Chapter-6:

Reliability and validity of health state valuation tools.

In the first section we describe our assessment of reliability of the health state valuations done in this study. An important consideration for estimation of reliability is measurement stability, which is linked to the theoretical conceptualisation of the health state valuation function. We explore the ordinal rank measurements obtained through this study and conjecture that the health state valuation function is a multi valued function. The second section discusses issues regarding validity of these measurements.

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[1]. One implicit assumption in the example given above is that the object of measurement, namely, the quantity of oil being measured, 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[2]. 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 $N(0, \sigma_e^2)$. 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

[1] .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.

[2] 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

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.

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 45^0 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 1994). 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 (1994) 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, 1981).

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.

The feasibility of various test - retest reliability measures for health state valuation instruments can be summarised using a scheme from Van Agt et al (1994) with some further modifications in the light of above discussions (Table-6.1). 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.

Table-6.1: 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 ρ Kendal's τ Weighted Kappa
Group	ANOVA: Classical test theory (ICC)	ANOVA: Generalizability Theory (Generalizability coefficient)	ANOVA: Friedman's by ranks	

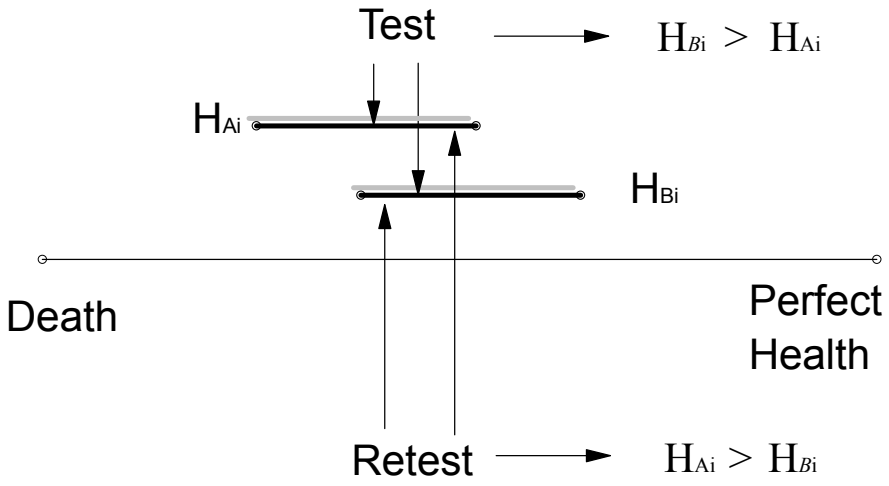
[1] Source: Adapted from van Agt HM; Essink-Bot ML, and Krabbe PFM (1994)

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 $h_i = H_i + e_i$ 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-6.1 : Potential for rank reversal of health states in retesting.

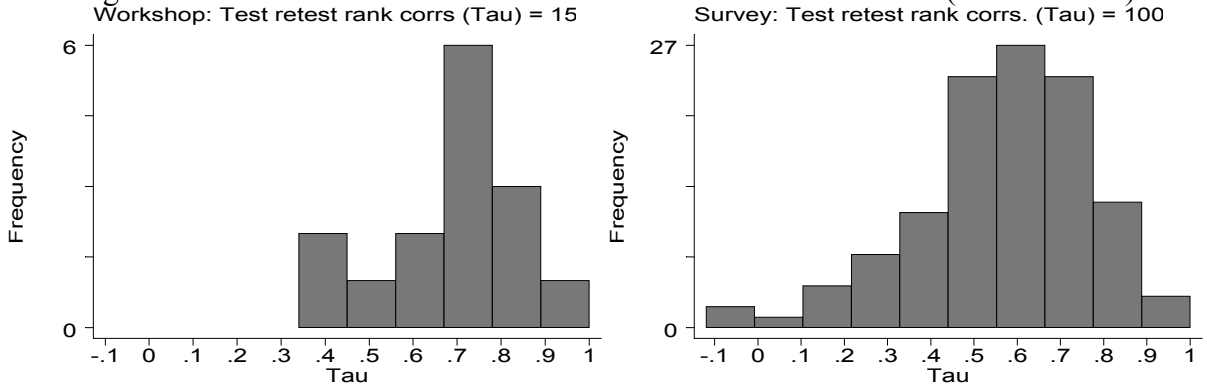


For example, consider health states A, and B with overlapping true value sets as shown in Figure-6.1. Suppose a person with these true value sets is asked to express his / her valuation on an occasion (test). (S)he may report health as being state B as being better than health state A by tapping into his / her true value set in the manner shown in Figure-6.1. On another occasion, (s)he may sample his / her true value set in the manner shown for the retest, in Figure-6.1, 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.

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-6.2 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.

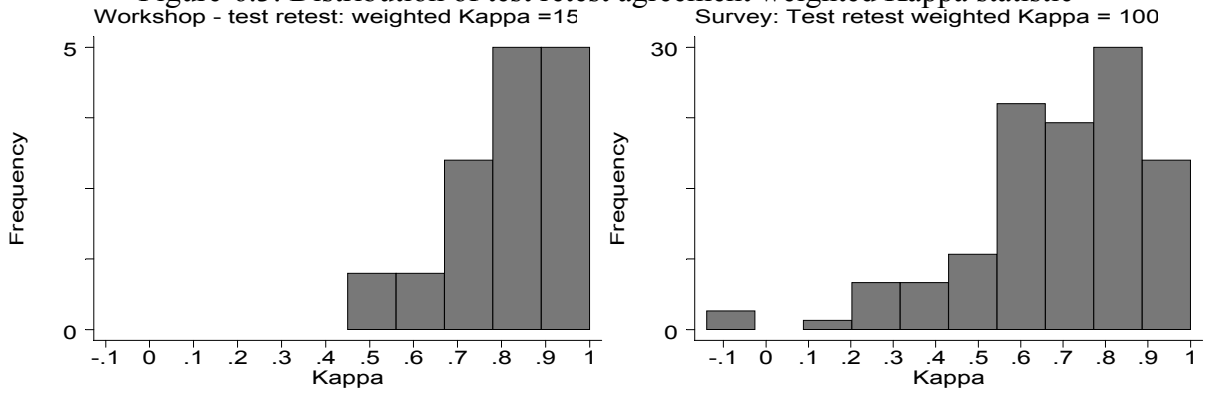
Figure-6.2: Distribution of within valuer test retest rank correlation (Kendal's Tau)



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[3]. Figure-6.3 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.

Figure-6.3: Distribution of test retest agreement weighted Kappa statistic



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

[3] The Kappa statistic is described in Streiner and Norman (1995, pp116-118).

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.

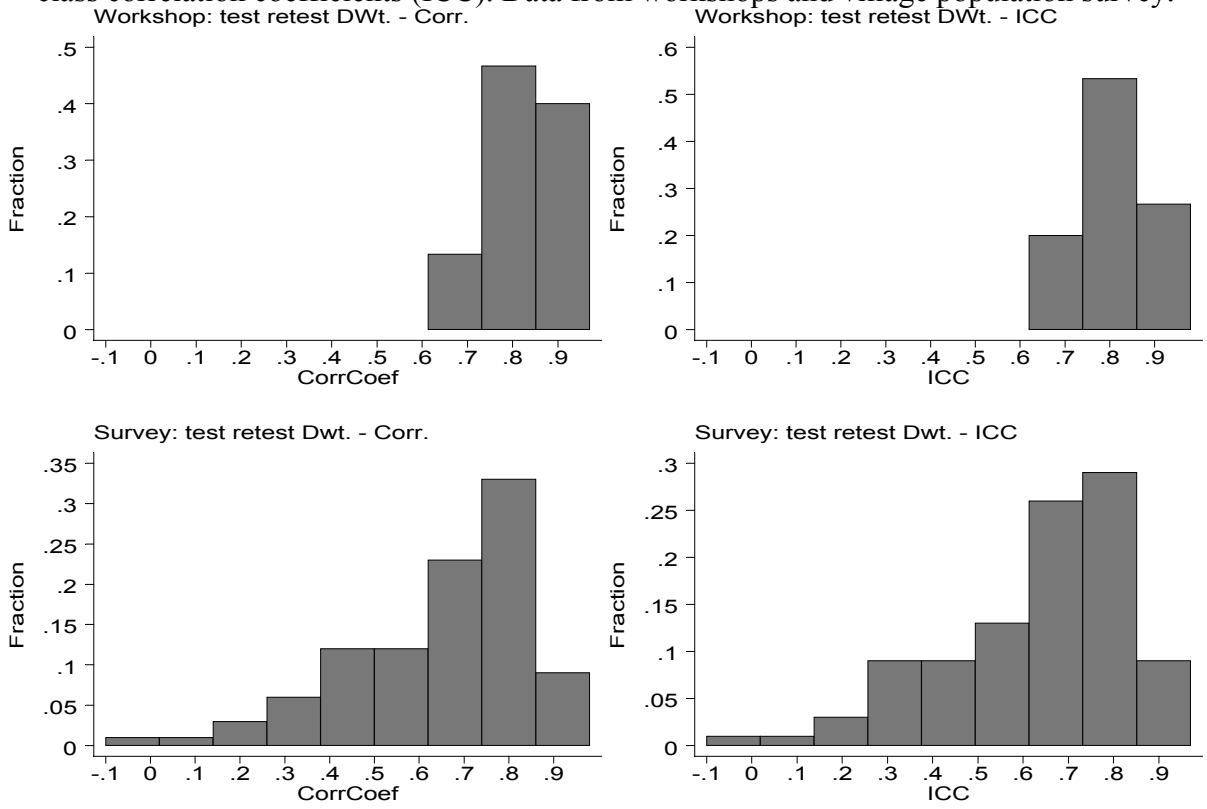
The above findings are consistent with our hypothesis. Certain additional conjectures appear. We find that stability of expressed valuations are somewhat better for the educated group and in case of "Own health" state for both groups. In case of the community survey 85% of persons who ranked their own 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,* 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:

Figure-6.4: Distribution of within valuer test-retest product moment correlation and intra class correlation coefficients (ICC). Data from workshops and village population survey.



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-6.4 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.

Table-6.2: 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.

Intra class correlation coefficients for all health states combined and by each health state ICCs were calculated (Table-6.2) 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.

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-6.2 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-6.2 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 $v \times h \times o$ 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. $MS_v = \frac{abc \sum_{v=1}^{a} (\bar{x}_{v..} - \bar{x})^2}{a-1}$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Table-6.3: 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 svariance 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-6.3. 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-6.4: 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

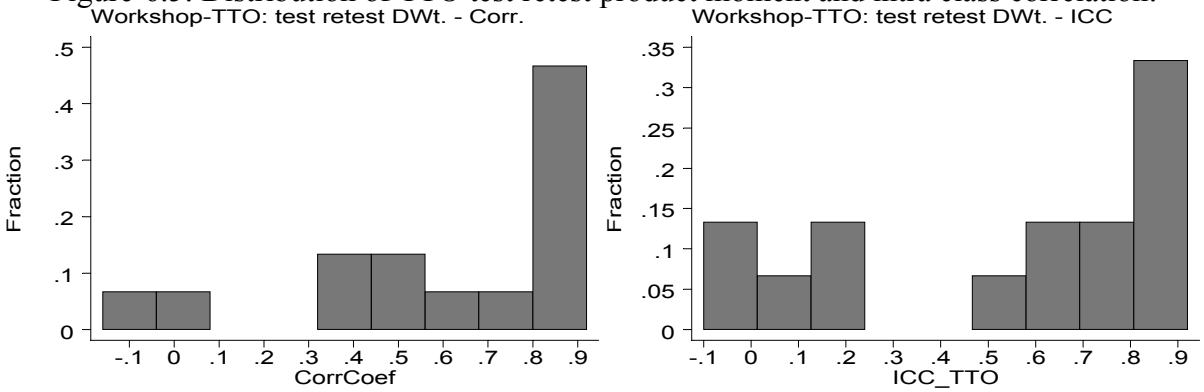
[1] Standard EuroQol instrument. Postal survey in Rotterdam, Netherlands, Jan. 1991.
[2] Ordinal rank consistent VAS. Medical students in Japan, 1999.

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-6.4 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:

Figure-6.5: Distribution of TTO test retest product moment and intra class correlation.



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-6.5 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.

Intra class correlation coefficients were estimated by health state (Table-6.5). 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-6.5: 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.

Table-6.6: Correlation of health state values obtained from different methods and effect of deliberation.

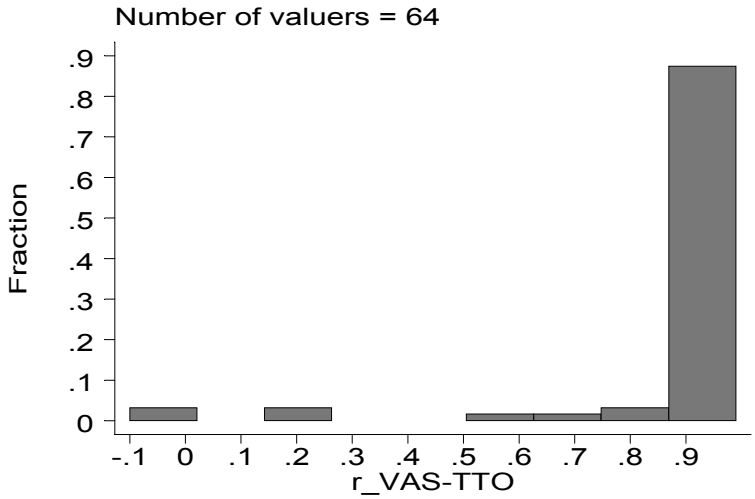
Method ↓→		TTO	PTO
First TTO / PTO valuation			
VAS	All valuations (180 valuers)	0.51 (1969)	PTO1: 0.02 PTO2: 0.03 (319)
	Ordinal rank consistent (162 valuers)	0.52 (1782)	PTO1: 0.47 PTO2: 0.28 (275)
TTO	First iteration (179 valuers)		PTO1: -0.00 PTO2: 0.13 (319)
Last round of TTO / PTO valuation			
	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).

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-6.6 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.

Figure-6.6: Frequency distribution of within valuer correlation between VAS and TTO valuations.



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-6.6 "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.

Table-6.7: 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		

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-6.6 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.

Let us now consider to convergence at the health state level. Table-6.7 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.

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-6.8). 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-6.8 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.

Table-6.8: 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	5	7	11%	34%	10%	22%
	Amputation below the knee(both legs)	433221			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	4	4	2%	14%	0%	0%
	Amputation below the knee (one leg)	322211			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.

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-6.8 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-6.9. 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-6.9: 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

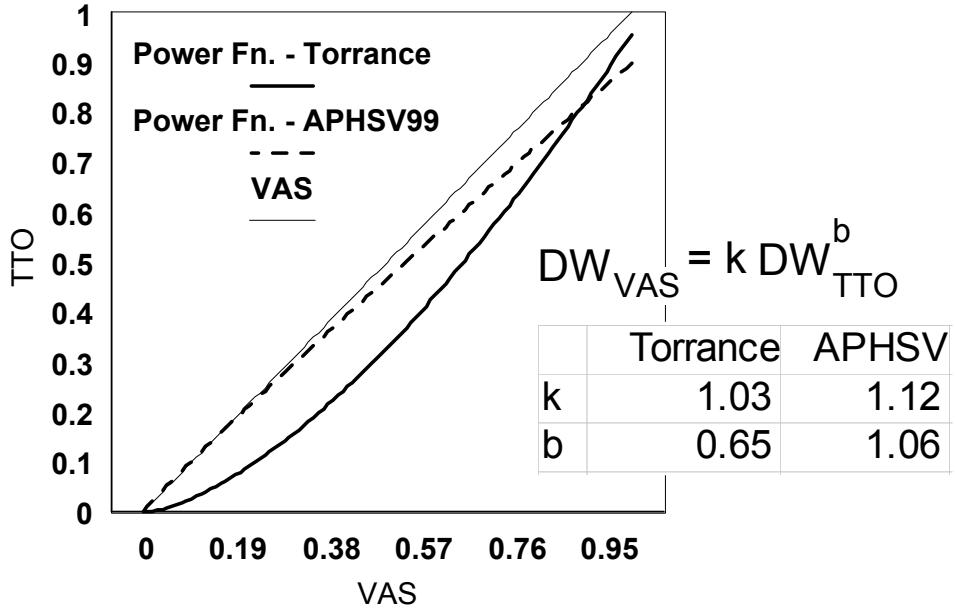
Some researchers (for example, Torrance, 1976) have argued that the valuations obtained through VAS and TTO are not comparable. One hypothesis implied by Torrance (1976) is that the valuations from trade-off techniques reflect the true stimulus in the mind of the valuer. What we measure is the response 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[4]. If we assume that the VAS valuation is the response and trade-off 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; $DW_{VAS} = k\, DW_{TTO}^{b}$.

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

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.

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

Figure-6.7: Power function models of TTO from VAS: Torrance (1976) and APHSV99.



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 $b = 1$.

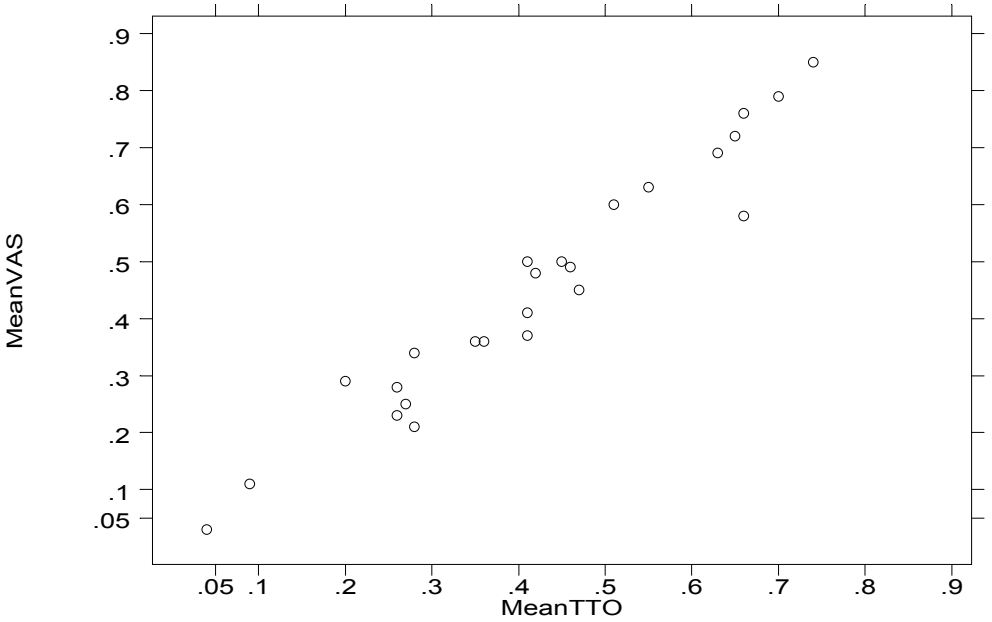
Figure-6.7 shows the plots of TTO disability weights predicted from our VAS-based measurements using the model estimated by us from this study (made up of dashed line), and the model estimated by Torrance (thick continuous line). The thin straight line represents

VAS measurements without any transformation. Clearly, we did not find differences between VAS and TTO measurements, to the extent observed by Torrance (1976) among the Canadian population.

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-6.8: Two way scatter plot of mean disability weights from TTO and VAS



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-6.8 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-6.9. 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-6.10). 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.

Figure-6.9: One-way scatter plot of mean disability weights for different health states obtained by visual scaling (VAS) and time trade-off (TTO).

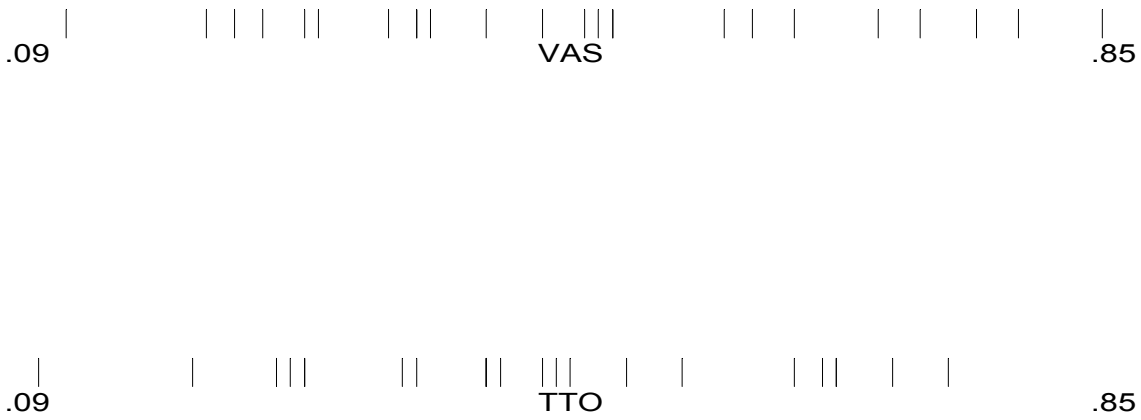
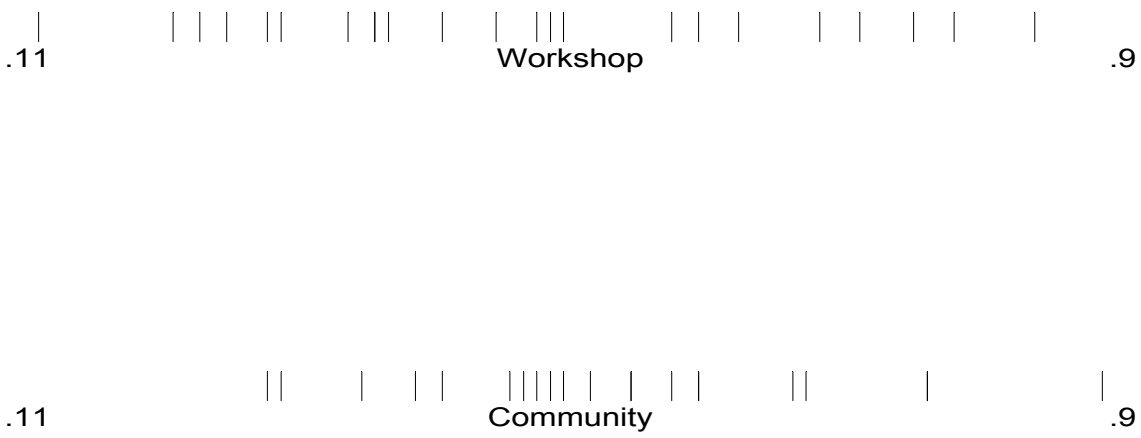


Figure-6.10: One-way scatter plot of mean disability weights for different health states obtained by visual scaling (VAS) from MDHSV Workshops and Community Survey



References:

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.

Bass EB; Steinberg EP; Pitt HA and others. Comparison of the rating scale and the standard gamble in measuring patient preference s for outcomes of gallstone disease. Medical Decision Making. 1994; 14:307-314.

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 RA; Carter WB, and Gilson BS. The Sickness Impact Profile: Development and final revision of a health status measure. Medical Care. 1981; 19:787-805.

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.

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.

Chronbach LJ, Gleser GC, Nanda H. and Rajaratnam N.; The dependability of behavioural 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.

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.

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

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

Krabbe PFM, Essink-Bot ML, Bonsel GJ. The comparability and reliability of five health-state valuation methods. Social Science Medicine, Vol 45, No. 11, pp. 1641-1652. 1997.

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.

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).

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.

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.

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

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.

Streiner D. L. and Norman G. R. Health measurement scales: a practical guide to their development and use. New York: Oxford University Press; 1995.

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.

van Agt HM; Essink-Bot ML, and Krabbe PFM. Test-retest reliability of health state valuations collected with EuroQol questionnaire. Social Science and Medicine. 1994; 39(11):1537-1544.

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.

-------o-------