Chapter - 6

Health State Valuations Study in Andhra Pradesh: Review of Literature and Methods

Prasanta Mahapatra, Josh A Salomon, Lipika Nanda and KT Rajshree

This chapter presents a review of available literature, followed by description of the methods of the health state valuation study in Andhra Pradesh as well as the reliability and validity of health state valuation measurement. The next chapter describes the results of the community survey and summarises inferences from the health state valuation study. The review of literature is organised into four subsections, starting with a brief overview of theoretical considerations regarding the use of health state valuations. The next subsection poses the question "how to describe health states for valuation?" and reviews relevant conditions for the definition of a health state description space. The following subsection describes various approaches to measurement of health states. The fourth subsection presents an overview of instruments being used for general health status measurement. The second section pertains to the study method and starts with a subsection that describes the 6D5L health state description system that we developed for the study. The description system is built upon the EuroQol, but with important extensions by way of an additional dimension, two additional levels and a graphical description system. This is followed by two subsections describing the two major arms of the study, namely: (a) the multi-method deliberative health state valuation workshops, and (b) the community-based health state valuation. In the third section socio-economic characteristics of the respondents are presented along with an overview of data collected from the two arms of the study. In the fourth section, the reliability of the health state valuation measurements that we had undertaken, is analysed. In the course of estimating reliability of our measurements, the data was found to be inconsistent with conventional notions of personal valuations as single valued functions. Hence we consider a re-examination of the nature of value functions for health states and propose that multivalued function is more consistent with empirical observations. The fifth and final section analyses validity of the health state valuation measurements.

Review of Literature:

Theoretical considerations about use of health state valuations:

Theoretical interpretations about the object of measurement have some impact on the methodological path leading up to health status weights. Three different interpretations have been made in the literature, namely: (a) individual preference, (b) descriptive measure of health state and (c) social preference weight attached to the health state. Culyer (1989) distinguished between "welfarist" and "extra-welfarist" approaches to health status measurement among economist. The "Welfarist" approach consists of viewing individual preferences (utility) as the source of all social welfare. Health state or disability weight is viewed to represent individual preference for different health states. Since health outcomes at a personal level are characterised by uncertainty, Von Neuman and Morgenstern's (VNM) expected utility theory[1] is applied. Thus health state weights are viewed as VNM utility. Viewing health state weight as a measure of personal utility means that an equivalence with utility of other goods and services is straightforward. Advocates of personal utility interpretations may not, however, emphasise this equivalence in deference to strong emotional responses against valuation of human life in money terms. "Extra-welfarists" view health as the principal outcome of health services. The health state weight describes this output i.e. the health -related quality of life (Bleichrodt, 1997). The third interpretation, that quality adjustment weights are values under a social welfare function ,is proposed by Nord (1994). Patrick et al (1973) also recognised that quality adjustment weights derived from the equivalence (person tradeoff) method would represent valuation under some social welfare function.

Another analytic issue is concerned with how to treat the large amount of illnesses that exist but for which people do not seek treatment. It has been observed that aggregate measures like the sickness rate (number of persons falling ill per time period) or illness episodes per time period based on simple count and the implied equal weightage to illnesses of all severity, are too stable and non-responsive to variations of incidence of more severe illnesses (Logan and Brooke 1957). This is in fact the primary motivation behind the search for a set of unequal health state weights. The relative weight to illness of different severity will depend on the current concept of health. At a time characterised by survival as the dominant concept of health, weights attached to all forms of morbidity are nearly zero. As the concept of health evolves to include absence of disease, more severe forms of disability begin to receive higher weightage,

[1]Von Neumann and Morgenstern (1944) proposed the expected utility theory. This theory is described in any text book on micro economic theory. One good exposition can be found in Mas-Collel, Whinstone and Green (1995).

both in the minds of the patients and public. Further evolution of the concept of health to include quality of life would naturally enhance the weightage received by illnesses considered minor and trivial in an earlier era. Rosser (1983) notes that even though Logan and Brooke, in 1957, sought to increase the sensitivity of aggregate indicators of morbidity by splitting down some categories (thereby assigning zero weightage to excluded conditions), such an approach would not be relevant in view of increasing concerns about these so-called minor and trivial illnesses.

Health status measures differ in sensitivity to minor and trivial illnesses. For example, the Quality of Well Being scale (QWB) is known to be more sensitive to small departures from perfect health (McDowell and Newell, 1996 p.483-491). This is because the QWB construct includes a symptom complex dimension. For example, Erickson et al (1989) found that the QWB scale classified 95% of the 45-64 year -old population in less than perfect health compared to 75% when activity limitation was used as the criterion. On the other hand EuroQol is less sensitive to small departures from perfect health (McDowell and Newell, 1996 p480-483). This instrument differentiates between more severe forms of morbidity but lumps all morbidities at the healthy end of the scale.

Thus, the choice of instrument in assigning health state weight will have an impact on the kind of policy application to which the resultant aggregate measure of disease burden can be assigned. For example, generic health status measures insensitive to small variations in functional status in a particular dimensions or sub-dimension, would not be suitable to demonstrate the efficacy of a new therapy acting to improve the quality of life in that dimension or sub - dimension alone. This is the rationale for many condition-specific health status measurement scales. Here the policy application of health status measurement is in approval of new formulations and procedures, etc. A different kind of policy application involves the allocation of resources at the macro economic level. For this purpose, a generic health status measurement allowing for comparison of a large number of health care interventions would be useful. Within the class of generic health status measurement, scales may vary in their sensitiveness to different levels of disability. For example, if a health state valuation instrument is not sensitive to minor and trivial illnesses, the resultant disease burden estimates would not be useful for planning of ambulatory services.

How to describe health states for valuation:

When asked to value time spent in a health state without information about the key domains of health, a respondent must guess what the level in those dimensions will be. This will inevitably introduce measurement error and

a potential for bias in the results. Another consideration is how to convey relevant information about a hypothetical health state to the individual undertaking the valuation, who might not have personally encountered the state. Disease labels are short and parsimonious, but do not convey adequate information about functional status[2]. Moreover, disease labels are vulnerable to different interpretations based on cultural and personal settings. Issues relevant to development of a health state description systems have been described by Boyle and Torrance (1984), and Froberg and Kane I-IV (1989). Briefly, four important considerations guide us in defining the description space (number of dimensions) and the inclusion of specific attributes[3] of human health.

1. Conceptual definitions of health and deduced description systems.
2. Empirically gathered health-related attributes, and description systems induced by them.
3. Attention span and cognitive capacity of the human mind to process multi-dimensional information.
4. Statistical analysis of multi-attribute measurements.

Conceptual definitions of health and deduced description systems:

An ideal definition of health provides us the goal towards which a formal health state description system should work. Deductive lineage from an ideal definition, gives the description system its content validity. Hence a description system should incorporate as much of our ideal notion of health, as is practically feasible. Concepts of health as well as support to and criticism of various definitions of health has been reviewed by Fanshel (1972), Patrick et al (1973), Boyle and Torrance (1984), Noack (1987), Goldberg and Dab (1987), Stewart (1992) and Patrick and Ericson (1993). Based on an overview of concepts of health Noack (1987), highlighted two common elements in various definitions, namely (a) that health is a holistic concept, and (b) that health is a multi-dimensional concept. Intuitively appealing this view is invariably shared by researchers dealing with the subject of health and its measurement.

---

[2]It is now widely recognised that health states should be described in terms of functional status. Functional status information can be presented either in a narrative or structured format. For example, Sacket and Torrance (1978) used brief scenarios written up with help of clinicians, to describe various health states. EuroQol uses a structured approach where each health state is described in terms of five dimensions and threeseverity levels within each dimension (Brooks, EuroQol Group, 1996). The Health Utilities Index uses a structured approach consisting of an eight dimensions and 5-6 severity levels within each dimensions and 4-5 severity levels with in each.

[3] "Attributes" and "dimensions" are used in health status measurement literature, , interchangeably.

The holistic and multidimensional character of health is emphasised by the WHO definition of health. WHO's constitution defines health as a state of complete physical, mental and social well-being and not merely absence of disease or infirmity. This is a very inclusive definition. The definition certainly motivates health workers to integrate their role into a social well-being world-view. From the analytic perspective, there could be some doubt as to whether including social well-being within the health construct helps or hinders analysis. For example, restricting the concept of health status to physical and mental health would allow for testing of research questions as to how actions in the health care sector affect social well-being. On the other hand, an inclusive definition would make it difficult to identify the effect of actions in the health sector on an overall social well-being. Many generic health status measurement tools have drawn inspiration from the WHO's definition of health, using the physical, mental and social well- being triad as a starting point for the inclusion of dimensions and items within them. Some examples are: the EuroQol (Brook, 1996), the health status index (Fanshel and Bush, 1970) which has since evolved into the more commonly known Quality of Well Being Scale (Kaplan and Bush, 1982) and the McMaster Health Index Questionnaire (Chambers, 1976). In the EuroQol instrument, for example, mobility and self-care would map to physical functioning; usual activities are linked to social functioning; and anxiety and depression would represent mental health.

Empirically gathered health-related attributes, and description systems induced by them:

Table 6.1: Mapping of selected health status description systems to EQ-5D.

SIP	QWB	NHP	EQ-5D
Ambulation	Mobility	Physical abilities	Mobility
Mobility	Physical activity		
Body care and movements			Self care
Eating	Social activity - self care		
Work	Social activity - major		
Home management			Usual activities +
Recreation and pastimes	Social activity - other		
	Pain		Pain - discomfort
Emotional behaviour	Emotional reactions		
Sleeping and rest	Sleep		Anxiety - Depression
Social interaction	Social isolation		
Communication			
Alertness behaviour	Energy level		

+ Includes main activity and leisure which were separate in early versions of EuroQol.

Authors of the Quality of Well Being scale first abstracted "several hundred" case descriptions from medical texts. Then they consulted various survey instruments including the Health Interview Survey of the US National Center for Health Statistics, Alameda County Population Laboratory's community social surveys. Items from the survey instruments were selected to cover the range of disturbances in functional status (Patrick, Bush and Chen, 1973a).

Development of the Sickness Impact Profile (SIP) began with accumulation of statements describing behavioural changes attributable to sickness. These statements were collected from a sample of enrolees in a prepaid group practice and persons attending a few other outpatient facilities. Sampling of enrolees in the group practice continued until the yield of new and usable statements diminished markedly (Bergner, 1976b). A basic catalogue of 1100 statements was reduced to 312 unique items in 14 categories. The Nottingham Health Profile (NHP) generated its pool of items through a survey of 768 patients with acute and chronic ailments (Hunt et al, 1981). Items from the SIP were used in addition. The NHP contains 38 items grouped into six sections, namely physical abilities, pain, sleep, social isolation, emotional reactions and energy level.

The EuroQol group (1990) reviewed the health state description systems developed by the above studies to arrive at a parsimonious set of dimensions. The group sought to develop an instrument of generic health status measurement across multiple cultures. Table 6.1 shows the dimensions arrived at by studies leading to the three scales described above and the five dimensions adopted by EuroQol (EQ-5D). Except alertness and energy level, all other dimensions from SIP, QWB and NHP scales are represented in the EQ-5D system. Note that cognition did not appear as a distinct dimension in any of these scales.

The Rand Health Insurance Experiment, followed by the Medical Outcomes Study (MOS), systematically collected items to describe various aspects of health and studied their properties for construction of a generic health status measurement tool (Stewart, 1992, Brook et al 1979). The Short Form-36 (SF-36) instrument is an outcome of these extensive studies. SF-36 includes multiple items organised under eight dimensions[4] (Table- 6.2). Cognition appeared in these studies as a distinct dimension. Most of these map to the EQ-5D system, except cognition, health perceptions, energy-fatigue, and physical-psychological symptoms.

The EQ-5D description system appears to be strongly rooted in its conceptual lineage to an ideal definition of health and its linkage to empirically rooted health status descriptions. Its emphasis on cross-cultural validity and

[4] The number of dimensions inSF-36 can be viewed as four. Please see page 112-113 more discussionof SF-36

feasibility of measurement are very attractive. However, the lack of cognitive dimension and the restriction of severity levels within each dimension to three, leaves us with some handicaps. Cognition, changed ratings for conditions with lower levels of disability in her dimentions. Valuations for conditions with severe levels of disability in other dimensions did not change much. Restriction of severity levels to t hitherto taken for granted, is clearly an important attribute of human health. Diseases affecting cognitive functioning are now being recognised. Recent research in a EuroQol member centre suggests that the addition of cognition as the sixth dimension, would make the EQ-5D system more comprehensive (Krabbe et al, 1998). These authors found that the inclusion of cognition changed ratings for conditions with lower levels of disability in other dimensions. Valuations for conditions with severe levels of disability in other dimensions did not change much. Restriction of severity levels to three may be a reason for the insensitivity of EuroQol to minor and trivial illnesses.

Table 6.2: Mapping of MOS* dimensions to EQ-5D.

Medical outcomes study (MOS)		EQ-5D
Mobility	Getting around in the community	Mobility
Physical functioning	Walking, climbing stairs	
	Self care	Self care
Role functioning	Performance of usual role activities such as working at a job, housework, child care, community activities and volunteer work	Usual activities
Pain	Subjective feeling of bodily distress of discomfort such as headaches, backaches.	Pain-Discomfort
Social functioning	Functioning in normal social activities with family, friends, neighbours, marital functioning, sexual problems.	
Psychological distress /wellbeing	Positive and negative psychological states including anxiety, depression, behavioural emotional control, loneliness, positive affect, feelings of belonging.	Anxiety - Depression
Sleep	Quantity, disturbance, adequacy of sleep	
Health distress	Psychological distress due to health	
Cognitive functioning	Cognitive problems, such as forgetfulness, difficulty in concentrating.	
Health perceptions	Personal evaluations of health in general, including current and prior health, health outlook, resistance to illness.	
Energy / fatigue	Feelings of energy, pep, fatigue, tiredness	
Physical / psychological. symptoms	Subjective perceptions about the internal state of the body, such as stiffness and coughing.	

*Source: Stewart Anita L.; The Medical Outcomes Study framework of health indicators, in Anita L Stewart and John E. Ware Jr. Eds, Measuring Functioning and Well being, Duke University Press, Durham, 1992, pp23-24.

Attention span and cognitive capacity of human mind to process multidimensional information:

Research in the field of psychology suggests that there is a limit to our capacity to process information (Saariluoma, 1998). Miller (1956) suggested that human beings process about 5 - 9 attributes (chunks of information) at a time. More recent evidence from research in working memory suggests that human capacity to simultaneously process multi-attribute information may range from 3 to 5 rather than 5 to 9 as was thought earlier (Halford, 1998). These findings imply that the number of dimensions used to describe the states should be kept as minimum as feasible, to allow adequate processing of health state descriptions by valuers. Recognising the need to keep the information load on valuers within manageable limits, researchers have tried to simplify health state description systems. For example, Brazier and others (1998) simplified the SF-36 profiles to a six-dimension (SF-6D) description system, which was used to obtain a holistic valuation of health states to be used for estimation of QALYs. Froberg and Kane (1989) propose that the number of dimensions in a description system should not exceed nine, and should preferably be less. Reviewing empirical evidences on the mode of presentation of health states, Froberg and Kane (1989) conjecture that "moderately detailed health state descriptions yield more accurate judgements of preference than either very scant descriptions or very lengthy descriptions that run the risk of overloading the rater's information processing capacity".

Statistical analysis of multi-attribute measurements:

The number of dimensions have implications about the type of statistical analyses that can be done on directly measured health state values. Froberg and Kane (1989) have referred to Fischer's overview (1979) which found that with six or fewer dimensions, functional measurement and explicit decomposition procedures assigned similar values to a health state. The reliability of multi-attribute judgements deteriorate with larger number of dimensions. Froberg and Kane have referred to other investigators (Llewellyn-Thomas et al, 1984; Lyness and Cornelius, 1982) who found that when only a few dimensions are involved, multiattribute judgements are more reliable than decomposed judgements. Thus, parsimony of dimensions is important to retain the holistic property of a description used for operational purposes.

How to convey health state descriptions effectively to an individual undertaking the valuations:

Effective communication of the description to individuals acting as valuers has many difficulties. The descriptive system must be comprehensible to the young, middle-aged and older adults with widely varying levels of educational

attainment, socio-economic and cultural backgrounds. For example, differences have been found between using paragraphs written in the first person in describing a health state, and using straight lists of levels in each domain of health (Llewellyn-Thomas et al. 1982). The descriptive system should be meaningful across cultures. Translation of instruments should produce equivalence in terms of word meanings and idioms i.e. semantic and idiomatic equivalence; equivalence in terms of situations and concepts evoked in the descriptions i.e. experiential and conceptual equivalence, respectively (Guillemin et al, 1993). The description system should enable communication with semi-literate as well as illiterate persons. The description systems used so far have been developed for literate societies like North America and Europe. Even here, studies have experienced communication difficulties due to language barriers. For example, in the Canadian study by Sackett and Torrance (1978), about 12% of the randomly selected sample had to be excluded, because the interviewees could not communicate in English. One way to deal with this problem is to supplement written descriptions with appropriate graphical representations. Some researchers have used multimedia methods for valuation exercises (Lennert and Hornberger 1996, Lennert and Soetikno 1997). One problem with multimedia solutions is that the computer may be a source of distraction, particularly where the general community has limited experience with multimedia. In any case, multimedia solutions need a graphical description system to start with. So description systems for partially literate and multi- lingual communities should ideally include a graphical description sub system.

Approaches to measurement of health state values:

Measurement strategy:

Torrance (1986) gives an overview of measurement of health state utilities for economic appraisal. In a four-part series, Froberg and Kane (1989 I-IV) have summarised methodology for measurement of health state weights. The first article deals with measurement strategy, which refers to the overall structure for posing questions to respondents and the corresponding method of data analysis. They divide extant measurement strategies into (a) holistic and (b) decomposed approaches. Torrance (1986) viewed these as alternative description of health states and used a similar classification, labelling the decomposed approach as health state classification system. In the holistic approach, each of the full range of health states is described as a whole, including all its attributes. The respondent is presented with the description for all health states one after the other and asked to assign values. Thus the respondent has to judge each health state as a whole and all health states in the scale. As a result, the procedure becomes cognitively demanding for the respondent. In decomposed designs, the respondent does not have to value all health states

in the profile. Decomposition may be explicit or achieved by statistical modelling. For statistically inferred decomposition, the respondent is presented with the multi-attribute description of a health state as in case of holistic approach. But only a few health states are presented to a single respondent, thus reducing cognitive overload. An algebraic model of multi-attribute health states is constructed using statistical inferences from respondent evaluations. Explicit decomposition for health state measurement is rooted in multi-attribute utility theory (Torrance 1982, 1986). Here the respondent is asked to evaluate each dimension of health separately, thereby keeping it cognitively simple. Froberg and Kane (1989) recommend the statistically inferred strategy in view of its simplicity for respondents.

Scaling methods:

In the second article, Froberg and Kane (1989) list six scaling methods. These are; (a) standard gamble, (b) time tradeoff, (c) rating scale, (d) magnitude estimation, (e) equivalence or person tradeoff, and (f) willingness to pay. Rating scales and magnitude estimation methods are psychometric in nature. Standard gamble, time tradeoff and willingness to pay are all preference- based. Person tradeoff is preference-based and has psychometric origins as well.

Standard gamble:

The standard gamble measures present a random prospect (gamble), consisting of a best and a worst outcome, and the alternative of a certain prospect intermediate in desirability between the best and worst outcomes. The random prospect is completely defined by probability p of one of the two possible events. The other event will have a probability $(1-p)$. For example, the best outcome in the random prospect can be "perfectly healthy" and the worst outcome can be "dead". The certain prospect of intermediate desirability will then be a specific health state: for example, living with a particular morbidity. To obtain quality adjusted weights, the random prospect is defined by assigning probability p to the "perfectly healthy" outcome. The two alternative prospects defined by choice of disease state for the certain arm and p for the random arm are presented to a respondent with some initial value of p. The value of p is then varied till the respondent is indifferent between the certain prospect and the random prospect. The p that satisfies this condition is taken as the quality adjustment weight for the disease condition defined in the certain arm. The disability weight for this health state is given by $(1-p)$. Instead we can define the random prospect by assigning probability q to the worst outcome (death in this set up). In that case the value of q satisfying the indifference condition is the disability weight. The standard gamble was proposed by Von Neuman and Morgenstern (1944) as a tool to measure expected utility. The difficulty with this method is that it is not easily understood by many respondents.

Time tradeoff (TTO):

Time tradeoff was designed by Torrance et al (1973) for health status measurement, as a simpler and cognitively less demanding alternative to the standard gamble. Here the valuer's preferences for health states is assessed indirectly through the time (number of years, months, days, hours) (s)he wants to trade in to lead a completely healthy life, as against life with a particular less-than-perfect health state. The subject is offered two alternatives. Alternative 1: state *i* for time *t* (local life expectancy of an individual with the chronic condition) followed by death. Alternative 2 is a perfectly healthy state for time *x*, where *x* is less than *t*. Time *x* is varied until the respondent is indifferent between the two alternatives, at which point the required preference value for state *i* is given by $h_i = \frac{x}{t}$

Rating scales:

The rating scale consists of a range of values with clearly defined end-points (anchors). For holistic health status measurement, "perfect health" and "death" act as natural anchors. The range of values could be continuous or discrete. The visual analogue scale (VAS) consisting of a graduated line segment, with one end labelled as death and the other labelled as perfect health, is a continuous rating scale. Another form of continuous rating scale uses adjectival labels to describe the intermediate points of a line anchored at both ends (for example, see Fig 4.4b at page 34 in Streiner and Norman, 1995). Except for the intermediate labels and smaller line length, they resemble the VAS. Rating scales using equally appearing intervals i.e. discrete points along the scale are called category rating scales (for example see fig 4.4a at page 34 in Streiner and Norman, 1995). Streiner and Norman (1995) describe rating scales as direct estimation methods (Chapter 6) and category or continuous rating scales are described as adjectival scales. Rating scales are most frequently used to measure health state weights. Specter (1992) describes the theoretical basis as well as practical steps in construction of rating scales for general psychometric use. Streiner and Norman (1995) provide a similar description for the construction of rating scales in the field of health status measurement.

The visual analogue scales are simple in construction, but respondents may not always agree. For example, Streiner and Norman (1995) cite a study by Huskisson (1974) in which 7% of patients were unable to complete a VAS against 3% for category rating scale. Bosi Refaz et al (1990) found that illiterate respondents had more difficulty with VAS as compared to category rating. In a comparative study of VAS and person tradeoff method, Nord (1991) asked his respondents to describe the meaning they attached to points in VAS chosen by them. Both ends of the VAS was anchored by worst and best imaginable health states respectively. Sixty seven (out of a total of 105) respondents answered

this question, of which nineteen persons said that they viewed it as a percentage of best imaginable health state, eleven said it did not mean anything and the remaining 37 did not answer the question directly but described specific dimensions of given condition taken into consideration by them. Nord observed that subjects expressed strength of preferences, through the VAS, in addition to ranking of health states. Person tradeoff implied by VAS was lower than the ratios obtained from directly asked person tradeoff questions. With this finding, coupled with the small size (19) of respondents reporting a conscious attempt to relate given state in percentage terms to one end of the scale, Nord concluded that one should not lay too much emphasis on the numerical values obtained from VAS. But such an interpretation is flawed by many limitations of this study. Firstly, Nord appears to be denominating the 19 persons, consciously trying to assign a ratio number, with the 67 persons who gave some response to the question on meaning of valuations. More than half (37) of them did not give a direct answer to this question. So the appropriate denominator to appreciate the relative size of the 19 consciously trying ratio raters would be the 30 persons who gave a direct answer to this question. Secondly, Nord uses person tradeoff as a criterion to compare results from VAS. The person tradeoff methodology is itself very sensitive to sample size, framing and start point bias found subsequently by Nord (1995).

Magnitude estimation:

In magnitude estimation a reference state is identified and described. A numerical value (numerical estimation) or a line segment of certain length (line production) is assigned to the reference state. Respondents are presented with the health state of interest and asked to assess how it relates to the reference state— more specifically, how many times better or worse the given health state is in comparison to the reference state. They then assign a number proportional to the reference number (numerical estimation) or produce a line (line production) to show how it relates to the reference state. A higher number or larger length of line would mean better health state and corresponding a smaller number or line would mean a worse health state. Calibration of respondents, ability to reproduce magnitude estimates or at the least a practice session to orient them towards magnitude estimation is usually called for. Apart from the brief description in Froberg and Kane (1989b), McDowell and Newell (1987, 1996), the magnitude scaling method is described by Lodge (1981).

Person tradeoff (PTO) or Equivalence:

Patrick et al (1973) adapted the method of adjustment or equivalent stimuli in psychometrics (Guilford, 1954; Torgerson, 1958) and devised the equivalence method for health status measurement. Froberg and Kane (1989b) opine that the person tradeoff (PTO) method is conceptually similar to magnitude

scaling. In recent literature the same is referred to as person tradeoff method (Nord 1992, 1995). Here, respondents are presented with two groups of persons distinguished by their health state and a constraint to the effect that only one of the two groups can be helped. One of the two groups is characterised by a reference health state (standardgroup): for example, people with perfect health. The other group (evaluation group) is characterised by a specific health state of interest. The group with perfect health constitutes the standard stimuli and number of persons in the evaluation group provides the variable comparison stimuli. Respondents are asked to choose between the two groups who should receive help. The number of persons in the evaluation group is varied till the respondent is indifferent between the two. Miles (1977) compared the person tradeoff method with category ratings. Difference in health states between the two methods was not significant. Patrick et al (1973) reported that this method was too complex. The unrealistic assumptions and the emotive nature of the task offended some judges. Similar observations were made by Nord (1995). In addition, Nord (1995) noted that the techniques needed fairly large groups of respondents to keep measurement error within limits, was susceptible to start point bias, and was sensitive to question framing. To overcome influence of question framing, respondents should be induced to arrive at a reflective equilibrium. To achieve this, Nord (1995) suggested a multistep procedure, through which respondents are presented all relevant arguments and to reconsider initial responses in the light of such arguments. A problem with this solution is: what arguments are considered relevant? For example, Froberg and Kane (1989c) observed that "when an interviewer takes an active role, the potential for influencing the rater is increased". For purposes of population health status measurement and resource allocation, this method is assumed to be valid by construction, since it asks respondents to weigh the claims of one group of persons with respect to another, distinguished only by their health states (Patrick et al, 1973; Nord 1991, 1995; Murray 1996). Knowledge of its reliability is limited, since the method has not been used widely enough (Froberg and Kane, 1989d).

Murray (1996) used the PTO method to derive disability weights for the fifth revision of GBD estimates. A good deal of effort was taken to minimise the usual problems associated with this method. To induce deliberation and reflection, two PTOs were designed using different situational characteristics. In the first method (PTO1), respondents trade life extension for disabled persons with life extension for healthy persons. In the second method (PTO2), respondents trade improving health related quality of life to complete well being with life extension for perfectly healthy persons. In general, respondents were found to assign a lower disability weight to the same health state in PTO1 compared to PTO2. Once the respondent completes the two PTOs for a given

condition, the results are fed back to him while pointing out inconsistency, if any, between them. The respondent is then encouraged to revise the estimates to resolve the inconsistency. After evaluating all of the 22 indicator conditions, respondents are asked to give an ordinal rank to these indicator conditions. This ordinal ranking is compared with the ranking implied by the reconciled PTO mentioned above. Any inconsistency is fed back to the respondent with instructions to reconcile them. Group discussions are encouraged at different stages to facilitate a consensus. Details of this protocol (GBD PTO protocol) are described in Murray (1996, pages 35-41 and appendix 1 at pages 90-98). A major deficiency that persisted is lack of standard description of health states. Since most participants were chosen from health professions, it was assumed that they would have a shared understanding of the health states in question. Interactions between interviewers and respondents as well as the group discussions were not fully structured.

Willingness to pay:

Willingness to pay is a straightforward application of welfarist notion of health status as a personal utility similar to other goods and services in the economy. A person's willingness to pay for curing or avoiding impairment, disability and handicaps associated with a health state, is taken as a measure of quality adjustment weight for that health state. Willingness to pay to avoid the risk of death can be used as a denominator to derive quality adjustment weights in the range of [0,1]. Apart from ethical objections concerning the welfarist approach to health and economic valuation of human life, the measurement of willingness to pay is complicated by lack of a perfectly (or nearly perfect) competitive market in the health sector. Various alternatives like the contingent valuation method and cost of illness method are used to get around this problem. A comprehensive description of these methods can be found in Tolley, Kenkel and Fabian (1994).

Controlling context effects:

Froberg and Kane (1989c) in their third article review observations about the effect of respondent characteristics and other contextual characteristics on health status measurement. They group these factors into three clusters, (a) differences among population (respondent characteristics); (b) inconsistencies due to the nature of human judgement process (framing); and (c) inconsistency due to situation-specific variables (situational differences). They observe that respondent characteristics do not have any significant effect on valuation and hence can be ignored. Framing effects can be reduced by presenting the problem in more than one way and asking the rater to reconcile inconsistencies. This is identical to the multistep approach (reflective equilibrium) suggested by Nord

(1995). The solution to situational differences proposed by Froberg and Kane (1989c) is to standardise them.

Cross-cultural validity:

The validity of an instrument across different cultural settings will depend on the conception of health from which it arises and the nature of dimensions included in it. If the judgement on some dimensions is influenced by specific cultural characteristics, the instrument would not perform well in settings outside the place of its origin. Schumacher and Naughton (1995), recognising the need for portable health status measurement instruments for international use, proposed that domains of health-related quality of life be restricted to those "universally most essential to one's ability to pursue valued life goals". Most instruments purporting to measure health-related quality of life do include such universally useful dimensions such as physical, social and psychological functioning, mobility and self care, and emotional well being. Although this principle is not in doubt, there is scope for wide ranging interpretations of the very generic characterisation of universally useful dimensions.

Another practical issue is ease and accuracy of translation. The translated instrument should retain its original validity and reliability characteristics. Leplege and Verdier (1995) have described methodological aspects of translation of health status measurement instruments. Generic instruments, which have been translated to different languages (Shumaker and Berzon 1995 appendix-1) include (a) the Short Form Health Survey (SF-36); (b) Nottingham Health Profile (NHP); (c) Sickness Impact Profile (SIP); (d) EuroQol and (e) Dartmouth COOP functional health assessment charts. Out of these five, EuroQol is the only instrument to generate a composite index of health status. All other instruments produce general health profiles (McDowell and Newell, 1996).

A further issue pertains to the feasibility of administration in partially literate population. Here the scope for self-administration of questionnaires is limited. Local experience in the measurement of health-related quality of life is almost non existent. So even for interviewer- administered questionnaire, the items need to be simple and straightforward in order to ensure acceptable levels of compliance by interviewers and interviewee.

An overview of instruments for general health status measurement:

A number of instruments are now available for health status measurement. McDowell and Newell (1987, 1996) provide an overview of these and distinguish two general classes, namely (a) instruments for measurement of general health status and (b) instruments designed for specific dimensions of health: for example physical disability, psychological well being, pain etc. Our interest here pertains to the first category. McDowell and Newell (1996,

Chapter 9, pgs 380-492) list 21 instruments for measurement of general health status. Eighteen of these instruments, called general health profiles, generate a profile of scores in different dimensions included in the instrument. Another three allow for the computation of a single index from out of the scores in component dimensions. These are called health indices.

We will briefly describe four of these instruments, namely: (a) the Sickness Impact Profile (SIP); (b) Short-Form-36 (SF-36) Health survey; (c) EuroQol, and (d) the Quality of Well Being scale (QWB). The first two only give a general health profile. The last two produce general health indices in addition to profiles.

The sickness impact profile (SIP) seeks to measure changes in a person's behaviour on account of illness. Scoring is done along 12 categories or sub-dimensions. Respondent behaviour in each category is assessed by a set of questions graded according to severity or intensity, along the sub-dimension. There are altogether 136 such graded questions. The 12 sub-dimensions can be grouped into (a) physical health consisting of ambulation, mobility and body care; (b) psycho- social health consisting of social interaction, alertness, emotional state and communication; and (c) five independent categories, namely (i) sleep and rest, (ii) eating, (iii) work, (iv) home management and (v) recreation. Item weights were arrived at from more than 100 judges with scaling procedures at equal interval. The profile can be administered either by self or by an interviewer. It takes about 20 to 30 minutes to complete the questionnaire. The scale was developed by Bergner and others (1976a, 1976b, 1981).

The SF-36 instrument measures eight dimensions, namely (a) physical functioning, (b) role limitations due to physical problems, (c) pain, (d) social functioning, (e) general mental health, (f) role limitations due to emotional problems, (g) vitality, energy or fatigue, and (h) general health perceptions. Physical functioning and role limitation due to physical problems can be viewed as one dimension. Similarly (d), (f) and (g) can be viewed as social functioning. Thus the dimensions covered in this instruments can be summarised as (a) physical function status, (b) social function, (c) psychological well being and (d) pain. Each dimension is assessed by a category rating of multiple items which are themselves graded by severity or intensity. The form uses the preceding one month as the time frame for all questions. Alternative forms using shorter time frames for acute conditions have also been used. Reliability and criterion validity (using common sense criteria like ability to work, symptoms, etc.) have been found to be fairly high (McDowell and Newell, 1996). It was developed out of the health insurance experiment and medical outcomes study (MOS) conducted by the RAND corporation in US. The SF-36 instrument has been

described by Ware and Sherbourne (1992), Ware and others (1993), McHorney and others (1993, 1994), Aaronson and others (1992).

The EuroQol is a summated rating scale consisting of five dimensions. These are mobility, self care, usual activities, pain or discomfort and anxiety or depression. Each dimension is rated by a three-point category rating scale. Weights for computation of composite index were developed by using valuation of 10 core health states on visual analogue scales. The instrument has four parts as follows: (a) description of patients own health (page 2) along five dimensions; (b) overall rating of own health using a visual analogue scale (page 3); (c) valuation of a standard set of health states (pages 4-7); and (d) background information about the respondent (pages 8-9). Parts (a) and (b) are required to collect data on health-related quality of life. A general health index can be computed using weights derived by the EuroQol team using responses from the valuation part (pages 4-7). For local weights these parts have to be implemented as well. The EuroQol instrument is described by the EuroQol group (1990), Brooks and the EuroQol group (1996), and McDowell and Newell (1987, 1996). The full instrument is reproduced in Shumaker and Berzon (1995 appendix -2).

The Quality of Well Being (QWB) scale is a summated rating scale consisting of three dimensions, namely (a) mobility, (b) physical activity, (c) social activity, and (d) the symptom problem complex. Estimation of QWB index proceeds in the following three steps: (a) assessment of functional status profile, (b) Scaling of responses to derive dimension-specific weights for the composite index, and (c) estimation of transition probabilities to derive expected duration in each health state. Construction of synthetic measures like DALY, already takes into account the expected duration in each health state. So measurement of disability weight or its complement quality adjustment weight requires only the first two steps i.e. assessment of functional status and dimension specific weights. The authors of the scale have derived a set of dimension-specific weights from valuations by 867 raters and using an equal appearing interval rating procedure (Kaplan, Bush and Berry 1976; 1979; Patrick, Bush and Chen 1973a-b; Blischke, Bush and Kaplan 1975). This instrument has been described by Kaplan, Anderson and Ganiats (1992), and in McDowell and Newell (1987, 1996). QWB was used to gather community valuations for different health states by the Oregon health services commission (OSHC). The scale was criticised when it was found that weights assigned by it to certain health states were clearly counter intuitive. However, it has been pointed out that the QWB scale was not properly applied by the OSHC and hence the counter-intuitive results could not be attributed to the scale.

Review of literature on health status measurement and health state

valuation gives an idea about the efforts by many researchers to deal with the various theoretical and practical issues in measurement of health state values. Many measurement tools have been tried, mostly on captive populations recruited from university campuses. The visual analogue scale has been used in population surveys on account of its simplicity and feasibility. The description of health states for effective communication in partially literate multilingual communities is an important issue that needed to be addressed for this study. In the absence of a perfect instrument for measurement of health state values, reliability and validity of obtained measurements assume significance. In the following section we describe the health state valuation study in Andhra Pradesh.

Methods of the health state valuation study in Andhra Pradesh:

General study design:

A study was conducted in Andhra Pradesh, at the Institute of Health Systems (IHS), to measure peoples' preferences about various health states. The broad goals of this study were:

1. To strengthen the methodological foundation for population-based measurement of health state weights. Because population-based empirical assessments of health states are extremely limited, new surveys are needed. In addition, valuation protocols must be adapted for use in partly literate populations like those in India.
2. To measure local preferences for health states, to be used for estimation of national and state burden of disease in India.

Two distinct sources of assessment were made. One arm of the study consisted of a series of multi-method deliberative health state valuation (MDHSV) workshops for educated persons from different backgrounds. Participants of these workshops valued health states using more than one procedure card sort, including, visual analogue scale, time tradeoff, and person tradeoff methods. The second arm of the study involved measurement of valuations given by general population drawn from a rural area, done by conducting household surveys in Kondakkal village in Ranga Reddy district of Andhra Pradesh. Respondents in the community survey were requested to give their valuations using card sort followed by visual analogue scales.

Selection of index health states for the study:

Although synthetic health status measures would require health state weights for a large number of conditions, a health state valuation study has its limitations in generating direct valuations for a large number of conditions. Health state valuation exercises are usually characterised by high cognitive load on the valuers. Hence a valuer is usually given a limited set of conditions. Increasing the number of conditions would lead to valuer fatigue, resulting in poor validity and reliability of the measurements. A practical alternative is to use a small set of conditions to measure peoples' preference and then statistically infer the health state values for other states. The small set of conditions used to gather measurements on peoples, valuation of health states, is referred to as index health states or synonymously indicator conditions.

Guidelines for choice of indicator conditions:

Before proceeding with the selection of indicator conditions, we take stock of principles and guidelines relevant for choice of index health states. Parts of the health state valuation exercise require the valuer to deliberate about all health states at hand and assign values to each. Hence, the number of health states presented to a single valuer should be kept within manageable limits to minimise incidence of cognitive overload. This limit is a function of human working memory[5] and cognitive behaviour, and is independent of sample size. The limit is all the more significant for population-based surveys, where multiple sessions would not usually be feasible. Gudex et al (1996) limited the number of health states per valuer to 15, for a population-based survey in UK. Sacket and Torrance included 10 health states for their population based survey in Hamilton, Otario Canada (Sackett and Torrance, 1978). The EuroQol group (Brooks, 1996) found, it was feasible for a person to value about 12-16 health states using the EuroQol instrument.

Secondly, study design and data analysis considerations should be kept in view. For example, the set of index health states should represent widest range of health state profiles. This will allow statistical inference of health state weights for conditions for which no direct measurement is available. Hence it is important that the set of indicator conditions maximise the independent variation in each dimension of health status. The need to maximise independent variation along each dimension and the constraint imposed by the ability of valuers in dealing with a number of conditions, run counter to each other. Since the limit on number of health states for valuation by an individual is fixed on psychometric grounds, the only scope to increase variation in dimensions of health status is to increase sample size and present different sets of conditions to different

[5]See the discussion earlier about attention span and cognitive capacity of human mind to process multi dimensional Information.

persons. Comparability and linkage between different health states would require that some health states be common for all valuation sets: these are known as core health states. The inclusion of a few pairs of dominating and dominated health states will facilitate inference about the validity of measurement protocol. A state that is unequivocally worse than another state in at least one dimension and at most equivalent in all others, is considered a dominated condition. One that is unequivocally better than another state in at least one dimension, and at the least, equivalent in all, is considered a dominating condition. Examples of dominating and dominated pairs of health states are: amputation of one leg below the knee, and amputation of the both the legs below the knee, quadriplegia and paraplegia. Then the proportion of times that ranks for dominated states are inverted can be used as a measure of (or lack of) the health state comprehension.

Finally, local relevance and familiarity of health states and its effect on valuer motivation has to be considered. Health states corresponding to diseases not found or found very rarely in the local population should be avoided. Since a descriptive label for each health state is retained to facilitate synthetic comprehension by the valuer, going for conditions that are hitherto unheard-of would not be helpful. Such conditions might demotivate valuers due to lack of apparent purpose of the exercise. Instead, at least some health states corresponding to diseases highly prevalent and / or considered important public health problems in the local area should be included. This is required to increase valuer's motivation and allow for a sense of purpose in the valuation exercise. On the other hand, health states associated with disease labels known to be a taboo, to provoke a sense of outrage or subject to widely prevalent stereotypes, should be avoided. The disease label is likely to dominate the valuer's thought process to near total exclusion of the health state profile. There would appear to be some conflict between the need to avoid stereotypes and the goal of seeking locally prevalent health states, described earlier. Indeed, there would be some diseases which are highly prevalent and carry a strong stereotype. So, quite naturally, the art of selecting indicator conditions requires careful weighing of potential confounding due to various factors, and the feasibility of getting valuer's co-operation.

Health states chosen for this study:

The number of health states to be valued by a single valuer was limited to 11, to minimise incidence of cognitive overload, and valuer fatigue. The study, however, sought to directly measure disability weights for a larger set of conditions. According to original plan, 22 health states were chosen for the study, apart from the valuer's own health state. Six of these health states were used as the core, common to all valuers. The remaining 16 health states were divided into four subsets. Each subset added to the core subset, made a set of

health states. Accordingly, the study planned for four sets of health states. The six core conditions were selected to represent a broad range of 6D5L profile and disability severity weights. While making up these lists, conditions known to be prevalent in Andhra Pradesh were included, and conditions not found in AP were excluded, while trying to keep the list as comparable with the ones used at other study sites for international comparability. Thus, the selection of indicator conditions involved several rounds of discussion between local investigators at IHS and study co-ordinators at WHO. Sets were systematically assigned to balance assignment of valuers to different sets of health states.

The 6D5L health state description system:

The 6D5L description system is developed by expanding upon the EuroQol (EQ-5D) description system. Cognition has been added as the sixth dimension. Severity levels in each dimension are described using five levels instead of three. The EQ-5D system allowed for a maximum of 244 distinct health states[6]. This restricted the system's ability to discriminate between moderate to small differences in functional status. The 6D5L system gives rise to distinct health states. Some of these states may not exist in practice, for example 555555 (a person with total loss of cognitive function would not be anxious). Even then, the system provides for description of a fairly large number of distinct health states. We hope that this will improve 6D5L's sensitivity to minor illnesses.

The 6D5L health state description system, developed for the AP health state valuation study, consists of the following distinct parts, each of which is described below.

1. A written description of dimensions and severity levels.
2. A Telugu language version of the dimensions and severity levels.
3. Locally valid graphical representation of dimensions and severity levels.
4. Identification protocols. Procedure to identify health state descriptions of diseases, clinical and epidemiologically encountered conditions.
5. Coding schema to represent different health states.

Written description of dimensions and severity levels:

Since most valuers were to come in contact with the description system for the first time, we anticipated that they may have difficulty in interpreting the six dimensions and discrimi- nating between them. Hence a set of explanatory notes entitled "What this dimension represents?" was developed to reliably communicate aspects of health represented by respe- ctive dimension. These notes first explain what are included in the dimension. Then an example of a

[6]Five dimensions with three levels in each give rise to 3^5=243 permutations. Death is added to this.

condition, that does not affect the dimension at all, is given, followed by an example of conditions that may affect the concerned dimension. Published literature on functional status measurement including Activities of Daily Living (ADL), Instrumental Activities of Daily Living (IADL), Pain Measurement questionnaires (McDowell and Newells 1987, 1996), health-related quality of life measurement scales like the EuroQol (Brook, 1996), SF-36 (Ware and Sherbourne, 1992), etc. were reviewed to cull out expre- ssions that may explain, elucidate, clarify or discriminate the concerned dimension. Such expressions have been used in the "What this dimensions represents" part of the descriptive system.

These expressions have been taken from many articles and functional status measurement scales. Often more than one article or scale provided similar expressions. Hence it has not been feasible to acknowledge all sources of these expressions. During the study, we found that the expression "usual activities" is easily confused with self-care in the Indian context. Hence the third dimension, namely usual activities, was assigned an alternative label of work and leisure[7]. Box 6.1 shows the written description of mobility dimension. Descriptions for all dimensions is provided in Appendix 6.1.

Box 6.1 Written description of mobility dimension

Mobility (Position = 1):

A. What this dimension represents:

1. Transfers: Includes the management of all aspects of transfers to and from bed, mat, toilet etc. More simply getting in and out of bed.
2. Ambulation: Includes coming to a standing position and walking about,
3. Stairs and environmental surfaces: Ability to handle environmental barriers, and includes climbing stairs, curbs, ramps or environmental terrain,
4. Community mobility: Ability to manage transportation.
5. Example of a condition that does not affect mobility: Vitiligo
6. Example of conditions that may affect mobility to various degrees: Back ache, paralysis of lower limbs.

B. Severity Levels and Codes (SLC):

1. Independent, i.e. no assistance required and no problem with mobility. Ability to run / flight in times of need. SLC =1
2. Occasional or very few problems in moving about. SLC =2
3. Some problems in moving about. SLC=3
4. Many problems in moving about. SLC=4
5. Unable i.e. totally dependent for mobility. SLC=5

[7]Since this fact was found mideway through the study, all instruments and printed material continued to have the label 'usual activities' but interviewers and workshop co-ordinators were instructed to clarify to valuers as to the correct meaning of this dimension

Telugu version of the written description system:

A panel of doctors and nurses practising in the area were invited for a health state description workshop. Tasks assigned to the workshop included (a) Telugu translation of the 6D5L description system, and (b) Telugu translation of disease labels. The Telugu translations obtained from the health state description panel was the initial document. The draft of the translated document was further worked upon with help of other faculty knowledgeable in Telugu, to arrive at a provisional draft. The provisional draft was then discussed with experts in Telugu literature. They were requested to provide alternate translations. Subsequently persons who were not aware of our list of health states were given the provisional Telugu drafts and were asked to translate them back to English. The translations that resulted in the original English version were chosen for the Telugu version, which has been reported elsewhere (Mahapatra and others, 2000).

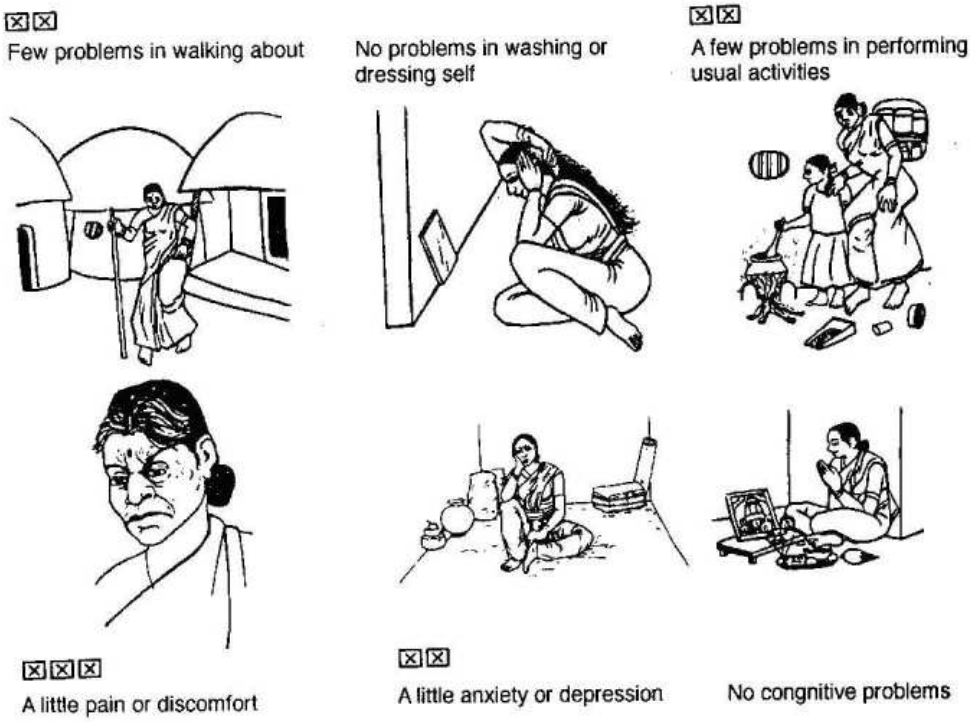
Locally appropriate graphical representation of dimensions and severity levels:

To facilitate communication of the 6D5L description system to semi-literate and illiterate valuers in the general population, we planned to develop a graphical description system for the 6D5L profiles. First an Artist brief was prepared, explaining the 6D5L description system, and describing the nature of task at hand. The brief gave examples of some what similar graphical representations, such as the Dartmouth Coop Function Charts (Nelson and others, 1987) and Faces scale (Andrew and Withey, 1976). The art team's task was to arrive at the most appropriate pictorial representation of the severity levels under each of the six health dimensions. A team of Fine Arts students from the School of Performing and Fine Arts, University of Hyderabad were identified with the help of the School's faculty. This team of artists worked to draw multiple set of the graphics of the five severity levels in each of the six dimensions. We found that the scaling and reproducibility of graphics using more of lines and less of shades was better. To minimise gender bias, separate sets of graphics were developed using female and male characters respectively. To facilitate preparation of health state description cards etc., once again art works with more of line drawings and less of shading were preferred. Artists were asked to make sure that the characters used in the pictures satisfied following criteria.

1. Features of characters are similar to the local population.
2. Dress is consistent with the dress pattern prevalent in rural areas of Andhra Pradesh for the respective gender.
3. Background, foreground and other artefacts in the pictures that are consistent with the rural scenario in Andhra Pradesh, and
4. Activities shown in the picture are consistent with usual roles for respective gender, currently prevalent in the area.

Figure 6.1 : 6D6L Graphics for Self Care

Figure 6.2 : Continuous Moderate Back Pain



The pictures were reviewed many times. Persons not directly involved in the study were shown different sets of the pictures and asked to interpret them. The pictorial systems that were perceived by these judges to represent the severity levels of the corresponding dimension were selected and used. The graphical description system consists of a collection of two sets of pictures, with six subsets in each. Each subset in turn represents one of the six health dimensions and consists of five pictures to represent the five severity levels. Thus the basic element of the 6D5L graphical description system is a picture meant to convey a given level of severity in a particular health dimension. Altogether 30 such pictures consisting of five for each of the six health dimensions, and with a female person as the primary character constitutes the graphical description system for females. A similar set is prepared for the males.

Figure 6.1 shows a set of five pictures for a single dimension (self-care) using male characters. A health state can be pictorially described by choosing the appropriate picture from each of the six subsets. So a graphical 6D5L profile would consist of a set of six pictures. For example, Figure 6.2 shows the 6D5L graphical profile for continuous moderate back pain in a female character. Figures with Telugu labels were used for the general population survey and those with English labels were used in MDHSV workshops. A complete set of these pictures and Telugu labels have been reported elsewhere (Mahapatra et al, 2000).

Identification of 6D5L profiles:

Identification of 6D5L profiles may be required in the following two situations. Firstly, description of typical functional status of disease states. Here, we use the term disease state to include clinically and epidemiologically encountered conditions, which may not necessarily be considered disease states. Secondly, identification of labels for specific 6D5L profiles to facilitate holistic processing of 6D5L information by valuers. The need for association of labels to 6D5L profiles for purposes of valuation and how we arrived at the labels used in this study, is described later. We will discuss here the need for mapping of specific disease states to 6D5L profiles, and then proceed to describe our efforts to implement the same.

Health state valuations are usually obtained for incorporation into summary measures, which may be computed to allow for disaggregated analysis. If disaggregated analysis is required, then identification of 6D5L profiles for disease states becomes necessary. Summary measures of population health combine cause of death, descriptive epidemiological data on incidence, prevalence, duration etc. and health state values. Cause of death data is invariably tabulated according to disease labels. Descriptive epidemiological information is largely available for health states identified by specific disease labels. The system of labelling causes of death is usually similar to the system

of nomenclature of morbidities. Where there is some variation a mapping of disease state labels to the cause of death label is usually feasible. Hence it becomes imperative for most researchers to use the disease categories as a convenient classification mechanism for disaggregated analysis of summary measures. Disaggregation by risk factor is usually achieved by tracing the incidence of mortality and morbidity to the risk factor through different disease categories. Thus, to incorporate health state values into a summary measure of population health status that allows disaggregated analysis, we need to arrive at health status or disability weights for disease categories included in the computation. If health states were valued separately for each of the disease categories, similar to incidence prevalence measurements, then the computations will be straightforward. Although valuation of health status of persons suffering specific disease conditions is feasible, such measurements are not used for summary measures of population health status, for various reasons. The valuation of health states is known to be conditioned by the locus of the valuer. The valuation of the same health state by a person in that state is usually different from the valuations given to that state by doctors and nurses. These two valuations differ from the ones given by the general population. Since summary measures are used for health policy analysis and allocation decisions, valuations by the general populations are preferred.

To cope with various methodological difficulties, direct measurement of health state values is carried out for a limited set of indicator conditions followed by statistical modelling to infer health state values for other 6D5L profiles. We need to relate the health state values thus arrived at, to disease states used for disaggregated analysis of summary measures. Hence the need for a protocol to identify the 6D5L profile corresponding to disease states.

We decided to use expert judgement arrived at by a consensus development method for identification of 6D5L profiles for identified disease states. A workshop was organised to bring together a panel of physicians and nurses from various fields working in public and private hospitals. Altogether a group of nineteen physicians and four clinical nurses participated. All panel members had clinical positions in local hospitals. Details of the workshop proceedings have been reported elsewhere (Mahapatra et al, 2000). The panel recommended 6D5L profiles to each of the 22 diseases.

While planning the study, we had provisionally selected a list of indicator conditions along with their 6D5L profiles. We set aside the 6D5L profile of indicator conditions till recommendations of the description panel was available. We then compared the provisionally identified 6D5L profiles with the description panel recommendations. In four out of twenty two cases the two matched. These were: Watery diarrhoea 111211, Infertility 111131, Mild hearing disorder 112121, Paraplegia 444431. There was discrepancy for other conditions. We discussed these discrepancies among ourselves and sought additional expert opinion wherever necessary. Finally, we accepted panel recommendations for five states, adopted a modified profile partially accepting panel recommendations for six

cases and maintained our provisional profile for seven conditions. Appendix 4-2 shows provisional panel recommendations and final identification of 6D5L profiles for chosen disease states. Whereever there was a difference between the provisional, and panel recommendations, we have validated our rationale for choosing the final profile as it stands now.

Labels:

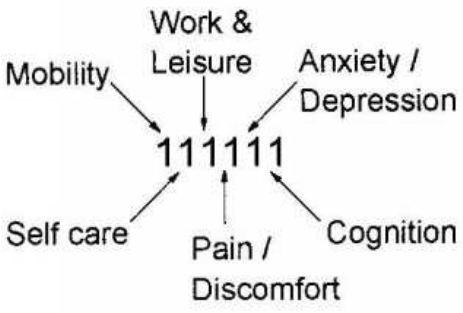
Ceteris paribus, disease labels have been found to affect the value attached to a health state by the valuer. For example, Sacket and Torrance (1978) found that labels had statistically significant effects upon health state utilities in both the positive (tuberculosis preferred over an unnamed contagious disease) and strongly negative directions (mastectomy for injury preferred over mastectomy for breast cancer). Pilot testing of the valuation exercises using the descriptions arrived so far, indicated that valuer's responses were clearly based on their stereotyped understanding of the disease labels, without paying much attention to the 6D5L description. For example, people appeared to value tuberculosis as worse than what its 6D5L profile would justify. We could not ascertain whether the worse valuation was an authentic response or an effect, purely, of the label. In any case, to minimise effect of the label as much as we could, we decided to have longer descriptive labels emphasising the 6D5L profile. Table4.3 shows the evolution of labels for selected health states. The first column shows the disease labels that we began with. The second columns shows the label used for the MDHSV workshops. The last (fourth) column shows the longer labels, used for the household survey. Details for all health states and the Telugu version of the labels is available elsewhere (Mahapatra et al, 2000).

Table 6.3: Short and long disease labels used in health state valuation exercises.

Disease labels	Labels used in the MDHSV workshops	6D5L Profile	Long labels used in the household survey
Diabetes	Mild diabetes, no symptoms	111121	Mild diabetes with no symptoms, controlled with pills
Tuberculosis	Mild tuberculosis with treatment	111221	Tuberculosis under treatment with very mild symptoms limited to occasional cough
Unipolar major depression	Unipolar major depression	124142	Depression, with loss of pleasure from most activities, low energy and slight difficulties in thinking and concentrating
Congestive heart failure	Severe heart failure (congestive)	434531	Extreme chest pains and breathlessness caused by severe heart failure

Coding schema:

Figure 6.3: 6D5L code and dimensions



A health state is described by a string of six ordered digits, such that position of the digit represents a particular dimension and value of the digit ranging from 1 -5 represents the severity level. For example health sate 111111 would mean perfect health. Positions in the ordered sequence of six digits first to sixth are respectively, mobility, self-care, usual activities (work and leisure), pain / discomfort, anxiety / depression, and cognition (Figure- 6.3).

Multi-method deliberative health state valuation (MDHSV) workshops:

The concept of values given to different health states is essentially psychometric. We need to understand the reliability and validity of measurement instruments before we can interpret the valuations obtained through them. Our aim is to measure health state valuations in the general population. Health state valuation instruments for widespread use in the general population has to be simple and easily understood by most people. Experience has revealed that anything more complicated than rank ordering and the visual analogue scale is usually not feasible. Assessing the validity of psychometric instruments, particularly for health state valuation, is a difficult task. Unfortunately we do not have a gold standard to test criterion-based validity of health state valuation instrument. Instead, we have to rely on the convergence of measurements from multiple instruments and the logical consistency of measurements to draw inferences about instrument validity.

Measuring valuation of different health states by individuals is a complex task. The measurement process requires valuers to imagine the quality of life implied by the health state descriptions, and then deliberate on some thought experiments comparing the health states to anchor points like perfect health and death. The scaling method itself may require additional deliberations and thinking, so that the valuer is able to express his / her valuation reasonably well. These tasks require the valuer to understand the measurement set up

and devote time to execute the complex cognitive tasks of valuing each health state. Asking for valuations using more than one method will require still more time.

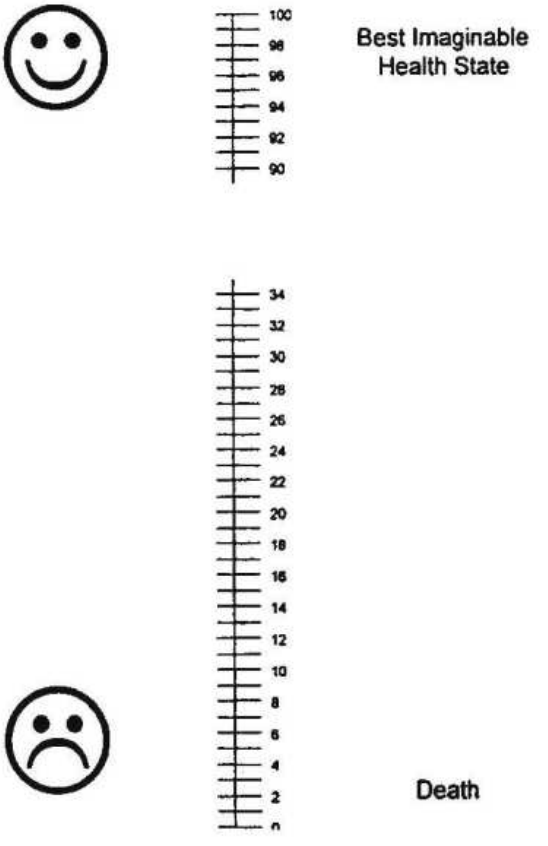
Health state valuations in a workshop setting provides an opportunity for full attention of participants on the valuation tasks. Workshops which collect a group of valuers at the same place and time allows researchers to present and explain details of each valuation instrument. Some of the alternative health state valuation instruments like the time tradeoff and person tradeoff require the valuer to imagine decision situations that the valuer might not face regularly. The workshop setting provides a cost-effective format to clarify doubts of valuers. The demerit of the workshop format is that it is not easily amenable to population-based studies. So it was decided that properties and behaviour of measurement instruments (including the one to be used for population- based surveys), would be studied using the workshop format. We planned to test three health state valuation instruments, namely (a) ordinal rank consistent visual analogue scaling, (b) time trade off, and (c) person tradeoff. This sub-section describes the detailed steps involved in recruitment of valuers, and the organisation of the workshops.

Altogether thirteen health state valuation workshops took place with a total of 180 participants. On an average, each workshop had fifteen to seventeen participants. Valuer's were chosen from different backgrounds including college teachers, community medicine faculty and postgraduate students, high school teachers, hospital administration students, primary health centre medical officers and health system researchers.

The workshops usually started with an introduction about objectives of the study and role of the participants. A written guideline was provided. At the very beginning, objectives of the study were reviewed and role of the participants as valuers of health states from the societal perspective was explained. Each participant was exhorted to perform a quasi-judicial function in assigning weights to different health states. First each participant described his / her own health state using the 6D5L description system. Then they moved on to rank ordering of the 11 health states by card sort, followed by visual analogue scaling of each health state, time tradeoff and if time permitted, person tradeoff. Before beginning each type of scaling exercise, valuers were given detailed explanations on how to use the respective instruments. At the end of each session the participants were given an opportunity to check and compare their valuations with card-sort rankings and make necessary changes. Several iterations were carried on until the valuations matched. Participants were allowed to discontinue the valuation exercise if they felt fatigued or showed signs of frustration. A vote of thanks as well as an honorarium was given to all the participants of the workshop.

Each participant followed the same sequence of valuation activity. The first task for each of the valuers was to describe his / her own health state using the 6D5L description system. Each participant was given a worksheet containing the written 6D5L description system, along with check boxes against the severity levels under each dimension. The valuers first checked appropriate severity level, applicable to him / her, under each of the six dimensions. Each valuer was supplied with two cards showing "Your own health today" title and the six dimension labels preceded by blanks to fill in the severity levels. Valuers filled in the blanks based on the severity levels they had already filled in the written 6D5L check list. The "Your own health today" cards were then added, respectively, to both sets of cards containing description of the health states assigned to the valuer. Next, the valuer worked with the free pack of health state cards including the "Own health state" card just prepared and was asked to order the cards from best health state in the pack to worst health state in the pack. The valuer then recorded rank order of the card on the card sort log provided in the workshop packet. The data derived from this yielded rank order within the respective set of health states.

Figure - 6.4 A scaled-down picture of the visual analogue scale (Actual size = Legal)



After completion of ordinal ranking of given health states, the participant moved on to scaling by visual analogue. The VAS platform comprised the picture of a visual analogue scale (VAS) labelled at either ends to represent the two extremes of the health continuum, namely best imaginable health state and death— a vertical straight line with divisions ranging from 0 to 100, and every even division labelled. Where 0 represented death and 100 represented perfect health. A picture of a happy face near 100 on one side and a picture of a sad face near 0 lent further focus to the valuer (Figure 6.4). The valuers were instructed to plant the pin-mounted health state cards along the scale according to the magnitude of their severity. The assigned host collected the card sort log and the VAS platform. The host first transferred the scale values to a VAS measurement log. A workbook was created for the valuer using the data entry deliberative iterative tool (DEDIT[8]) template. Data from the card sort log and VAS score log were then transferred to the spreadsheet. The Card sort - VAS report from the DEDIT tool titled "Reflections: Reconcile Card Sort and Scale Based Valuations" was then printed. The report indicated whether the card sort and VAS ranks matched, or, if they did not match, the nature of discrepancy. The report was printed and furnished to the valuer. In case of persisting discrepancy, the valuer was requested to review his / her VAS locations of various health states. The revised valuations were then entered into the HSV-DEDIT program and the process repeated. These iterations are continued till the two valuations match.

Time Tradeoff (TTO):

For the time tradeoff exercise a worksheet was provided with about ten rows of suggested choices to convey the idea of the time tradeoff. For example, if a person was faced with a less than perfect health state with a life expectancy of say 12 years, the time tradeoff alternative could start from a period slightly less than 12 years and progress downwards or could start from a very low value of say 0.6 years and progress upwards. Beginning with the alternatives from the upper end and progressing downwards or vice versa for all valuer's might have biased the valuation. To avoid this bias, two sets of worksheets had been prepared for each set of health states, one showing alternative-2 progressively decreasing and the other with the same alternative progressively increasing from the other end. These subsets were assigned to valuers alternatively. A sample of the TTO worksheets is given in Appendix 6-3. Detailed steps for preparation of the worksheets has been reported elsewhere (Mahapatra and others, 2000).

[8]This is a spreadsheet program in Lotus 123. For more details see Mahapatra et al, 2000.

Figure - 6.5: A report generated by the HSV Workshop Data Entry Program

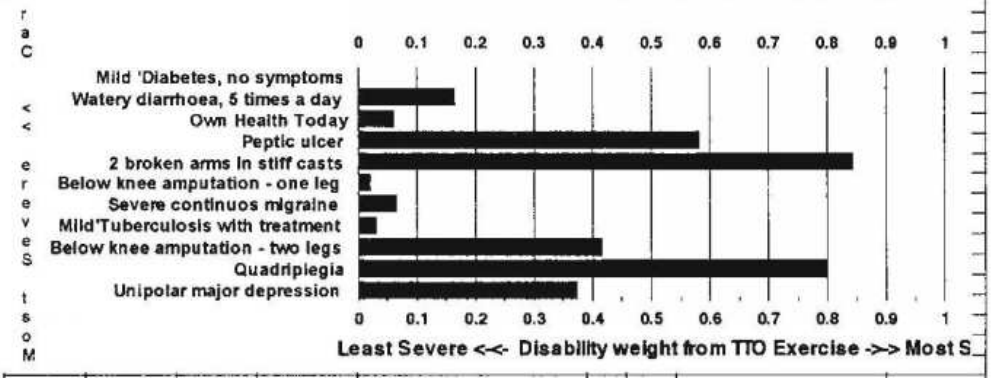
Time tradeoff - reflections: Review the magnitude of disability weights

Date: 13/10/99

Participant ID: 75_3 Name: Kondaveeti PADMAJA MOHAN Attempt: 2

Thank you for valuing the given health states using the Time Trade Off exercise. We have computed the disability weights implicitly assigned by you to each condition. These are shown on the right side of this note. The level of disability implied by your choice in the Time TradeOff exercise is also shown in the graph below. You may recall the rank ordering of conditions done by you in the card sort exercise earlier. We have arranged the conditions according to your card sort rank, so that you can reflect if your current valuations are consistent with the ordering of severity judged by you earlier.

Sl	Health State Condition	DWt
1	Mild 'Diabetes, no symptoms	0.00
2	Watery diarrhoea, 5 times a day	0.17
3	Own Health Today	0.06
4	Peptic ulcer	0.58
5	2 broken arms in stiff casts	0.84
6	Below knee amputation - one leg	0.02
7	Severe continuos migraine	0.07
8	Mild'Tuberculosis with treatment	0.03
9	Below knee amputation - two legs	0.42
10	Quadriplegia	0.80
11	Unipolar major depression	0.38



Please reflect upon the degree of severity you have determined for each condition above, as well as any discrempancies between the card sort and Time Tradeoff exercises. Based on this reflection you may wish to revise your Time Tradeoff evaluations. We would recommend that you do not revise the card sort at this stage, since you have already reflected on it adequately. Now please reflect and revise your valuations and let us have your revised estimations.

Before moving on to the TTO exercise, the workshop co-ordinator explains the method using a health state not included in the indicator conditions. We used paraplegia as an example in most cases. The valuers are guided through a sample worksheet with the condition stated in the example. This is done to familiarise participants with the TTO methodology. Most valuers, would usually have questions about the methodology. We found it helpful to explain the valuers that if a health state was considered very severe then a person would usually be willing to trade in that health state for a relatively much shorter duration of life in perfect health. This helped valuers to grasp the nature of the TTO exercise. Valuers were encouraged to reflect on their initial TTO valuations, if the rank ordering of conditions implied by it did not match with the card sort ranks. Initial valuation data was fed into the DEDIT spreadsheet and a "Reflection reports" (Figure 4.5), pointing out discrepancies along with a graphical comparison of the TTO valuations with card sort was printed and given back to the concerned valuer. S/he was asked to revise the TTO valuations in the light of discrepancies pointed out in the Reflections report. Subsequent valuations were entered into the DEDIT spreadsheet and a fresh Reflection report generated.

In the initial phase of the study, we continued this iterative process till the valuer reconciled the card sort ranks and TTO ranks completely. However, we did not succeed in our efforts completely. Despite our insistence, it was found that some people were not able to reconcile the two valuations even after thirteen additional iterations. We found that after the first few iterations, most participants stopped deliberating and thinking about the issue. Instead, they tried to figure out a way, by hit and trial, to some how match the TTO valuations with card sort rank. Thus participants appeared to reflect and deliberate for the first few iterations and then simply gave up. We then decided, not to insist on complete matching of card sort ranks with TTO valuations. From the fourth workshop onwards, we asked participants to continue for as many iterations as they felt comfortable with and then stop. By this time we had also improved the DEDiT Reflection reports to include a bar chart of TTO valuations arranged according to the card sort rank order (Figure-6.5). This visual tool appeared to help communicate the discrepancy more effectively. We found that given freedom to stop at will, valuers did not purse the TTO exercise beyond about four iterations. The number of valuers with complete match of TTO valuation with card sort ranks reduced, after participants were given the option to stop at will. For future studies, it would appear desirable to plan on a maximum of six iterations including the first round and improve rate of matched valuations by further improving the feedback communication strategy and workshop co-ordination skills.

PTO Exercises:

The person tradeoff method as described in Murray and Lopez (1996) was used in this study. This requires the valuer to consider the alternatives from two perspectives namely PTO1 and PTO2. Contents of intervention-A helping the reference unit of healthy persons (x) remains the same for both perspectives. PTO1 and PTO2 differ only in the manner in which the intervention - B is framed. In PTO1 Intervention - B will extend the life of a larger number of individuals (y) but with a less than perfect health state for one year. In other words this intervention will not be able to cure the disability. But it will enable the persons living with the disease to live longer by one more year. Here the health state value for condition i is given by $h_i = \frac{x}{y}$. In PTO2 the Intervention - B will cure a larger number of individuals (y) of the adverse health situation. In other words these persons will be able to enjoy perfect health for one year. Here the health state for condition i is given by $h_i = \frac{y-x}{y}$. As can be seen above, PTO1 presents the perspective of extending the life of a group of persons with some disability. Thus the valuer has to factor in his / her mind the net gain (h_i) in healthy years to the community and compare the same to intervention-A. In PTO2, the valuer is buying an intervention that cures the disability. So the valuer has to mentally factor in his / her estimate of disability with the number of persons

being benefited (life year gained = (1-h_i) * No of persons benefited) to compare with the persons benefited by intervention-A. The two perspectives are presented to facilitate deliberation by the valuers.

The steps for the PTO exercise were mostly similar to TTO, except for the fact that the TTO worksheets were replaced by PTO worksheets. The PTO exercise has two sub exercises within it, namely PTO1 and PTO2. So separate worksheets were provided for PTO1 and PTO2. A sample of PTO1 and PTO2 worksheets for two broken arms in stiff cast with progressively increasing alternative - 2 is shown in Appendix 6.4. Before moving on to the PTO exercise, the workshop co-ordinator explains the method. Most valuers did not get to the PTO exercise test. Only 28 valuers attempted PTO. Of these 18 were females and 10 males, 22 were between 20 to 29 years age and the remaining six were between 30 to 36 years age. Out of the 28 persons who attempted the PTO exercise, the PTO rank orders matched with card sort rank order only for 7 persons. A retest workshop was conducted with 15 participants. The gap between the original workshop attended by them and the repeat workshop was about six months. Only VAS and TTO were repeated. PTO was dropped from the repeat tests, since the number of original PTO valuations satisfying ordinal rank consistency were small (seven) and finding valuers from this small group for the repeat test was difficult.

Community-based health state valuations in Kondakkal village:

To select a village for the study we approached another civil society institution[9], which is working to improve educational status of people. We requested them to help us identify a typical village, with more than 1000 households. Kondakkal was chosen on the basis of their suggestions and an exploratory visit by faculty from the IHS.

Kondakkal has a population of 2342 adults and is typical of villages in Telengana area of Andhra Pradesh. The electoral (voters) list for females and males containing 1127 and 1215 entries respectively was used as the sampling frame. A simple random sample of 550 persons from each of the two strata was drawn. Altogether, a list of 1100 potential valuers was prepared, consisting equal number of females and males. In the community survey 1010 out of the 1100 randomly sampled persons were actually interviewed (Non response proportion = 0.08). Age-sex distribution is on expected lines and appears representative of the adult population of the village.

The survey was conducted over a period of 12 days between October 15 - 26, 1999. As most of the valuers were engaged in agricultural labour, we were advised to conduct the surveys during early mornings as well as in late evenings. Typically, the surveyors had to start as early as 6:00 am in the morning

[9]The MV Foundation which operates in many areas of the Ranga Reddy district in Andhra Pradesh

for the early morning sessions. Each surveyor took around 1 to 1 1/2 hours to complete one household survey. Each surveyor was provided with a local escort recruited from the village. The local escort helped in locating households and introducing the surveyor to the interviewee. The surveyor usually spent some time in cultivating an acquaintance with the valuer and his / her household. After that, s(he) would fill out the personal information form and then proceed to the valuation session. Typically, a valuation session with a villager proceeded as follows.

The session would start with description of "own health state" by the valuer. Surveyors used a set of Velcro-mounted 6D5L graphical description cards appropriate for the interviewee's gender. The surveyor explained each of the six dimensions, one after the other, using the Velcro- mounted cards. Within each dimension, the five levels of severity were explained. The valuer was then asked to pick up a picture from a set of five pictures in the mobility dimension, that described his / her own mobility status. The valuer then stuck this picture to a blank "Your own health today" platform. The process was repeated for all six dimensions. Thus the valuer built a "Your own health today" card by picking up appropriate pictures to represent severity levels in each of the six dimensions. Once done, the valuer was asked to take a fresh look at the "Your own health state today" card just built by him /her and review the same if necessary. The labels on each picture were read out to the valuer. Once the valuer made up his / her mind and froze the "Your own health today" card, a pin-mounted version of the same card was made by the surveyor. These two cards were added to the free and pin-mounted set of cards to make up the complete set of 11 cards. Now the valuer is ready to do the card sorting. The surveyor would present cards in the assigned set, explaining the levels in each of the health states one by one. The valuer was requested to select the health condition, including his / her own health state, that was the worst of all the conditions. After selecting one s(he) was then asked to choose the next worst from the remaining 10. This process continued until there were no cards left. If the valuer had difficulty in identifying the next to most worst health state, then (s)he was asked to find the best health state from the remaining cards. Thus in some case the sorting proceeded step by step from worst to best, or in some other cases from best to worst, and in yet other cases in a cyclical fashion: worst, best, worst. Yet another way of getting valuers to order the cards was by pair-wise comparison. Surveyors had to resort to one of these methods, depending on the valuer's comfort level. After ordering of all 11 health state cards, the surveyor read out the health states one by one in a sequential order from worst to the best, and sought confirmation by the valuer. Although the sorting exercise proceeded in different manners according to the valuers choice, the rank orders assigned by the interviewee was recorded from best to worst. Thus the best health state received

rank order 1 and the worst state got rank 11. Finally the valuers would take up visual analogue scaling. The pin-mounted set of cards were sorted according to the card sort rank. A cork board visual scale was presented and the valuer then asked, one by one, to pin the cards showing his / her opinions for these health states as being near or far from "best imaginable health state" and "death" as well as from each other. Surveyors were instructed to look for any qualitative remark or information given by the valuer and record them. After this the surveyor thanked the valuer and gave a token gift. A valuation session is now considered to be complete.

Socio-economic profile of valuers and overview of data collected from two arms of the study:

All participants in the MDHSV workshops had graduate or equivalent (15 years) level of schooling. This is in accordance with the study design to recruit educated persons for these workshops. Table 6.4 shows that both genders were evenly represented. Since participants were sought from work places, all of them belong to 20-59 years age. Relatively younger adults are slightly over-represented.

Table - 6.4: Age and Literacy of health state valuers.

	Workshops			Survey		
Characteristic	Females	Males	All	Females	Males	All
Number of valuers	88	92	180	491	519	1,010
Age Group						
15-19	0	0	0	2.51	2.98	2.75
20-29	67.05	61.96	64.44	31.11	41.67	36.52
30-44	23.86	23.91	23.89	35.70	33.13	34.38
45-59	9.09	13.04	11.11	19.83	15.28	17.50
60-69	0	1.09	0.55	9.19	5.56	7.32
70	0	0	0	4.18	4.37	4.27
Years of schooling						
0	0	0	0	84.93	56.07	70.1
1-5	0	0	0	4.28	12.91	8.71
6-9	0	0	0	5.7	11.18	8.51
10-12	0	0	0	4.07	17.53	10.99
13-15	45.45	16.3	30.56	1.02	1.73	1.39
16-18	43.18	65.22	54.44	0	0.58	0.3
19+	11.36	18.48	15	0	0	0

Table 6.5: Number of valuations obtained for each health state

Health state	GBD*	Sets	6D5L	Survey	Workshop	Both
Own Health Today		C		1010	180	1190
Watery Diarrhoea 5 times a day	Yes	C	111211	1006	180	1186
Mild diabetes, no symptoms		C	111121	1004	180	1184
Mild Tuberculosis with treatment		C	111221	1008	180	1188
Severe continuous migraine	Yes	C	113431	998	180	1178
Unipolar major depression	Yes	C	124142	1000	180	1180
Quadriplegia	Yes	C	554341	1008	180	1188
Bronchitis		1	112311	279	45	324
Pain and stiffness in joints		1	222311	278	45	323
Urinary incontinence		1	113331	279	45	324
Schizophrenia		1	234244	280	45	325
Infertility	Yes	2,5	111131	250	45	295
Angina	Yes	2,5	111321	253	45	298
Blindness	Yes	2,5	323122	247	45	292
Severe Hallucinatory Fever		2	444333	253	31	284
Peptic Ulcer		3,6	112321	258	45	303
Amputation below the knee (one leg)	Yes	3,6	322211	255	45	300
Amputation below the knee(both legs)		3,6	433221	258	45	303
Two broken arms in cast		3	154321	256	30	286
White marks on face	Yes	4	111131	235	45	280
Mild hearing disorder		4	112121	233	45	278
Continuous moderate back pain		4	212321	233	45	278
Severe heart failure (congestive)		4	434531	229	45	274
Common cold*		2,5	112211	0	14	14
Moderate Anaemia*	Yes	3,6	112211	0	15	15
	All			11110	1980	13090

* These two were included in the respective sets in the beginning. Later, they were replaced by hallucinatory fever and two broken arms in cast, respectively. Hence the number of valuations for these two is relatively small and confined to the workshop only. Similarly the number of valuations in workshops is less for the replacement conditions.

Yes if the condition was included among 22 indicator conditions in GBD study (Murray & Lopez, 1996)

As Table 6.4 shows, about 70% of the valuers were illiterate. The Constitution of India recognises certain castes as socially disadvantaged. These castes included in the appropriate schedule of the Constitution are referred to as Scheduled castes. The Constitution provides for another schedule of tribes and aboriginal people living mostly in remote areas and these are called scheduled tribes. Backward classes are groups recognised by the State

Government as economically backward. All others not covered in any of the above three groups are classified under the residual category of "Other castes".

The list of health states, their assignment to different sets, and the number of valuations obtained for each from the survey and workshops respectively is shown in Table-6.5. Valuation through the community survey was done for twenty two of these health states. Since two of the 22 originally planned indicator conditions were substituted mid course but before the community survey started, the table lists a total of 24 health states. In addition, each of the respondents valued their own health state. Ten out of the 24 indicator conditions overlap with the indicator conditions in the GBD96 study. The 24 conditions consist of 22 unique 6D5L profiles. Only two profiles were represented more than once. These were 111131 (white marks on face, watery diarrhoea) and 112211 (moderate anaemia, common cold). If we restrict to the community survey alone, the 22 indicator conditions consist of 21 unique 6D5L profiles. Profile of indicator conditions represents a fairly broad range of severity levels in each of the six dimensions. All five severity levels of the mobility, self-care , and pain dimensions are represented. Four levels of usual activities, anxiety and cognition are represented. Since the six core health states were included in all the four sets, the number of valuations for these is considerably higher.

Reliability of health state valuation tools:

Concept of reliability of a measurement tool and its measurement:

The reliability of an instrument refers to the reproducibility of its measurements when applied to the same object. Reliability is to be distinguished from the concept of validity. An instrument may measure reliably but may not be valid. Reliability is a necessary but not sufficient condition for validity. In the physical world let us take the case of a one litre liquid measure that has a dent in it, reducing its volume by, say, 10 ml. Such a measure when applied to say 10 litres of edible oil will reveal the result as 10.1 litre. If the same quantity of oil is measured by the same liquid measure repeatedly, the result will consistently be distributed around 10.1 litre, except for random errors, and assuming that we have a set up that allows no spillage. This is a reliable but not valid measure of volume. To illustrate the concept of validity using the same example, now suppose we have a reliable measure with its volume exactly equal to one litre. This is both a reliable and valid measure of volume. But this is not a valid measure of weight[10]. One implicit assumption in the example given above is that the object of measurement, namely, the quantity of oil being measured,

[10]The example of oil is chosen deliberately to illustrate the concept of validity. One litre of oil does not equal one kilogram of oil. Volume of a liquid is related to its weight as a function of the temperature at which the measurement is taken and the specific gravity of the liquid.

remained the same (hence the assumption of no spillage). This is referred to as measurement stability. In the physical world, the measurement stability is not usually a problem, although even this can surface. When we apply the concept of reliability to psychometric measurements, the assumption of measurement stability may have to be tested.

In psychometrics, the concept of test (i.e. measurement instrument) reliability and its measurement is visualised using two theoretical models, namely: (a) the classical test theory, and (b) generalizability theory (G theory). More detailed introductions to the concept of reliability, validity, classical test theory and generalizability theory can be found in Carmines and Zeller (1979) and Shavelson and Webb (1991). Streiner and Norman (1995) describe these concepts in the context of health measurement scales[11]. Deyo et al (1991) describe the concept of reliability in the context of health status measurement and supplement it with some interesting computational formulae for computation of reliability coefficients.

Using classical test theory in the context of health state valuation, we would assume that the valuer (i.e. the respondent whose valuation we are measuring) has a true valuation for each of the health state being evaluated by him / her. We cannot observe this true valuation. The health state value assigned by a valuer (this is what we observe) contains within it both the true valuation and a random error. More formally; $h_i = H_i + e_i$ where h_i is the value assigned by valuer i to a health state. H_i is the true valuation in the mind of the valuer i for the health state, and e_i is the random error in measurement that is unrelated to the health state and the person. We further assume that the random error is distributed as normal with mean error of zero and an error variance $N(0,\sigma^2_e)$. These reasonable assumptions would imply that if a health state is valued by many individuals, the variance of the observed health state values will consist of the true variance of the valuations given to the health state by different individuals and the error variance. We can arrive at estimates that separate the error variance from the true variance if we have parallel measurements (for example test-retest). In psychometrics, reliability is assessed by various types of parallel measurements including (a) test-retest, (b) alternative forms of the same instrument, (c) split halves method and (d) internal consistency of test items. Except the test-retest method, all other methods are appropriate only to instruments consisting of multiple items. The health state valuation instruments like the VAS, TTO and PTO are single items scales. Hence the test-retest appears to be the only method appropriate for our purposes.

[11] Note that the equation in page 110 illustrating computation of Sum of squares (error) is a printing error. The authors have confirmed this. However their result is correct. A correct version of this equation would be a follows: Sum of squares (errors) = (6-7-5+6)^2 + (4-5-5+)^2 + ...+ (7-7-6+6)^2 + (5-5-6+6)^2 + ... + (8-7-8+6)^2 = 10

If we plot test and retest measurements from a perfectly reliably instrument, our theoretical conception of reliability implies that the result will be a straight line from origin with a slope (concordance). This will also imply a perfect correlation between test and retest measurements. Hence traditionally, correlation coefficients (Pearson's and/or Spearman) between test and retest valuations have been used to describe the reliability of tests (Torrance, 1976, Bergner and others 1981, Bass and others 1993). The advantage of correlation coefficient is that it is quite well recognised as a measure of association. The problem in the context of reliability measurement is that the correlation coefficient would continue to be high even if the retest results are systematically different. If the retest valuations are systematically different then the difference in test-retest group means would be systematically different also. This can be tested by a *t* test for paired differences or difference in means. A statistically insignificant *t* statistic coupled with a high correlation coefficient would serve as evidence in support of test reliability. For example, see its usage by Torrance (1976), and Bass et al (1993) cited above. A near-perfect correlation coefficient coupled with statistically insignificant difference in test and retest mean does not, however, assure complete concordance of test and retest valuation. A linear combination retest valuations allowing for a non zero intercept can give rise to perfect correlation with not very dissimilar means. We could probably deal with this situation by testing statistical significance of the intercept from a theoretical value of zero. An intra-class correlation coefficient (ICC) is a single statistic that combines correlation between test retest and concordance of the two means. ICC measures not only the strength of correlation, but also the deviation of the slope and intercept from that expected for replicate measures (Deyo et al, 1991). An ICC under the classical test theory is defined as: $ICC = \frac{\sigma^2_{Persons}}{\sigma^2_{Persons} + \sigma^2_{Occasion} + \sigma^2_{Error}}$ where σ^2 is the total variance and other variances correspond to their respective subscripts, the occasion meaning test or retest. An ICC can be computed for interval level measurements. Its counterpart for ordinal rank ordered measurements is the weighted kappa (Streiner and Norman, 1995; Kramer and Feinstein, 1998).

According to classical test theory, the variance of measurements by an instrument is visualised to consist of two parts, namely the systematic difference in valuations by the persons using the instrument to indicate their valuations (this is the primary object of our measurement) and error. All other variance components except the one attributable to the object of measurement are wrapped into the error variance. Any source of systematic variation other than the variance within subject is considered to affect the validity of the measurements but not the reliability. The generalizability theory relaxes this restriction by allowing for teasing out of some more components from the error

variance. These are variance components that can clearly be attributed to aspects of the measurement process. This implies that systematic effects on the valuations is further decomposed into aspects of measurement and residual systematic effects if any. The residual systematic effect, if any, affects validity of the measurements. For purposes of generalizability studies, residual effects, if any, are clubbed with the error term. Generalizability theory (Cronbach and others 1972) assumes that there are multiple sources of error in the measurement process. Each of the recognisable sources of error is considered a facet of measurement. These facets may be fixed (fixed facets) or we may want to generalise results to the universe of the concerned facet (facets of generalisation). The object of measurement is considered the facet of differentiation. Using analysis of variance computations, the variance components for each of the facets and their interactions are computed. The generalizability coefficient calculated from the variance components is a measure of reliability. The intra class correlation coefficient (ICC) described above happens to be a special case of the generalizability coefficient, in a one-facet model.

Table-6.6: Feasibility of test - retest reliability measures for health state valuation instruments

	Interval		Ranks	
	Per health state	All health states	Per health state	All health states
Individual	Irrelevant	Person's correlation	Irrelevant	Spearman's *p* Kendal's τ Weighted Kappa
Group	ANOVA: Classical test theory (ICC)	ANOVA:. Generalizability Theory (Generalizability coefficient)	ANOVA: Friedman's by ranks	

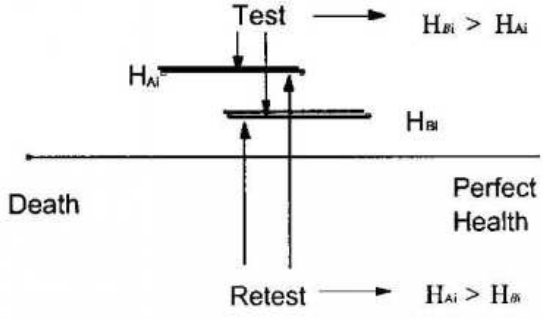
[1]Source: Adapted from van Agt HM; Essink-Bot ML, and Krabbe PFM (1994)

The feasibility of various test - retest reliability measures for health state valuation instruments can be summarised using a schema from Van Agt et al (1994) with some further modifications in the light of above discussions (Table 6.6). Although, we have retained Van Agt et al's columns for rank ordered data, test-retest reliability of rank orders assigned to the same set of health states by the same individual is not at issue. We treat the rank ordering of health states as the primitive expression of individual's true preference of ordering at the time of exercise. Hence we assume rank orders to be reliable and instead use the reliability measures to estimate concordance of rank ordering from test to retest. The test-retest concordance of rank orderings will help us assess the extent of measurement stability which is fundamental to interpretation of reliability measures.

Measurement stability in health state valuation:

The reliability of the health state valuation instrument is predicated on stability of expressed valuations. Consistency in expression of value for a health state is dependent on the nature of the true valuation of the health state in the valuer's mind. Let us take another look at the classical measurement model described earlier. It is conventionally assumed that H_i (the true quantity in the valuers mind) is a single value. Differences in the observed value h_i are wholly due to e_i, i.e. the error component. Such an interpretation assumes that every person has a well-formed and crystallised value attached to every health state, irrespective of the incidence of occasions and events encountered by him / her that would be cause for deliberation about this matter. Such an interpretation, however, does not appear to be the case in reality. People do not directly confront such questions in their daily life, although they do handle situations that implies choices between different health alternatives. It would appear more plausible that the true valuation in most person's mind is a fuzzy set consisting of a range of values for each health state. Thus we view H_i to be a multivalued set, and each attempt by the valuer *i* to express a value would actually be a sample endogenously drawn by the valuer from this set. For some health states, this range may be narrow, as is the case for extreme disabilities such as quadriplegia or unequivocally trivial illnesses such as common cold. For some other health states, the true valuation set may consist of a wider range. The set may narrow down as the individual deliberates on the characteristics of the health state, its relationship to other health states, and its implications for a person. If the true valuation in the minds of persons is a fuzzy set, and each valuation attempt is a sampling from that state, then there would be scope for instability in expressed valuations.

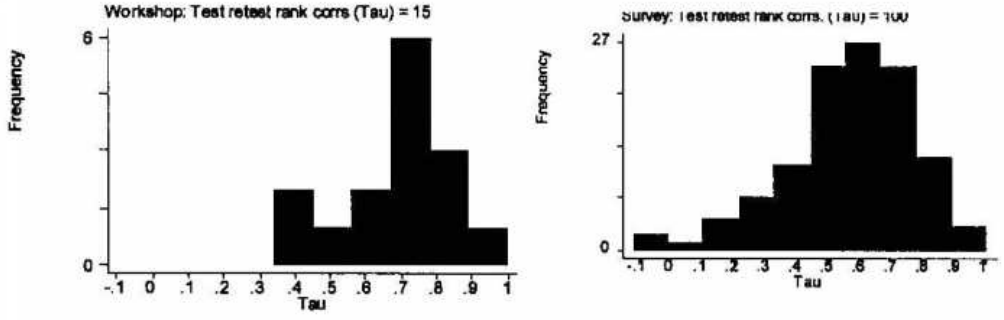
Figure 6.6 : Potential for rank reversal of health states in retesting.



For example, consider health states *A*, and B with overlapping true value sets (Figure 6.6) . Suppose a person with these true value sets is asked to express his / her valuation on an occasion (test). (S)he may report health as being state *B as being* better than health state *A* by tapping into his / her true value set in the manner shown in Figure-6.6. On another occasion, (s)he may sample his /

her true value set in the manner shown for the retest, in Figure 6.6, in which case (s)he will report state *A* to be better than state *B*. Since the health state valuation instrument measures the expressed valuation the stability of the expressed valuation of health states is confounded with the reliability of the measurement instrument.

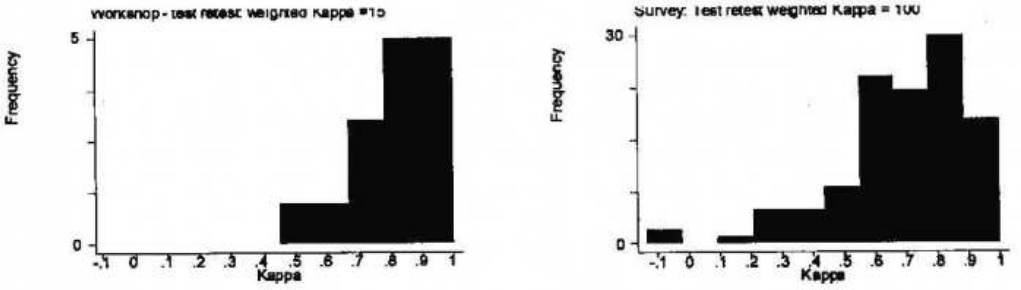
Figure - 6.7: Distribution of within valuer test retest rank correlation (Kendal's Tau)



The rank orderings provide us with a means of testing the above hypothesis as to whether people's valuation of health state is a single valued quantity or a multivalued set which a person maps each time (s)he intends to assign a numerical value to the health state. We assume that ordinal ranking suffers from minimal measurement error. Thus major changes in ordinal ranking from test to retest would support the above argument that the nature of true valuation in peoples mind is in the form of a fuzzy set of values with different degrees of clarification for different health states. We computed both Spearman's Rho and Kendal's Tau and performed tests of statistical significance on both. Each of these correlation coefficients was computed from 11 pairs of valuation by an individual. Figure 6.7 shows distribution of individual level rank correlation of test retest ranks. We choose Kendal's Tau for the graphical presentation since the magnitude of the Tau statistic is less sensitive to extreme values giving rise to a more normal spread compared to distribution of the Rho. The left graph shows distribution of rank correlation coefficients estimated for the 15 test retest valuers from the workshop. The right graph shows the same for 100 test-retest valuers in Kondakkal village. Note that none of the valuers could exactly reproduce their original rank orderings. Rank correlations from the workshop valuers are more tightly distributed around 0.7. The rank correlations from the community survey is centred at 0.6 but are more dispersed. We tested the null hypothesis of no correlation between test and retest rank orders using the Spearman's Rho and Kendal's Tau. For the workshop participants, we were able to reject the null hypothesis at 95% level of confidence, in 13 (87%) out of 15 test-retest cases. For the village population we were able to reject the null

hypothesis at 95% level of confidence, in case of 69 (69%) correlations out of the total of 100. In the balance 31 cases we fail to reject the null hypothesis of no correlation.

Figure-6.8: Distribution of test retest agreement weighted Kappa statistic



One possibility, if rank order of health states is not fully retained from test to retest, is that the valuers are randomly drawing from an undefined number space (0,1)? That would mean that persons do not have a value set and instead are simply giving a totally random response in the interval (0,1). To test this we turn to some measure of agreement, that allows us to test statistically whether the test-retest agreement is purely due to chance. Note that the reliability measures we discussed earlier are essentially measures of concordance. When we can assume that measurement stability is satisfied, we use these statistics to measure reliability. Turning these measures around to a situation where we assume that the measurements (rank orderings here) are reliable, we can use them as measures of concordance. The rank correlation measures trend in each variable, whereas we are interested in concordance. This is the motivation for the intra-class correlation coefficient as a more appropriate measure of reliability in case of interval level data. Its equivalent for rank ordered data is the weighted kappa statistic. The weighted kappa (Cohen 1968) statistic computes the actual agreement between test-retest ranking and compares it with the agreement expected at random[12]. Figure 6.8 shows frequency distribution of quadratic weighted kappa by the value of the statistic for the test retest cases from the workshops (left graph) and the community survey (right graph). Kappa coefficients from the workshops are more tightly distributed around 0.9, where as these are more spread out in case of the community survey. We tested the null hypothesis that the agreement between test and retest rank orders is not any different from what would be expected at random, given the set of health states valued by respective valuers. We rejected this hypothesis for all 15 test retest valuers at 95% confidence. In case of the village population, we rejected the hypothesis for 82 valuers at 95% confidence and failed to reject hypothesis for the remaining 18 persons.

[12]The Kappa statistic is described in Streiner and Norman (1995, pp1 16- 118)

Finally we turn to the description and rank ordering of own health states by the valuers in test and retest. Out of fifteen test-retest valuers in the workshop, only one person changed the own health state rank. The remaining fourteen persons retained the ranking of their own health state. In case of the community survey 33 out of 100 test retest valuers changed the own health state rank and the remaining 77 retained the ranking from the first occasion. It may be of some interest to see the changes in 6D5L description of own health states. If all those who changed ranks also changed the 6D5L profile of their own health state, we cannot then rule out the possibility of real changes in valuer's own health, leading to a change in its ranking on the second occasion. In case of the workshop, 8 out of the 15 test retest valuers did not change the 6D5L profile of their own health state. The person who changed the own health state rank, however, did not change the 6D5L profile. In case of test- retest valuers from the village, 81 out of 100 persons ranked their own health state as one among the 11 health states ordered by them. 77 of these persons (85%) retained Rank 1 for their own health state, of whom 52 retained the 6D5L profile and the rest changed the profile to some extent. The four persons who changed the rank of their own health state from 1 to 2 also changed the 6D5L descriptions. The 19 persons who gave their own health state Rank 2 or higher (rank 1 = best and rank 11 = worst) in the test changed the ranks mostly to lower ranks and usually with some change in 6D5L descriptions. Most of the changes in ranking of own health status were associated with some change in 6D5L profile.

The above findings are consistent with our hypothesis. Certain additional conjectures appear. We find that stability of expressed valuations are somewhat better for the educated group and in case of "Own health" state for both groups. In case of the community survey 85% of persons who ranked their own health state as one, did not change the ranking in retest. In case of the workshop 93% did not change ranking of their own health state. It would then be reasonable to conjecture that the valuation set H_{Ai} is subject to continuous revision in the light of the persons experience and knowledge about the health state A_i and about other health states. The valuation space, that starts as a fuzzy set, would get gradually clarified to various degrees, based on different factors, including experience and knowledge about the health state or related matters. We surmise that education is an indicator of a broader range of cumulative experience. Similarly everyone, irrespective of educational status, has more intimate knowledge of their own health state. Hence the stability of expressed valuations in the presence of these two factors appear to be better. These conjectures and hypotheses will need to be investigated further. For example, one implication of the above hypothesis is that the valuation set is likely to be clarified by repeated measurement, since these will provide repeated occasions for the valuer to deliberate on the concerned health status. Of course, the extent of clarification

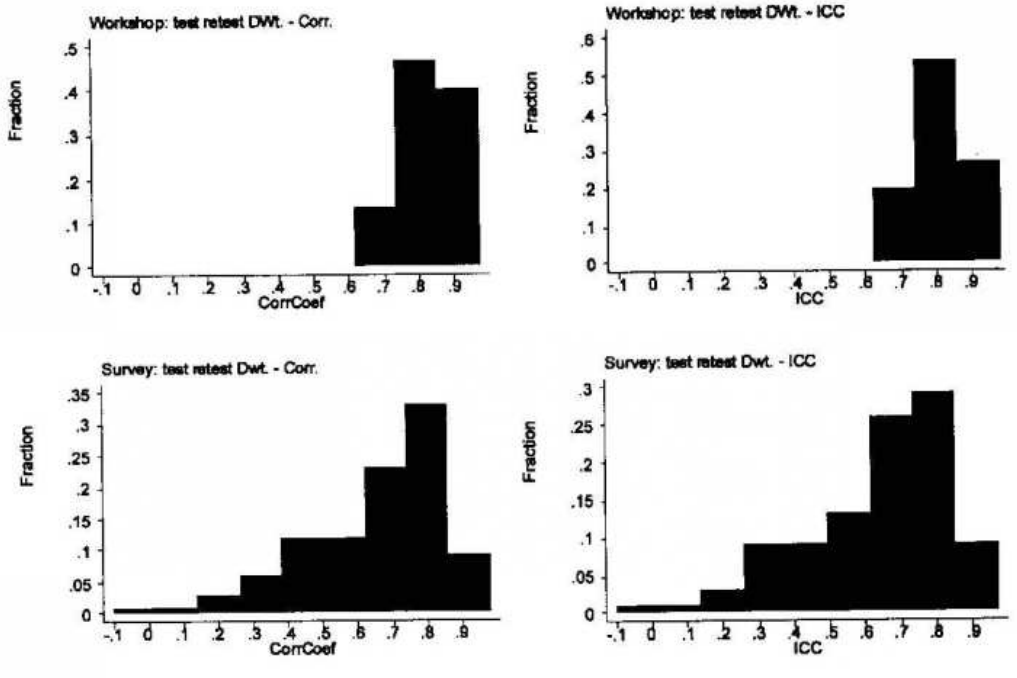
may not be the same for all health states. But some trend should be visible, if a large number of health states are measured.

Reliability of ordinal rank consistent visual analogue scales (VAS)

Conventional and Classical reliability measures:

The ordinal rank consistent visual analogue scale was used as one of the scaling methods in the multi method health state valuation workshops. This was the primary scaling method for the survey among Kondakkal village population. Retests were done with a sub sample of valuers to assess test-retest reliability in the Indian context. To assess reliability we first computed simple product moment correlation and intra class correlation coefficients separately for each individual. Figure 6.9 shows the frequency distribution of these coefficients. The top panel of two graphs show frequency distribution of correlation coefficient and ICC for the workshop participants (n=15). The bottom panel of two graphs show the same information for the 100 test retest valuers from the village population-based survey. Most of the correlation is reasonably high and distributed around 0.7 to 0.9. The ICCs are also distributed similarly. The difference in distribution of correlation from the workshop and the village population are noteworthy. The coefficients from the workshop are more tightly distributed when compared to the village population. Test valuations from some persons do not correlate at all with the retest valuations.

Figure-6.9: Distribution of within valuer test-retest product moment correlation and intra class correlation coefficients (ICC). Data from workshops and village population survey.



Intra class correlation coefficients for all health states combined and by each health state ICCs were calculated (Table-6.7) using the computational formula described by Deyo (1991). Finally, for all health states, the ICC was 0.81 for workshop participants and 0.6 for the village population. But ICCs by health states look puzzling. We would expect the ICCs by health state to be positive and high in case of health states where valuation in the community is diffused and positive but not so high for health states where the valuation in the community is tightly distributed around a central value. Measurement error will also drive the ICC value towards zero. Hence, it is difficult to separate measurement error from crystallised valuations in the community. This is an important problem with usage of ICC to measure reliability of health state valuation instruments.

Table-6.7: Intra class correlation coefficients (ICC) by health state from workshops and community survey.

Health State	Workshop		All villagers		Literate villagers	
	n	ICC	n	ICC	n	ICC
All states	165	0.81	1100	0.61	297	0.677
Mild diabetes, no symptoms	15	0.48	100	-0.01	27	0.22
Mild tuberculosis with treatment	15	0.44	100	0.24	27	0.21
Own Health	15	0.62	100	0.06	27	0.33
Quadriplegia	15	0.56	100	0.05	27	0.12
Severe migraine	15	0.49	100	0.16	27	-0.01
Unipolar major depression	15	0.60	100	0.26	27	0.19
Watery diarrhoea	15	0.26	100	0.14	27	0.09
Continuous moderate back pain	5	-0.81	21	0.20	5	0.01
Mild hearing disorder	5	0.93	21	0.00	5	0.15
Severe heart failure	5	0.81	21	0.07	5	-0.14
White marks on face	5	-0.75	21	-0.12	5	-0.18
Bronchitis	4	0.36	23	-0.15	10	0.25
Pain and stiffness of joints	4	0.63	23	0.15	10	-0.14
Schizophrenia	4	-0.17	23	0.45	10	0.67
Urinary incontinence	4	0.28	23	-0.27	10	-0.30
Below knee amputation - one leg.	2	0.76	26	0.04	5	0.12
Below knee amputation - two legs.	2	0.30	26	-0.02	5	0.35
Peptic ulcer	2	0.05	26	-0.06	5	-0.24
Two broken arms in cast	2	0.18	26	0.31	5	0.60
Angina	4	0.22	30	-0.14	7	0.02
Blindness	4	0.69	30	0.16	7	0.94
Infertility	4	-0.89	30	-0.08	7	-0.27
Severe hallucinatory fever	4	0.40	30	0.19	7	0.36

[1]n = number of valuations for the concerned health state.

We do not, however, expect negative ICC values. But the ICC for some health states are negative. We cannot dismiss these values as statistically not different from zero. For example, consider the ICCs from workshop participants for white marks on face (-0.75), infertility (-0.89), and continuous moderate back pain (-0.81). Note that all these ICCs are based on too few observations. Table-6.7 has been ordered by the number of observations on which the ICCs for each health state was computed. The top part of the table with more observations, does not show any extreme negative value of ICC. Among the top eight rows, where the sample sizes are relatively higher, mild diabetes has an ICC = 0.01 from community survey and severe migraine has an ICC = 0.01 for the literate sub group in community survey. Probably the extreme negative ICC values are just a matter of chance. In the light of our conjecture about measurement stability, we suspected that educational status may improve the ICC. So the ICCs were recomputed for the literate subset of test-retest cases from the village population. Only persons who had some formal schooling only were included in this subset. The last two columns in Table-6.7 show ICCs for this subset. We focus on the top eight rows only. No definite inference can be drawn. The ICCs improve in some cases and reduce in some others.

Generalisability study:

The generalisability study allows us to model the measurement situation as a multifaceted process where each facet has an effect on the measurement. The effect of each facet is identified. The object of measurement facet (facet of differentiation) and facets of generalisations are identified. In our case, health state is the object of measurement. Appreciation of the extent to which the instrument helps differentiate the object of measurement and effect of facets of generalisation gives us an idea about generalisability and dependability of the measurements. In the present case, we have three facets namely, (a) the valuers (v), (b) the health states (h) and (c) two occasions (o) of measurement. The valuers are a random sample of the universe of valuers. Similarly the occasions of measurement are random. We want to be able to generalise the measurements to any other occasion. The health states used for this study are a subset of a large number of health states for which the health state valuation instrument is to be applied. Since all three facets are random, we assume a fully crossed model of measurement. A fully crossed design requires that all test-retest valuers valued all health states which in our case did not happen. Each valuer worked on 11 health states including the own health state. The valuers were assigned health states from one of four sets. Thus those assigned to a set of health sets worked on the same health states on both occasions. In other words, if the

generalisability study is restricted within each set, the fully crossed design of measurement can be analysed. Within a given set we have measurements by all valuers assigned to that set, on all health states in that set and on both the occasions (test and retest). We first describe the computational steps and then present results for each of the four sets of health states.

As discussed earlier, the analysis for the generalisability study starts with partitioning of the variance components. For this purpose let *a, b, c,* respectively be the number of valuers, health states and occasions. In this case, we have 11 health states and two occasions. The number of valuers vary depending on the set of health states. Further let MS = Mean Squares i.e. mean squared deviation terms, and TSS = Total sum of squares i.e. sum of the squared deviation of each observation from the grand mean. MS with the appropriate subscript v, h or o and their combinations represents the mean square for the concerned facet or their interaction terms. To partition the variance components, we first compute the following mean squares from the given data.

$$MS_v = \frac{abc \sum_{v=1}^{a} (\bar{x}_{v\cdot} - \bar{\bar{x}})^2}{a-1}$$

$$MS_h = \frac{abc \sum_{h=1}^{b} (\bar{x}_{\cdot h} - \bar{\bar{x}})^2}{b-1}$$

$$MS_o = \frac{abc \sum_{o=1}^{c} (\bar{x}_{\cdot o} - \bar{\bar{x}})^2}{c-1}$$

$$MS_{vh} = \frac{abc \sum_{v=1}^{a} \sum_{h=1}^{b} (\bar{x}_{vh} - \bar{x}_{v\cdot} - \bar{x}_{\cdot h} + \bar{\bar{x}})^2}{(a-1)(b-1)}$$

$$MS_{vo} = \frac{abc \sum_{v=1}^{a} \sum_{o=1}^{c} (\bar{x}_{v\cdot o} - \bar{x}_{v\cdot\cdot} - \bar{x}_{\cdot\cdot o} + \bar{\bar{x}})^2}{(a-1)(c-1)}$$

$$MS_{vo} = \frac{abc \sum_{h=1}^{b} \sum_{o=1}^{c} (\bar{x}_{\cdot ho} - \bar{x}_{\cdot\cdot o} - \bar{x}_{\cdot h\cdot} + \bar{\bar{x}})^2}{(b-1)(c-1)}$$

$$TSS = \sum_{v=1}^{a} \sum_{h=1}^{b} \sum_{o=1}^{c} (x_{vho} - \bar{\bar{x}})^2$$

$$MS_{vho,e} = \frac{TSS - SS_v - SS_h - SS_o - SS_{vh} - SS_{vo} - SS_{ho}}{(a-1)(b-1)(c-1)}$$

We then arrive at estimates of variance components using the mean squares based on expected mean square equations from Shavelson and Webb (1991, p33).

$\hat{\sigma}^2_v = \frac{MS_v - MS_{vho,e} - bMS_{vo} - cMS_{vh}}{bc}$ i.e. variance component - valuers,

$\hat{\sigma}^2_h = \frac{MS_h - MS_{vho,e} - bMS_{ho} - cMS_{vh}}{ac}$ i.e. variance component - health states,

$\hat{\sigma}^2_o = \frac{MS_o - MS_{vho,e} - aMS_{ho} - bMS_{vo}}{ab}$ i.e. variance component - occasions of measurement,

$\hat{\sigma}^2_{vh} = \frac{MS_{vh} - MS_{vho,e}}{c}$ i.e. variance component - valuer health state interaction,

$\hat{\sigma}^2_{vo} = \frac{MS_{vo} - MS_{vho,e}}{b}$ i.e. variance component - valuer occasion interaction,

$\hat{\sigma}^2_{ho} = \frac{MS_{ho} - MS_{vho,e}}{a}$ i.e. variance component-health state occasion interaction, and

$\hat{\sigma}^2_{vho,e} = MS_{vho,e}$ i.e. variance component - random error.

Table 6.8: Estimated variance components for VAS-based health state valuations from village population.

Source	df	MS	Variance Component	% Total Variance	% total variance from sets 2-4		
					Set-2	Set-3	Set-4
Valuers (v)	22	0.05	0	-2	3	3	2
Health states (h)	10	2.58	0.05	60	56	67	56
Occasions (o)	1	0.06	0	0	1	0	-1
v * h	220	0.03	0	6	-29	-25	-31
v*o	22	0.08	0	6	2	0	3
h*o	10	0.05	0	1	0	0	2
vho,e	220	0.02	0.02	29	67	54	69
TSS	253	0.17	0.09	100	100	100	100

[1]Columns 2 to 5 shows details of variance components analysis for set-1 health states.

Results of the variance component analysis separately done for the four sets of health states is shown in Table 6.8. About 56 to 67% of variance is attributed to health states, the primary object of our measurement. The generalisability coefficient ranges from 0.56 to 0.67. Peculiarly some variance components are negative. Negative variance components could be due to misspecification of the measurement model or due to of sampling error (Shavelson and Webb, 1991). Most of the negative variance components in this study reside either in the valuers or interaction terms of valuers and health states. These negative variance components are compensated by a larger and positive error variance. This could be due to unstable valuations for certain health states, and differences in stability of valuations for different health states. The problem of unstable valuations and valuer health state interaction can not

be dealt by changes in the measurement model. Instead, a larger sample size may help improve stability of measurements at the group level.

Table 6.9: Generalizability of health state values by VAS. Variance components in percentages, reported by different studies.

Source	APHSV99	Van Agt et al (1994)[1]	Shibuya (1999)[2]
Valuers (v)	-2 to 3	2.87	3.8
Health states (h)	56 to 67	81.96	71.3
Occasions (o)	-1 to 1	0.05	0.3
v * h	-31 to 6	4.35	12.5
v*o	0 to 6	1.31	3.2
h*o	0 to 2	0.12	0.1
vho,e	29 to 69	9.33	8.8

[1]Standard EuroQol instrument. Postal survey in Rotterdam, Netherlands, Jan. 1991.
[2]Ordinal rank consistent VAS. Medical students in Japan, 1999.

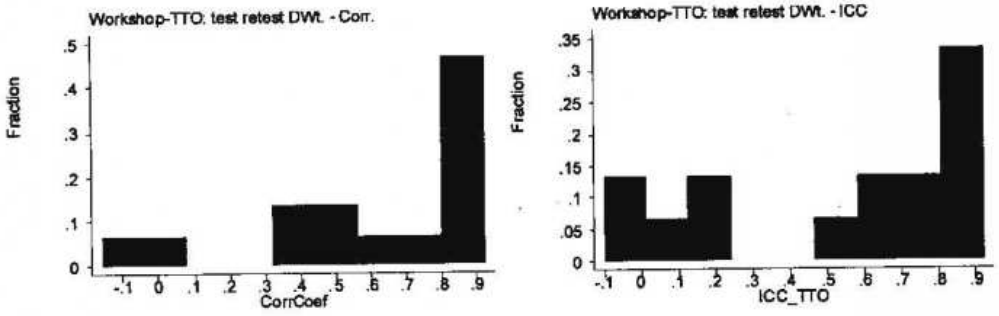
Generalisability study to assess reliability of health status measurements is being done recently. We are aware only of two studies in the area of health state valuation conducted hitherto. Helen van Agt and others (1994) did a generalisability study of health state valuation using different versions of the EuroQol instrument administered through a postal survey. Shibuya (1991) has compared different health state valuation methods used by medical students in Japan. A study by Krabee et al (1997) performed the generalisability analysis, but in the context of comparing different methods of valuation, and is therefore excluded from the comparisons here. Table-6.9 compares results from three studies, including the present one. Generalisability coefficients (simply the % variance due to health states expressed as a proportion) obtained in this study are comparatively lower than the other two studies. This difference could be due to the educational status of the valuer population. The Netherlands study was on educated urban professional and the study in Japan involved medical students where as this study was done in an Indian village, with many of the valuers being illiterate. Hence we conjecture that the slightly lower generalisability coefficient could be due to the difference in educational status of the valuers.

Reliability of health state valuations by TTO:

Product moment and intra class correlation coefficients were computed for valuations by the TTO method. Correlation between test and retest were computed for each of the test retest valuer. The same 15 workshop participants attending the retest workshop repeated the TTO exercise. Figure 6.10 shows the frequency distribution correlation coefficients (left graph) and ICCs (right

graph). The distribution is bimodal. A majority of participants had high correlation and concordance between their valuations on the two occasions. The poor correlation of test- retest valuations for some individuals could to some extent be attributed to their discomfort with the TTO valuation method and on account of unstable valuations for different health states.

Figure-6.10: Distribution of TTO test retest product moment and intra class correlation.



Intra class correlation coefficients were estimated by health state (Table-6.10). The overall ICC for all health states is about 0.44 which is lower than the level of overall concordance achieved by the VAS (0.81) for the same number of valuers and health states. Some health states show negative ICCs, suggesting unstable valuations for these states in the minds of the valuers. Two of such states, namely, continuous moderate back pain and vitiligo had negative ICCs under the VAS. For some health states, the ICC under VAS and TTO differed in the direction of agreement: for example, infertility and angina. Since the sample size is quite small in many cases (2 to 5) we can not attach much significance to the health state ICC statistics.

Table - 6.10: TTO test retest ICC by health states

Health States	n	ICC	Health States	n	ICC
All	165	0.438	Bronchitis	4	0.84
Mild diabetes, no symptoms	15	0.23	Infertility	4	0.36
Mild tuberculosis with treatment	15	0.25	Pain and stiffness of joints	4	0.23
Own health	15	-0.02	Schizophrenia	4	-0.52
Quadriplegia	15	0.01	Severe hallucinatory fever	4	-0.27
Severe migraine	15	-0.35	Urinary incontinence	4	-0.23
Unipolar major depression	15	0.20	Blindness	4	0.62
Watery diarrhoea	15	-0.10	Angina	4	-0.51
Continuous moderate back pain	5	-0.46	Below knee amputation - one leg.	2	0.91
Mild hearing disorder	5	0.25	Below knee amputation - two legs.	2	0.92
Severe heart failure	5	-0.15	Peptic ulcer	2	0.33
White marks on face	5	-0.94	Two broken arms in cast	2	0.98

Validity of the Health State Valuation Measurements:

An instrument is valid if it measures what it is intended to measure. Here we are trying to measure the value that people assign to life in different health states. We assess the validity by looking at the instrument's performance from different perspectives. The following three main perspectives are usually studied. Firstly, instrument's content, which would include administration protocol as well. Secondly, the comparative performance of the instrument in relation to a criterion: for example, performance in relation to a "gold standard" and finally, the consistency of measurements by the instrument with theoretical constructs around the subject of measurement. The above three perspectives are commonly referred to as content or face validity, criterion validity and construct validity respectively. Such descriptions have erroneously suggested the existence of a typology of validity. However, it is important to recognise that validity is a single concept. We are interested to know if an instrument is valid. We try to infer this by looking at the instrument from different perspectives.

Details of the health state description system used to describe the health states subjected to valuation have been described earlier. Adequacy, accuracy and effective communication of health states contribute to the content of these instruments. The theoretical basis of the three scaling strategies, namely VAS, TTO and PTO have already been referred to earlier. Each of these three scaling strategies were aided by card sorting. Ordinal ranking of a set of conditions is considered to be a primordial expression of preference and valuation. Hence we believe that modification of the three scaling techniques by seeking consistency with card sort by the same valuer enhances the content of these instruments. The written and spoken instructions, examples and the health state valuation protocol, all contribute to the instrument's content.

A "gold standard" measure of health state values does not exist. Hence a criterion-related validity assessment is not feasible. Instead, we turn to examine consistency of measurements with different constructs about psychometric measurements in general and health state valuation in particular. One construct frequently resorted to is convergence. If measurements from an instrument converge with measurements from other instruments appropriately built to measure the same concept, we take the evidence as a support to validity of the instrument. Table 6.11 shows correlation of visual analogue scale valuations with results from two other scaling methods, namely TTO and PTO. Correlation coefficient for VAS and TTO scores is 0.8. The correlation between VAS and PTO scores is 0.93 and between TTO - PTO scores it is 0.89. The correlation with PTO valuations is based on a small number of valuations (seven valuers and 77 valuations). We consider the correlation coefficients of about 0.8 between different scaling strategies as suggestive of convergence.

Table-6.11: Correlation of health state values obtained from different methods and effect of deliberation.

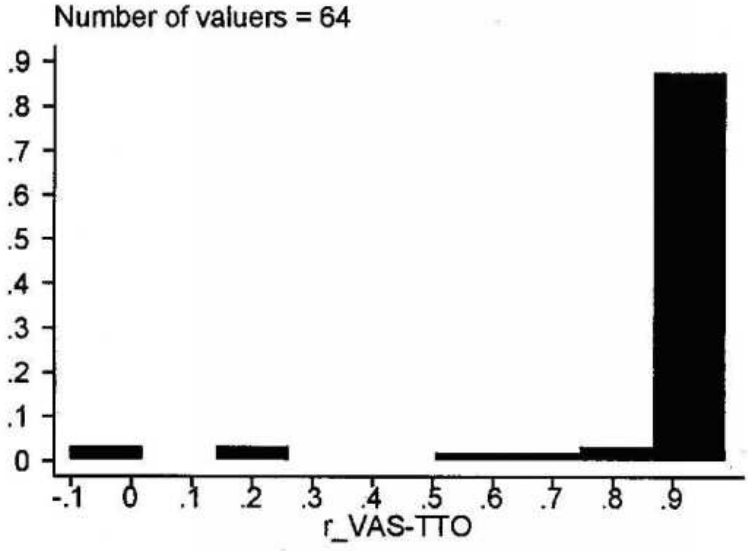
Method	TTO	PTO
First TTO / PTO valuation		
VAS All valuations (180 valuers	0.51 (1969)	PTO1: 0.02 PTO2: 0.03 (319)
Ordinal rank consistent (162 valuers)	0.52 (1782)	PTO1: 0.47 PTO2: 0.28 (275)
TTOFirst iteration (179 valuers)		PTO1: -0.00 PTO2: 0.13 (319)
Last round of TTO / PTO valuation		
All valuations (180 valuers)	0.62 (1969)	0.89 (82)
Ordinal rank consistent (162 valuers)	0.64 (1782)	0.89 (82)
All valuations (179 valuers)		0.84 (82)
Ordinal rank consistent (70) valuers)		0.87 (71)
All valuations consistent with card sort		
VAS	0.80 (770)	0.93 (77)
Time trade-off (TTO)		0.89 (77)

[1]Figures within parentheses show number of valuations based on which the correlation coefficient is estimated. This number divided by 11 rounded up gives the number of valuers which is equal to the min (row method valuers, column method valuers).

The reader may recall that valuation protocol encouraged a deliberative iterative process. Key components of this process were multiple iteration of valuation - consistency with card sort feed back loop. The top panel of Table 6.11 "First TTO / PTO valuation" shows the correlation of VAS scores with the first round of TTO and PTO values. The correlation of VAS and first round TTO scores stood at 0.51. This improves marginally to 0.51 if we restrict to ordinal rank consistent VAS scores . The improvement in correlation between VAS and PTO is more marked (0.04 to 0.47 and 0.07 to 0.28 for PTO1 and 2 respectively). Some valuers did not pursue the TTO / PTO exercise till their ordinal rankings matched completely with the card sort ranks while some others pursued these exercises till the rankings matched with card sort. The middle panel, "Last round of TTO /PTO valuation" shows correlation between VAS with the last round of TTO and PTO scores. The last round of TTO / PTO scores includes scores from valuers whose card sort ranking did not match fully and those whose ranking matched completely. Correlation of VAS with TTO scores at this stage improved from 0.51 / 0.52 earlier to 0.62 / 0.64 for amalgamated and ordinal rank consistent VAS scores respectively. Stronger correlation (0.89) between the VAS to PTO valuations appeared. These correlation improved further if the TTO / PTO valuations were restricted to completely matched cases only (i.e. card sort ranking and TTO / PTO ranking for these valuers matched completely). Correlation of VAS - TTO scores improved from 0.6 to 0.8. These findings are consistent with our belief that the ordinal rank consistency criteria and use of the deliberative interactive tools help clarify the valuation process.

Now let us consider at convergence at disaggregated level. We can dissagregate the correlation between different scaling methods by valuer and by health states. First we look within valuer correlation between scaling methods. Correlation coefficients were estimated for each valuer, between their valuations using VAS and TTO methods. The set of health states assigned to the individual did not change between instruments. There are 64 valuers in the data set whose valuations are ordinal rank consistent for both VAS and TTO valuations. Figure 6.11 shows the frequency distribution of these correlation coefficients. Clearly, valuations from the two instruments correlated very strongly at the individual level. Most of the correlation coefficients were about 0.9. Similar correlation between VAS and PTO scores showed that all seven (i.e. the number of valuers who completed ordinal rank consistent PTO) were 0.9 and above.

Figure-6.11: Frequency distribution of within valuer correlation between VAS and TTO valuations.



Let us now consider to convergence at the health state level. Table 6.12 shows correlation of VAS scores with TTO and PTO scores for each health state. The health states are shown in descending order of estimated correlation coefficient of VAS and TTO. Valuations by different instruments positively correlated with each other for most health states. However, VAS and TTO valuations for a few health states, like common cold, schizophrenia and below the knee amputation of two legs, did not correlate at all. It is difficult to say whether such lack of correlation would suggest systematic interaction between health state and scaling method, lack enough sample valuations or some unknown factor. This aspect needs to be investigated further.

Table-6.12: Within health state correlation of valuations by different instruments

Health state	VAS-TTO		VAS-PTO	
	n	*n*	*n*	*n*
Severe hallucinatory fever	5	0.99		
Bronchitis	14	0.86		
Infertility	15	0.82		
Urinary Incontinence	14	0.80		
White marks on face	14	0.79		
Watery diarrhoea 5 times a day	64	0.77	7	0.93
Blindness	15	0.71		
Mld diabetes, no symptoms	64	0.70	7	0.73
Mild tuberculosis with treatment	64	0.65	7	0.90
Severe congestive heart failure	14	0.62		
Unipolar major depression	64	0.62	7	0.67
Severe migraine	64	0.61	7	0.89
Continuous moderate back pain	14	0.60		
Mild hearing disorder	14	0.53		
Angina	15	0.44		
Pain and stiffness in joints	14	0.41		
Valuer's own health state	64	0.39	7	0.49
Two broken arms in cast	12	0.39		
Quadriplegia	64	0.29	7	0.83
Peptic ulcer	21	0.24		
Amputation of one leg below knee	21	0.21		
Moderate anaemia	9	0.11		
Common cold	10	0.01		
Schizophrenia	14	-0.01		
Amputation of both legs below knee	21	-0.10		

Logical consistency of valuations would throw some light on validity of the instruments. Our subject of valuations, namely the health states, differed in their 6D5L profiles. We identified pairs of health states such that one of the two 6D5L profile weakly dominates (i.e. is worse in at least one dimension and same in other dimensions). We have nine such dominant - dominated (dd) health state pairs (Table 6.13). We looked for valuations under a scaling method where DWt(dominant state) DWt (dominated state). Lets call such a valuation as counterintuitive. We counted such counter intuitive valuations under each scaling method.

Table 6.13: Incidence of counter intuitive valuations for dominating and dominated pairs.

Pair #	Dominating (first line) and Dominated (second line) Health State	6D5L	NDD[1]	distance	All Valuations		Card sort matched	
					VAS	TTO	VAS	TTO
1	Quadriplegia	554341			11%	34%	10%	22%
	Amputation below the knee (both legs)	433221	5	7	45	44	41	23
2	Severe Hallucinatory Fever	444333	6	8	19%	35%	21%	6%
	Blindness	323122			31	31	28	16
3	Amputation below the knee (both legs)	433221			2%	14%	0%	0%
	Amputation below the knee (one leg)	322211	4	4	45	44	41	23
4	Mild Tuberculosis with treatment	111221	1	1	19%	25%	17%	19%
	Mild diabetes, no symptoms	111121			180	179	163	70
5	Mild Tuberculosis with treatment	111221	1	1	19%	39%	19%	26%
	Watery Diarrhoea 5 times a day	111211			180	179	163	70
6	White marks on face	111131	1	1	62%	64%	62%	67%
	Mild diabetes, no symptoms	111121			45	45	42	15
7	Angina	111321	1	1	40%	49%	38%	38%
	Mild Tuberculosis with treatment	111221			45	45	40	16
8	Mild hearing disorder	112121	1	1	62%	56%	62%	60%
	Mild diabetes, no symptoms	111121			45	45	42	15
9	Severe continuous migraine	113431	1	1	38%	53%	35%	56%
	Urinary incontinence	113331			45	45	40	16

[1]NDD = Number of dominating dimensions.
[2]For each pair the top row shows % of valuations where value assigned to dominating condition was less than equal to the value assigned to the dominated condition. The bottom row shows the number of valuations for this pair i.e. the denominator for the % shown in top row.

Table6.13 shows occurrence of counter intuitive valuations for the nine dominating and dominated pairs under VAS and TTO methods. The two columns under "All valuations" show the occurrence of counter intuitive valuations for all valuers irrespective of whether their valuation was consistent with the ordinal ranks assigned by them to the same health states. The next two columns (right most) under "Card sort matched" restricts the denominator set to ordinal rank consistent valuations only. The column titled NDD shows the number of dimensions in which the dominant condition's profile is strictly worse than that of the dominated state. The NDD value can be considered as a measure of the

magnitude of dominance. The column titled "distance" shows difference in the equally weighted sum of the severity level codes contained in 6D5L profile of dominant and dominated condition. For example, the severity codes in the 6D5L profile for quadriplegia add up to 22 and the same for amputation of both legs below the knee add up to 15. So the distance between the two health state profiles is taken as 7. Note that this assumes an explicitly defined multifaceted health state value model consisting of six attributes and each attribute weighting equally. This need not be actually the case. The imperfect distance measure thus arrived, however, gives us some idea of the magnitude of difference in the two profiles.

The disability weight of the dominant state is expected to be greater than the value of the dominated state. If a dominant - dominated pair of health state is valued by many individuals using a perfectly valid instrument, the frequency of counterintuitive valuations will tend to zero as the number of valuers increase. In a less than perfect but real world, occurrence of counterintuitive valuations will be rarer as the validity of the instrument improves. One would also the expect that occurrence of counter intuitive valuations for a *dd* pair will be less as the *NDD* value and distance between the two increases. Earlier we have argued that requiring ordinal rank consistency improves instrument validity. That would imply that occurrence of counter intuitive valuations would be comparatively less for ordinal rank consistent valuations as opposed to amalgamated valuations containing both consistent and inconsistent measurements. Now let's examine the numbers in Table 6.13 in the light of the above theoretical expectations. Occurrence of counter intuitive valuations under card sort matched scaling was 0 to 21% for pairs with NDD values of 4 to 6 with distances from 4 to 8. Compare this with occurrences ranging from 17% to 62% for pairs with NDD value or distance of one. VAS appears to produce relatively less counterintuitive valuations in comparison to TTO. Ordinal rank consistency appears to slightly reduce the occurrence of counterintuitive valuations. But the magnitude of this effect is small for VAS measurements. Consistency of TTO measurements appear to improve much more when ordinal rank consistency is insisted. Further study is required to understand what aspect of these instruments, mode of administration, etc. need to be changed to reduce occurrence of counterintuitive valuations.

Relationship of VAS measurements to TTO valuations:

Mean disability weights, their standard errors (SE), and number of observations (n) obtained from different valuation methods in the workshops is shown in Table 6.14. Results from the PTO exercise are based on very few participants. Although we have shown the mean values here, we do not use them for comparative purposes in view of the small sample size.

Table 6.14: Mean disability weights from the workshops.

Health state	VAS			TTO			PTO		
	n	Mean	SE	n	Mean	SE	n	Mean	SE
Angina	45	0.46	0.03	45	0.47	0.03	1	0.70	
Below knee amputation one leg.	45	0.51	0.04	44	0.45	0.04	2	0.45	0.20
Below knee amputation two legs	45	0.69	0.03	44	0.64	0.04	2	0.85	0.05
Blindness	45	0.64	0.03	45	0.56	0.05	1	0.69	
Bronchitis	45	0.35	0.03	45	0.29	0.04	2	0.39	0.13
Common cold	14	0.12	0.03	14	0.09	0.02	1	0.09	
Continuous moderate back pain	45	0.36	0.03	45	0.36	0.03	3	0.47	0.12
Infertility	45	0.37	0.04	45	0.41	0.04	1	0.60	
Mild hearing disorder	45	0.21	0.03	45	0.28	0.04	2	0.19	0.01
Mild tuberculosis with treatment	180	0.42	0.02	179	0.42	0.02	8	0.68	0.06
Moderate anaemia	15	0.29	0.06	15	0.20	0.04	2	0.28	0.00
Mild diabetes	180	0.29	0.02	179	0.26	0.02	8	0.50	0.08
Own health	180	0.03	0.01	179	0.05	0.01	7	0.11	0.02
Peptic ulcer	45	0.36	0.03	44	0.35	0.03	2	0.60	0.00
Pain and stiffness in joints	45	0.49	0.03	45	0.42	0.04	2	0.79	0.03
Quadriplegia	180	0.86	0.01	179	0.75	0.02	7	0.86	0.04
Severe hallucinatory fever	31	0.77	0.04	31	0.67	0.06	0		
Severe heart failure	45	0.73	0.03	45	0.65	0.05	3	0.75	0.13
Severe migraine	180	0.50	0.02	179	0.46	0.02	7	0.51	0.06
Schizophrenia	45	0.79	0.03	45	0.71	0.04	2	0.91	0.02
Two broken arms in cast	30	0.59	0.04	29	0.66	0.04			
Unipolar major depression	180	0.60	0.02	179	0.51	0.02	7	0.71	0.04
Urinary incontinence	45	0.50	0.03	45	0.41	0.04	2	0.79	0.01
Watery diarrhoea	180	0.25	0.01	179	0.27	0.02	8	0.44	0.08
White marks on face	45	0.24	0.03	45	0.26	0.04	2	0.41	0.17

Some researchers (for example, Torrance, 1976) have argued that the valuations obtained through VAS and TTO are not comparable. One hypothesis implied by Torrance (1976) is that the valuations from trade-off techniques reflect the true stimulus in the mind of the valuer. What we measure is the response of the valuer to his / her endogenous stimulus. We would then model such a relationship with a power function from stimulus response theory in psycho physics[13]. If we assume that the VAS valuation is the response and trade-off

[13] See McDowell and Newell, 1996 p15-18 for a brief summary of results from psycho physics about te power function in the context of health status measurement.

valuation like the TTO is a more direct reflection of the endogenous stimulus in valuer's mind then the power function relating the two measurement can be written as;. $DW_{VAS} = k\ DW^{b}_{TTO}$

To test if the VAS and TTO valuations were different, we did a pair-wise difference of means test. The null hypothesis of no difference in means from VAS and TTO was rejected at 95% level confidence (p value = 0.0108). This gives credence to the view that the valuations from two methods are different. Hence we estimated parameters of the power function relating the two valuations. For this purpose, the power function described above can be linearised as $\ln(DW_{VAS}) = \ln(k) + b\ \ln(DW_{TTO})$. Ordinary least squares (OLS) estimation of this linearised model using data from this study, gives the model.

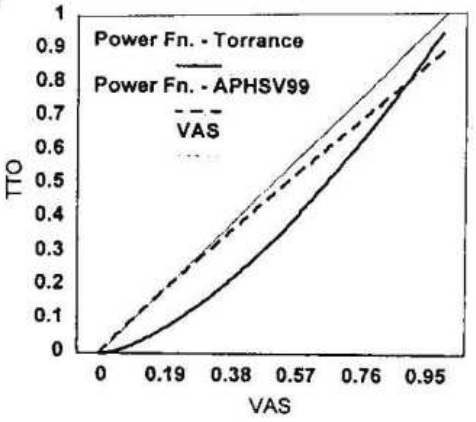
$$\ln(DW_{VAS}) = .1128692 + 1.063455\ \ln(DW_{TTO}).$$

Recovering k and b from the estimated model, we have the following equation to convert VAS valuations to a ratio scale; $DW_{TTO} = \left(\frac{DW_{VAS}}{k}\right)^{\frac{1}{b}} = \left(\frac{DW_{VAS}}{1.12}\right)^{\frac{1}{1.06}}$

where $k = 1.12$ and $b = 1.06$ (compare these with k=1.03 and b=0.65 obtained by Torrance, 1976). Although the model is statistically significant ($p<0.001$) and has a good fit (Adjusted $R^2=0.95$), the 95% confidence interval of the estimated parameter b is .9665101 to 1.160399 straddling one within it. Thus we can not reject the null hypothesis that the true value of parameter $b = 1$.

Figure 6.12 shows the plots of TTO disability weights predicted from our VAS-based measurements using the model estimated by us from this study (made up of dashed line), and the model estimated by Torrance (thick continuous line). The thin straight line represents VAS measurements without any transformation. Clearly, we did not find differences between VAS and TTO measurements, to the extent observed by Torrance (19976) among the Canadian population.

Figure 6.12: Power function models of TTO from VAS: Torrance (1976) and APHSV99.



$DW_{VAS} = k\ DW^{b}_{TTO}$

	Torrance	APHSV
k	1.03	1.12
b	0.65	1.06

Minimal differences between VAS and TTO based valuations found in this study could have more than one explanation. Firstly, it may be true that ordinal rank consistent VAS measurements of health state valuations are not very different from the TTO based valuations. Secondly, a design feature of this study might have blurred the differences between VAS and TTO based valuations. In the APHSV study, the TTO worksheets presented to each valuer had about 8 to 10 alternative durations of life in perfect health presented to the valuers. We did this to be more effective in communicating the time trade-off idea to the valuers. These alternatives had been calculated starting with say 95% of life expectancy at the age of onset and progressively decreasing to 5% of life expectancy at the age of onset of the concerned health state. One possibility might be that valuers did not consider a duration of healthy life beyond the lower and upper bounds contained in the worksheets. In such a case, relatively milder health states would be valued as worse, if the valuer did not consider to trade a duration of life less than 5% of the life expectancy at the age at onset. Severe conditions will be valued better, if the valuer did not consider trading a duration of life more than 95% of the life expectancy at age at onset. If this framing effect acted only predominantly for the milder conditions, then the TTO valuations have been biased upwards. However, if such a framing effect did in fact operate, then we would not have any observations of disability weight less than 0.05 or more than 0.95. We filtered the MDHSV workshop data, by excluding the valuations for own health state, and all valuations where the disability weight from TTO valuations was in the range [0.05, 0.95]. After filtering these out, we get 8% valuations where the assigned disability weight was either less than 0.05 or more than 0.95 given by 40% of the total workshop valuers. Thus 40% of the valuers did in fact chose valuations outside the range suggested by the alternatives given in the worksheets. A little more than half of these (22%) valuers chose valuations giving disability weights as less than 0.05. Milder conditions like watery diarrhoea and mild diabetes received many such valuations. That would mean that the framing effect, if any, of the specific alternatives in the TTO worksheets was either non existent or minimal. Although the TTO valuations observed in this study are not very much lower than the disability weights given by VAS, the direction is similar to the model estimated by Torrance (1976).

Figure 6.13: Two way scatter plot of mean disability weights from TTO and VAS

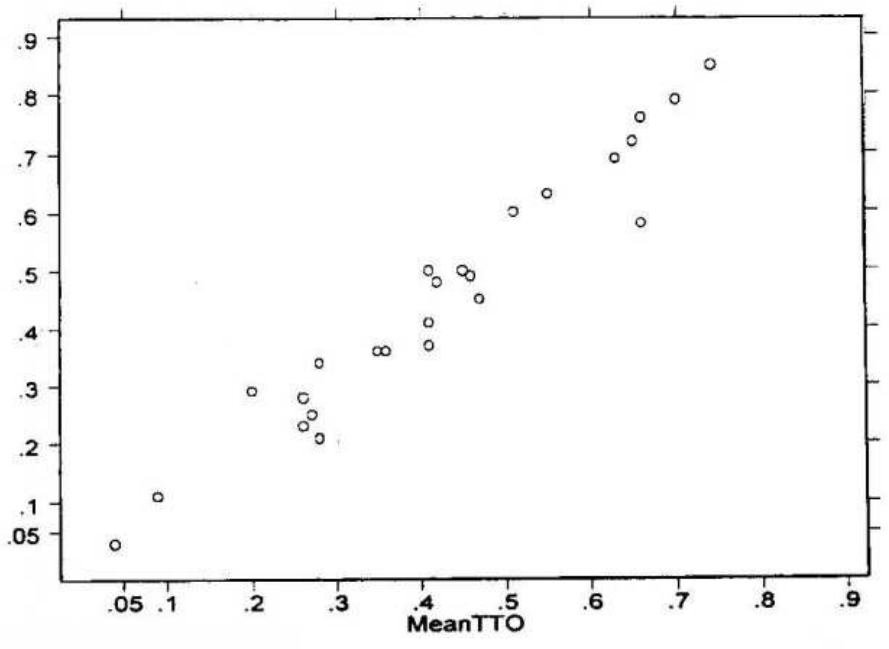


Figure - 6.14: One-way scatter plot of mean disability weights for different health states obtained by visual scaling (VAS) and time trade-off (TTO).

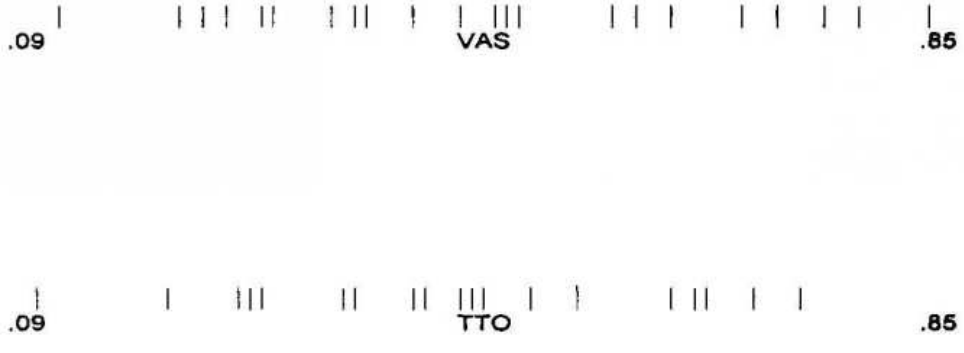
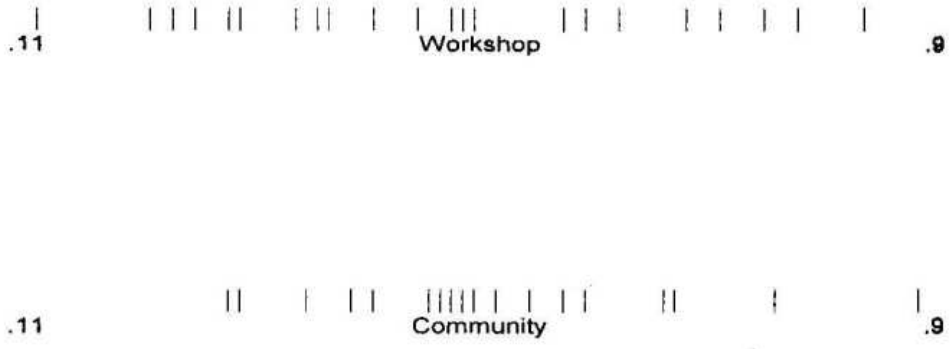


Figure - 6.15: One-way scatter plot of mean disability weights for different health states obtained by visual scaling (VAS) from MDHSV Workshops and Community Survey



Another potential problem with rating scales is that the valuations may tend to cluster towards the midpoint of the scale or at both ends. Figure 6.13 shows a two-way plot of mean disability weights from TTO and VAS. Hardly any difference is visible in the spread of mean valuations across the full range of 0 to 1 scale . Differences in spread can be better appreciated with one-way plots shown in Figure 6.13. The upper plot shows spread of the VAS valuations and the lower one shows the same for TTO valuations. There is not much difference in the spread of valuations from two methods. If at all, the VAS valuations are slightly more spread out than the TTO valuations. So transforming the VAS valuations using the power function model described earlier would either leave the spread of VAS values as it were or marginally narrow it down. The spread of mean disability weights for different health states obtained from the community-based survey shrinks towards middle part of the scale (Figure 6.15). This could be due to real differences between valuations by the community and the participants in the workshops, and/or due to measurement error. This will have to be investigated further.

References

Allen D., Lee R.H., and Lowson K. 1989. The use of QALYs in health service planning. International Journal of Health Planning and Management 4: 261-73.

Andrews FM and Withey SB. Social indicators of well being: Americans' perceptions of life quality. New York: Plenum; 1976. Cited in McDowell and Newell, 1987, 1996 ibid.

Berg Robert L. 1973. Establishing the value of various conditions of life for a health status index. in Health status indexes. Proceedings of a conference conducted by Health Services Research, Tucson Arizona, October 1-4, 1972. Chairman and Editor Berg Robert L. Chicago: Hospital Research and Educational Trust.

Bergner M., Bobbit R. A., Carter W. B., and others. 1981. The sickness impact profile: development and final revision of a health status measure. Medical Care 19: 787-805.

Bergner M., Bobbit R. A., Kressel S., and others. 1976b. The sickness impact profile: conceptual formulation and methodology for the development of a health status measure. International Journal of Health Services 6: 393-415.

Bergner M., Bobbit R.A., Pollard W. E., and others. 1976a. The sickness impact profile: validation of a health status measure. Medical Care 14: 57-67.

Bleichrodt Han. 1997. Health utility indices and equity considerations. Journal of Health Economics 16: 65-91.

Blischke W. R.; Bush J. W., and Kaplan R. M. Successive intervals analysis of preference measures in a health status index. Health Services Research. 1975; 10:181-198.

Boyle MH and Torrance GW. Developing multiattribute health indexes. Medical Care. 1984; 22:1045-1057.

Boyle Michael; Furlong William; Torrance George, and Feeny David. Reliability of the Health Utilities Index - Mark III Used in the 1991 Cycle 6 General Social Survey Health Questionnaire. Ontario: Center for Health Economics and Policy Analysis; 1994.

Brazier JE; Harper R.; Thomas K.; Jones N., and Underwood T. Deriving a preference based single index measure from the SF-36. Journal of Clinical Epidemiology. 1998; 51:1115-1129.

Brook RH; Ware JE; Davis-Avery A and others. Overview of adult health status measures fielded in Rand's health insurance study. Medical Care Supplement. 1979; 17(1).

Brooks Richard and EuroQol group. EuroQol: the current state of play. Health Policy. 1996; 37:53-72.

Carmines Edward G. and Zeller Richard A. Reliability and validity assessment. Beverly Hills, London, New Delhi: Sage publications; 1979; ISBN: 0-8039-1371-0.

Casella George and Berger Roger L. Statistical inference. Belmont, CA: Duxbury Press; 1990.

Chambers LW; Sacket DL, and Goldsmith CH. Development and application of an index of social function. Health Services Research. 1976; 11:430-441.

Chronbach LJ, Gleser GC, Nanda H. and Rajaratnam N.; The dpendability of behavioral measurements: Theory of generalizability for scores and profiles, Wiley New York, 1972, cited in Streiner and Norman (1995) and Shavelson and Webb (1991).

Cohen J. Weighted Kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin. 1968; 70:213-220.

Culyer Anthony J. 1989. The normative economics of health care finance and provision. Oxford Review of Economic Policy 5, no. 1: 34-58.

Deyo RA; Diehr P, and patrick DL. Reproducibility and Responsiveness of Health Status Measures: Statistics and Strategies for Evaluation. Controlled Clinical Trials. 1991; 12:142S-158S.

Dolan Paul. Modelling valuations for EuroQol health states. Medical Care. 1997(11):1095-1108.

Dolan Paul; Gudex C.; Kind P., and Williams A. The time tradeoff method: results from a general population study. Health Economics(5):141-154.

Euroqol Group. EuroQol. A new facility for measurement of health-related quality of life. Health Policy. 1990; 16:199-208.

Froberg D.G. and Kane R.L. Methodology for measuring health-state preferences, part-I: measurement strategies. Journal of Clinical Epidemiology, 1989; 42(4):345-354.

Froberg D.G. and Kane R.L. Methodology for measuring health-state preferences, part-II: scaling methods. Journal of Clinical Epidemiology. 1989; 42(5):459-471.

Froberg D.G. and Kane R.L. Methodology for measuring health-state preferences, part-III: population and context effects. Journal of Clinical Epidemiology. 1989; 42(6):585-592.

Froberg D.G. and Kane R.L. Methodology for measuring health-state preferences, part-IV: progress and a research agenda. Journal of Clinical Epidemiology.1989; 42(7):675-685.

Goldberg M. and Dab W. Complex indexes for measuring a complex phenomenon. in: Abelin T.; Brzezinxki Z.J., and Carstairs Vera D.L., Eds. Measurement in health promotion and protection. Copenhagen: WHO Regional Office for Europe; 1987; pp. 174-194.

Gudex C.; Dolan P.; Kind P., and Williams A. Health state valuations from the general public, using the visual analog scale. Quality of Life Research. 1996; 5:521-531.

Guilford J.P. 1954. Psychometric methods. New York: McGraw Hill; page 24. Cited in Patrick et al 1973.

Guillemin F., Bombardier C., Beaton D. 1993. Cross-cultural adaptation of health-related quality of life measures: literature review and proposed guidelines. Journal of Clinical Epidemiology 46(12): 1417-1432.

Halford Graeme S. Development of processing capacity entails representing more complex relations: implications for cognitive development. in: Logie Robert L. and Gilhooly Kenneth J., Editors. Working Memory and Thinking. Hove, East Sussex, UK: Psychology Press Ltd; 1998; pp. 139-157.

Hunt SM; McKenna SP; McEwen J. and others. The Nottingham Health Profile: subjective health status and medical consultations. Social Science and Medicine. 1981; 15:221-229. Cited in McDowell, ibid.

Kaplan R. M. and Anderson J. P. A general health policy model: update and applications. Health Services Research. 1988; 23:203-235.

Kaplan R. M.; Bush J. W., and Berry C.C. Health Status Index: category rating versus magnitude estimation for measuring levels of well being. Medical Care. 1979; 17:501-523.

Kaplan R. M.; Bush J. W., and Berry C.C. Health status: types of validity and the index of well being. Health Services Research. 1976; 11:478-507.

Kaplan RM and Bush JW. Health-related quality of life measurement for evaluation of research and policy analysis. Health Psychology. 1982; 1:61-80.

Kaplan Robert M. Using quality of life information to set priorities in health policy. Social Indicators Research. 1994; 33:121-163.

Kleinman, A. 1987. Anthropology and Psychiatry: the role of culture in cross-cultural research on illness. British Journal of Psychiatry 151: 447-454.

Krabbe Paul FM; Stouthard Marlies EA; Essink-Bot Marie Louise, and Bonsel Gouke. The effect of additing a cognitive dimension to the EuroQol multiattribute health-status classification system. Krabbe Paul M. The valuation of health outcomes. A contribution to the QALY approach. Rotterdam: Erasmus University Rotterdam, The Netherlands.; 1998 Jun; pp. 105-118.

Krabbe Paul M. The valuation of health outcomes. A contribution to the QALY approach. Rotterdam: Erasmus University Rotterdam, The Netherlands.; 1998 Jun.

Kramer MS and Feinstein AR. Clinical biostatistics LIV. The biostatistics of concordance. Clinical Pharmacology and Therapeutics. 1981; 29:111-123.

Lenert L.A. and Hornberger J.C. 1996. Computer-assisted quality of life assessment for clinical trials. Proceedings of the AMIA Annual Fall Symposium, 922-996.

Lenert L.A. and Soetikno R.M. 1997. Automated Computer Interviews to Elicit Utilities: Potential Applications in the Treatment of Deep Venous Thrombosis. Journal of the American Medical Informatics Association, 4(1): 49-56.

Leplege A., and Verdier A. 1995. The adaptation of health status measures: methodological aspects of the translation procedure. in The international assessment of health related quality of life. Theory, translation, measurement and analysis. Editors Shumaker Sally A., and Berzon Richard A.; Oxford, New York: Rapid Communications.

Llewellyn-Thomas H.A., Sutherland H.J. and Tibshirani R. 1984a. Describing health states: methodological issues in obtaining values for health states. Medical Care, 22: 543-552.

Magdelaine M., Misrahi A., Rosch G.; Un Indicateur de la Morbidite Applique aux Donnees dune Enquete sur la Consommation Medicale; Consommation, Annales du CREDOC Centre de Recherches et de Documentation sur la Consommation, 2, 3-41; cited by Rosser (1983).

Mahapatra Prasanta, Salomon Josh, Nanda Lipika, and Rajashree, KT; Measuring Health State Values in Developing Countries: Report of a study in Andhra Pradesh, India; Institute of Health Systems, HACA Bhavan, Hyderabad, AP 500004, India, Report.

Mas-Collel Andreu, Whinstone Michael D., and Green Jerry. 1995. Micro economic theory. New York, Oxford: Oxford University Press.

McDowell Ian and Newell Claire. Measuring health: a guide to rating scales and questionnaires. New York / Oxford: Oxford University Press; 1987.

McDowell Ian and Newell Claire. Measuring health: a guide to rating scales and questionnaires. New York / Oxford: Oxford University Press; 1996.

McHorney C.A., Ware J. E. Jr., Lu J. F. R., and others. 1994. The MOS 36 item short-form health survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. Medical Care 32: 40-66.

Murray Christopher J. L.; 1996. Rethinking DALYs. in The global burden of disease. A comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020. Editors Murray Christopher J. L., and Lopez Alan D. Boston: Harvard School of Public Health.

Murray Christopher J.L. and Lopez Alan D; The global burden of disease. A comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020. Boston: Harvard School of Public Health; 1996.

Murray Christopher J.L., and Lopez Alan D. 1996. The global burden of disease in 1990: final results and their sensitivity to alternate epidemiological perspectives, discount rates, age weights and disability weights. in The global burden of disease. A comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020. Editors Murray Christopher J. L., and Lopez Alan D. Boston: Harvard School of Public Health.

Nelson EC; Wasson J.; Kirk J and others. Assessment of function in routine clinical practice: description of the COOP chart method and preliminary findings. Journal of Chronic Diseases. 1987; 40 (Suppl 1):55S-63S, cited in McDowell and Newell, 1987, 1996 ibid.

Noack H. Concepts of health and health promotion. in: Abelin T.; Brzezinxki Z.J., and Carstairs Vera D.L., Eds. Measurement in health promotion and protection. Copenhagen: WHO Regional Office for Europe; 1987; pp. 5-28.

Packer A.H. Applying cost-effectiveness concepts to the community health system. Operations Research. 1968; 16:227-253.

Patrick and Erickson. Concepts of Health-Related Quality of Life. in: Patrick and Erickson. Health Status and Health Policy. 1993; pp. 76-112.

Patrick D. L.; Bush J. W., and Chen M. M.; 1973a; Towards an operational definition of health. Journal of Health and Social Behavior; 14:6-23.

Patrick D. L.; Bush J. W., and Chen M. M. Methods for measuring levels of well-being for health status index. Health Services Research. 1973b; 8:228-245.

Pennifer Erickson, Kendall Allen, Anderson John P., and Kaplan Robert M. 1989. Using composite health status measures to assess the Nation's health. Medical Care 27, no. 3 Supplement: S66-S76.

Rosser Rachel. Issues of measurement in the design of health indicators: a review. in: Culyer A.J., Editor. Health indicators. An international study for the European Science Foundation. Oxford: Martin Robertson; 1983; pp. 34-81.

Saariluoma Pertti. Adversary problem-solving and working memory. in: Logie Robert L. and Gilhooly Kenneth J., Editors. Working Memory and Thinking. Hove, East Sussex, UK: Psychology Press Ltd; 1998; pp. 115-138.

Sackett D.L. and Torrance G.W. The utility of different health states as perceived by general public. Journal of Chronic Diseases. 1978; 7:347-358.

Shavelson Richard J. and Webb Noreen M. Generalizability theory: A primer. Newburry Park / London / New Delhi: Sage Publications; 1991.

Shibuya Kenji; Quantifying the economic impact and health consequences of disease: implications for studies on smoking, unpublished thesis Harvard University School of Public Health, Boston, MA USA, 1999.

Specter Paul E. 1992. Summated rating scale construction. An introduction. Newburry Park, London, New Delhi: Sage.

Stewart Anita L. The medical outcomes study for work of health indicators. in: Stewart Anita L. and Ware John E. Jr, Editors. Measuring functioning and well-being. Durham//London: Duke University Press; 1992.

Streiner, D. L. & Norman G. R. 1995. Health Measurement Scales: A Practical Guide to their Development and use. New York: Oxford University Press.

Tolley George, Kenkel Donald, and Fabian Robert; 1994; Valuing health for policy. An economic approach; Chicago and London: University of Chicago Press.

Torgerson W.S. 1958. Theory and methods of scaling. New York: John Wiley; page 141. Cited in Patrick et al 1973.

Torrance G.W., and Boyle M. H. 1982. Application of multiattribute utility theory to measure social preferences for health states. Operations Research 30: 1043-69. Cited in Froberg and Kane 1989a.

Torrance George W. Measurement of health state utilities for economic appraisal. A review. Journal of Health Economics. 1986; 5:1-30.

Torrance George W. Social preference for health states: an empirical evaluation of three measurement techniques. Socio-Economic Planning Sciences. 1976; 10:129-136.

Torrance George W., Thomas Warren H., and Sackett David L. 1972. A utility maximization model for evaluation of health care programs. Health Services Research 7: 118-33.

Trotter, R.T., Ustun, B., Chatterji, S., Rehm, J., Room, R., Kennedy, C., Saxena, S., 1999. Cross-Cultural Applicability research on Disablement: Models, Methods and Contribution to Revision of the International Classification, Human Organization.

von Neumann J. and O. Morgenstern; 1944; Theory of games and economic behavior.; Princeton, New Jersey: Princeton University Press.

Wagstaff A. 1991. QALYs and the equity-efficiency tradeoff. Journal of Health Economics 10: 21-41.

Ware J. E., Snow K. K., Kosinski M., and others. 1993. SF-36 health survey: manual and interpretation guide. Boston, MA: The Health Institute, New England Medical Center.

Ware John E. Jr. and Sherbourne Cathy Donald. The MOS 36-item short form health survey (SF-36). Medical Care. 1992 Jun; 30(6).