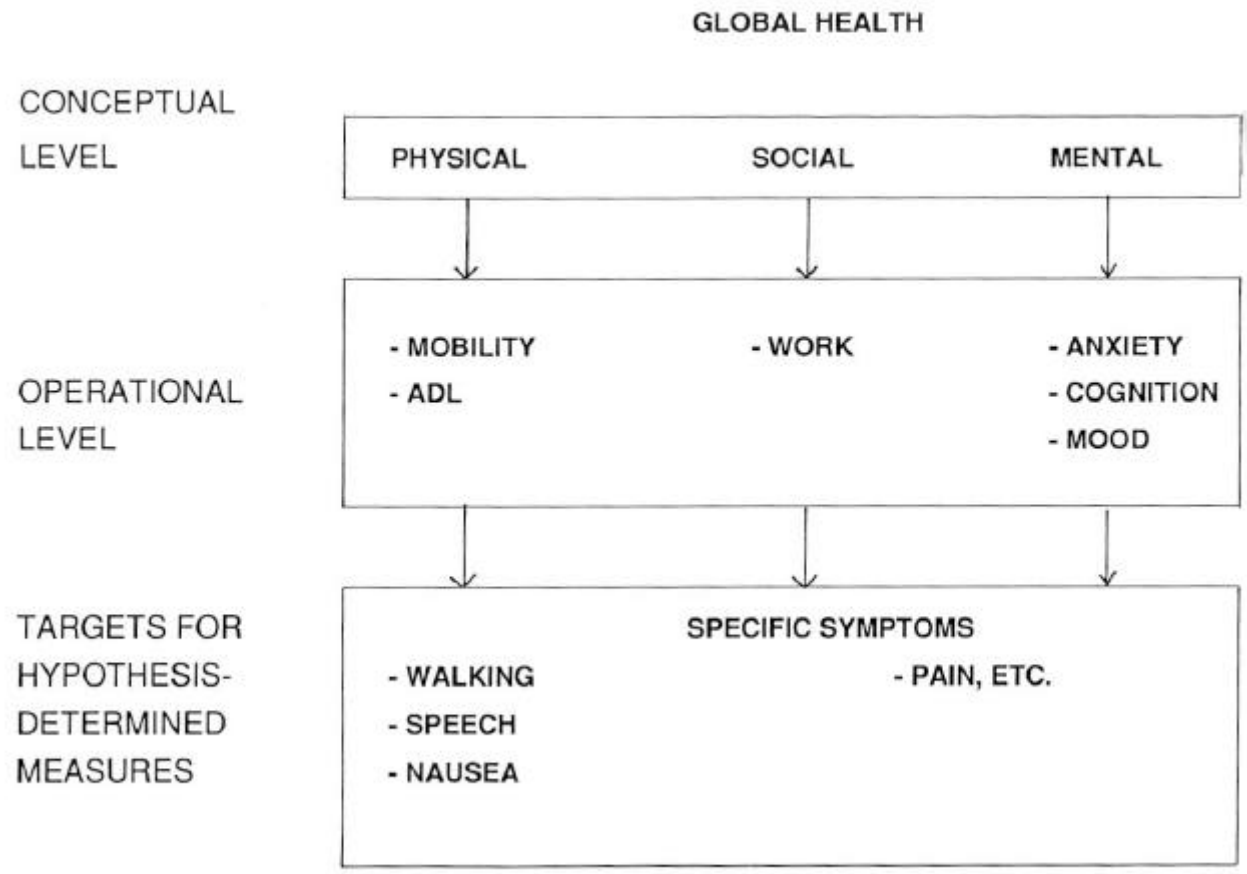
FIGURE 3
RELATIONSHIP BETWEEN EFFICACY OF AN INTERVENTION AND THE LONGITUDINAL VALIDITY OF AN EVALUATIVE INSTRUMENT

RESPONSIVENESS

Change in Self-Reported Health Status?

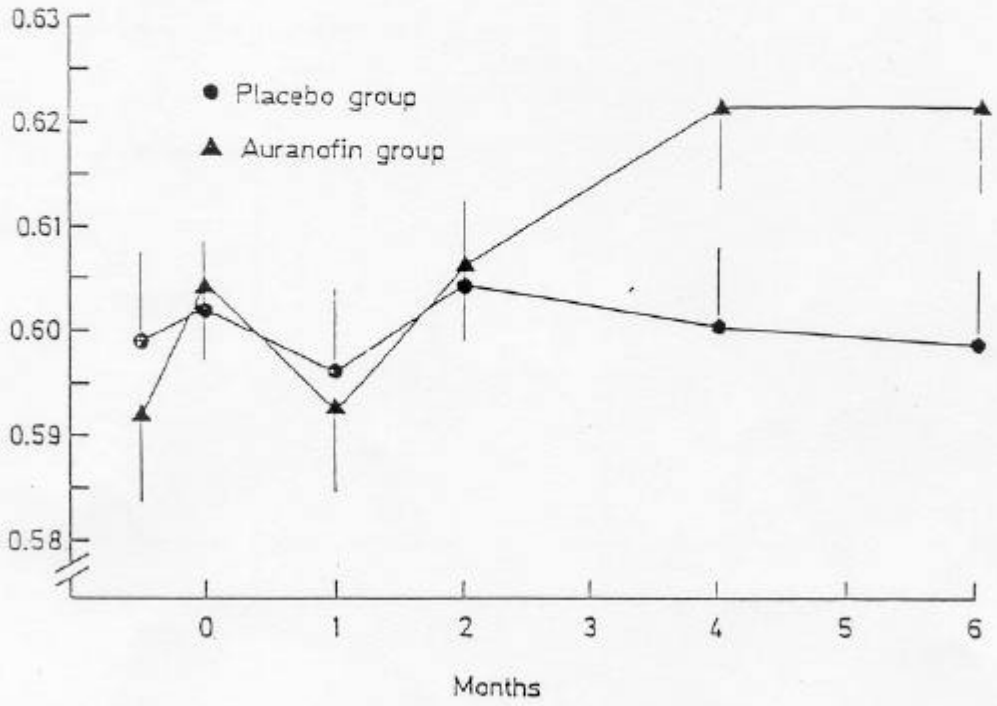
EFFICACY		YES	NO
Does it work?	YES	valid + efficacious	not valid + efficacious
	NO	not valid + not efficacious	valid/not valid + not efficacious

FIGURE 4:
LEVELS OF OPERATIONALIZATION FOR QUALITY OF LIFE AND FUNCTIONAL STATUS



Source: Spitzer, 1987 (reproduced with permission)

FIGURE 5
AURANOFIN TRIAL: SCORES ON QUALITY OF WELL-BEING QUESTIONNAIRE



Source: Paterson, 1988.