



Psychometric Validity and Multi Attribute Utility (MAU) Instruments

Professor Jeff Richardson

Foundation Director, Centre for Health Economics
Monash University

November 2010

Centre for Health Economics

ISSN 1833-1173

ISBN 1 921187 56 5

Correspondence:

Professor Jeff Richardson
Centre for Health Economics
Faculty of Business and Economics
Monash University Vic 3800
Australia

Ph: +61 3 9905 0754 Fax: +61 3 9905 8344

Email: Jeffrey.Richardson@monash.edu

ABSTRACT

The concept of validity is widely misunderstood and, particularly, the phrase 'an instrument has been validated'. This has been used to bestow far greater authority upon instruments than is justified by the empirical data. In principle, the concept of 'validity' is straight forward. A valid instrument measures what it purports to measure. In practice, validation is a complex subject. It is discussed in most textbooks concerned with instrument construction but widely ignored in economics. Standard textbooks of economic evaluation neither index psychometric validity nor discuss validity except in the context of epidemiological trials.

The present paper reviews the psychometric concept of validity as a prelude to its main focus which is upon the validity of multi attribute utility (MAU) instruments. The literature suggest that these have been developed with little regard for psychometric validity or even the tests which would suggest that the numbers produced are a valid representation of the 'utility' needed for the construction of Quality Adjusted Life Years (QALYs).

The paper was written for the AQoL website (www.aqol.com.au) and, consequently, it is written in the form of questions and discussion of the questions. The conclusion reached is that more care is needed in the claims made for instruments and particularly with respect to the universality of their validity.

TABLE OF CONTENTS

A. What is Validation	1
Validation in practice.....	3
Validity and parsimony.....	3
Categorising tests of validity	3
B. Some Common Issues.....	5
Face validity and construct validity.....	5
The need for multiple items	5
Does correlation demonstrate validity?	5
Economics versus Psychometrics: The GDP dilemma	7
Decision analytic versus psychometric valuation	10
Combining DA and PI in MAU instruments	11
C. Validating MAU Instruments	11
How is an MAU instrument validated?	11
Necessary and sufficient conditions for QALY validation.....	11
Have MAU instruments been satisfactorily validated?	12
Do MAU instruments correlate highly?	12
Why do MAU instruments have a low correlation?	13
Why is modelling necessary: Have utility models been validated?.....	15
Validating utility (scaling instruments)	15
Can there be negative utilities?	16
D. Validity of the AQoL Instruments	16
Why is another instrument necessary?	16
Have AQoL <i>utility scores</i> been validated?.....	17
Have the AQoL instruments been validated?.....	17
E. Conclusion	18
The need for caution.....	18
References	19

Figures

Figure 1 Concept and item overlap	2
Figure 3 Insensitivity/content invalidity.....	6
Figure 4 Construct validity.....	7

Boxes

Box 1 A common misunderstanding	1
Box 2 Definitions of validation tests	4
Box 3 Empirical results from two 5 instrument comparative studies	13
Box 4 Case Studies of Content Invalidity (see text for explanation).....	14

Psychometric Validity and Multi Attribute Utility (MAU) Instruments

A. What is Validation

In principle, the concept of 'validity' is straight forward. What is validity? A valid instrument measures what it purports to measure. A correctly calibrated ruler, for example, gives valid measurement of distance.

In practice, validation is a complex subject. It is discussed in most textbooks concerned with instrument construction. For a simple but comprehensive explanation see the excellent web-based notes by William Trochim [<http://www.socialresearchmethods.net/kb/measval.php>]

The present notes focus primarily upon validity of health related multi attribute utility (MAU) instruments; that is, evaluative questionnaires containing multiple attributes (dimensions/domains), where answers are combined to produce an index of the strength of preferences (utility). This context adds two additional layers of complexity. The first is the breadth of the concept. Health is multi dimensional and overall validity requires validity in all dimensions. Secondly, to qualify as 'utility', in the sense used by health economists, additional properties are required to those usually discussed in the psychometrics literature.

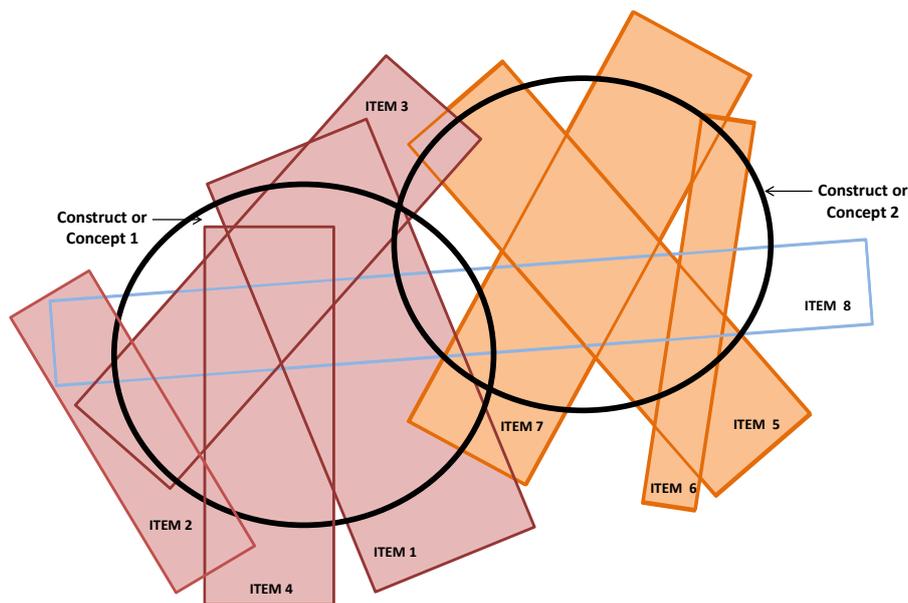
Box 1 A common misunderstanding

The concept of validity is widely misunderstood and, in particular, the phrase 'an instrument has been validated'. Many are misled by the compelling connotations of the word 'validated'. The term implies a generality and finality which is incorrect. In contrast with the connotations of the term 'validity', the property we seek is more aptly described as (degrees of) confidence in an instrument in a particular context, rather than a universal 'true-false' stamp.

With psychological constructs such as intelligence (IQ) or quality of life (QoL), establishing validity is problematical as there is no 'gold standard' as there is for physical measurement (see 'What is a 'gold standard'). A concept such as intelligence is commonly the result of a number of elements: verbal, numerical, spatial skills, problem solving, memory, etc. In turn, each of these may not be clearly identified by the answer to a simple question, but may require a series of questions and answers. Further, the precise meaning of terms and questions can vary between individuals and cultures in a way which is related to personal circumstances. As an example 'communication' may mean speaking to some, signing to others, face to face contact for some or texting for others. Happiness may be primarily dependent upon social relationships in a culture which is community oriented (Asia) but self represented in individualistic (western) culture.

To overcome this problem psychometric 'Classical Test Theory' uses some variant of factor analysis to create measurement instruments.¹ Answers to questions are analysed for their relationship, and answers which cluster around a concept – the answers correlate – are accepted as a measure of this concept. This is illustrated in Figure 1 in which two constructs or concepts are represented in 'content space' by the heavy bold circles. A series of questions and answers – items – are represented by the various rectangles. As shown, four of these heavily overlap Concept 1. Three overlap Concept 2. Item 8 crosses both concepts. In the terminology of factor analysis this last item 'cross loads' on the two concepts and would normally be eliminated from the items used in an instrument.

Figure 1 Concept and item overlap



Key:

- 'Item' = question with a series of possible response levels (eg How often do you feel sad?
a) never, b) rarely, c) some of the time, d) usually, e) nearly all the time)
- Concept = an abstract idea concerning some hypothesised attribute or characteristic
(physical fitness, mental health)
- Construct = A mini theory or created construct to explain observed behaviour.

Figure 1 illustrates a number of points. Concepts overlap. Statements overlap and do not exactly correspond with concepts. Importantly, single statements may cover only a small part of the content of a concept, ie language and concepts are imperfectly related. Finally, as shown, neither concept may be perfectly defined by the items. Some content may be omitted by the item description.

¹ It is necessary to distinguish CTT (Classical Test Theory) from IRT (Item Response Theory). CTT has recently used some variant of factor analysis to create measurement instruments. Historically, the foundational measured concepts (and strategies) for CTT were the item-total (or the item-remainder) correlations (that provided evidence of item discrimination) and a measure of internal consistency (eg Cronbach's alpha). In contrast, the 'foundational' measurement concept for IRT is the item response curve. IRT seeks items that relate to one single latent trait that satisfy the criterion of 'conditional independence' – no association between items over and above that explained by the one latent variable.

Validation in practice

Validation is a process of hypothesis testing: increasing the confidence we have in a scale and confidence in the inferences drawn from it. This implies that an instrument is never (fully) validated: we have more or less confidence in it. Importantly, the more demanding the test the greater the confidence. The less demanding the test the less the confidence.

In the absence of a gold standard, validation of a construct (construct validity) usually refers to content, concurrent or predictive validity. As described by Streiner and Norman (2003):

“A measure that includes a more representative sample of the target behaviour lends itself to more accurate inferences; that is inferences which hold true under a wider range of circumstances. If there are important aspects of the outcome that are missed by the scale, then we are likely to make some inferences which will prove to be wrong; our inferences (not the instrument) are invalid” (page 175).

This implies that validation in Context A does not necessarily mean validation in Context B. A test of hearing amongst the general population may capture the ability to communicate. But the same test would be invalid in the deaf community which uses signing.

Validity and parsimony

Brevity is not always a virtue. Adopting a ‘parsimonious instrument’ which attempts to measure both dimensions in Figure 1 with a single item – item 7 – will often result in data which correlate with the true value of both dimensions. However the correlation does not indicate content validity (see ‘Does Correlation Demonstrate Validity’). Outcomes may occur which do not affect item 7. The same conclusion is true within a dimension. Item 6 may accurately measure some changes within dimension 2 but not those described by item 4. (For example, the ability to run 100 metres does not measure dexterity, flexibility or physical endurance all of which contribute to the concept of physical fitness.) Similarly the absence of depression does not indicate good mental health as this concept includes self esteem, good life satisfaction, normal anger, good sleep and other sub-dimensions which are imperfectly correlated. Following from this it is likely that the more questions used, the greater the probability of fully capturing a concept but, as shown in Figure 1, this will result in significant overlap and ‘structural redundancy’. In the context of MAU instruments this will result in the double counting of disutility.

This implies that, in practice, a parsimonious instrument that is valid in one context for one group may be invalid in another context for another group. This will result in a systematic bias in economic evaluation favouring interventions where the instrument captures more of the content. Thus, if service 1 improved the facet of QoL described by item 1 in Figure 1, and service 2 improved the facet described by item 3, an instrument which equated construct 1 (the left circle) with only item 1 would wrongly favour service 1 over service 2 in an evaluation study using the instrument.

Categorising tests of validity

The types of validation tests have been categorised under different headings. These include face validity, content, predictive, convergent and discriminant validity (skeletal definitions are given in Box 2). These tests are commonly classified further. Face and content validity are often described as a sub-set of construct validity (literally the validity of the psychological construct). Predictive and concurrent validity are sometimes classified as ‘criterion related validity’. For detailed discussion see any relevant text or William Trochim’s website

[\[http://www.socialresearchmethods.net/kb/measval.php\]](http://www.socialresearchmethods.net/kb/measval.php)

Box 2 Tests of Validity

'Validity' would be better described as 'degrees of confidence'. The labels described below do not indicate different endpoints but are derived from the different tests or reasons which have been used to increase confidence that the inferences made from an instrument score will be correct. Restated, each of the types of validity provides a different reason why we should (or should not) have confidence in the conclusions drawn from an instrument score.

Types of validity:

- Translation or representation validity
 - Face validity
 - Content validity
- Construct validity
 - Convergent validity
 - Discriminant validity
 - Discriminant (extreme group) validity
- Criterion validity
 - Concurrent validity
 - Predictive validity

Translation or representation validity: A general term for the extent to which a construct (concept) can be successfully translated into, or represented by, specific tests.

Face validity: The instrument seems, at face value, to capture the construct, for example, by naming it. This is generally considered the weakest form of test. For example, asking 'are you good at arithmetic' with a response scale from very good to very bad is not likely to produce an accurate scale.

Content validity: The extent to which an instrument includes or covers a representative sample of the construct's behaviour domain, for example, determining arithmetic skill by asking for the answers to 3-4 questions for each domain of arithmetic – addition, subtraction, multiplication, division, fractions, decimals, etc. Examples of (context specific) content invalidity are given in two case studies in Box 4 'Why do MAU instruments have a low correlation'.

Construct validity: A general term for the success of a test or instrument in measuring a construct (concept). It commonly subsumes the convergent and discriminant validity.

Convergent validity: A specific test of construct or criterion validity. Instrument scores correlated, as predicted, with other instrument scores or some criterion score which are known to correlate with the construct.

Discriminant validity: Another specific test of content or criterion validity. Instrument scores do not correlate with instrument scores unrelated to the construct. For example QoL EQ5D scores would be expected to have low or no correlation with a person's blood pressure.

Discriminative validity: An instrument can distinguish between groups as expected, eg patients and the public.

Criterion validity: A general term for the use of some external criterion to test the concept.

Concurrent validity: An instrument correlates with a criterion variable or instrument, ie one known to measure the construct.

Predictive validity: The ability to predict what is expected. This includes the predictive tests above but is more general. For example an IQ test may predict subsequent income.

Strength of a test: As a rule, the higher the prior likelihood of a test's success the lower the strength of the test (ie there should be less confidence in the *theory* that the instrument has universal application). In physics, the observation of objects falling to the ground should give little support to the Theory of Relativity however often it is repeated even though the outcome is predicted. However the previously unexpected observation of light rays bending in a gravitational field led to overwhelming confidence in the Theory. Similarly the correlation of two general QoL scales is unsurprising. In contrast, the prediction of real people's choices between real health

programs on the basis of the measured utilities of health states would justify greater confidence in the scale especially if the prediction could not be made without using the utility scores.

Observing patients selecting treatments with lower life expectancy but better health in accordance with utility scores would be highly persuasive evidence that the scores represented the 'utility' of economic theory.

B. Some Common Issues

Face validity and construct validity

Face validity is the assessment of whether or not a test seems OK. Or (subjectively) whether it 'should be valid', possibly as a result of an analysis of the meaning of the words used. Questions of the form 'how good are you at arithmetic' or 'do you suffer from depression' represent weak tests of validity. Answers may be correct but they may easily be wrong if people only partially understand the implications of the concept or (consciously or unconsciously) deny or exaggerate the truth. A better test of depression would ask unthreatening questions about behaviour which covered each of the major symptoms of depression – sadness, pessimism, guilt, tiredness, lack of energy, poor sleep, etc. This would increase confidence in content validity. The Beck Depression Inventory asks 21 such questions without mentioning 'depression'. The McGill Pain Inventory isolates 20 possible facets of pain quality also without mentioning pain, its intensity or duration.

The need for multiple items

Most broad concepts relating to health have multiple sub-domains and the need for multiple questions may follow from this if content validity is to be achieved. The use of simple questions to measure a broad dimension may have face validity. Face validity is, however, unlikely to ensure content validity (see 'Face validity' and 'Construct validity').

Does correlation demonstrate validity?

Validity has been largely tested in the literature by convergent validity – a correlation between an instrument and a second 'validated' instrument. This is necessary but not sufficient for a satisfactory instrument (even if the second instrument is truly valid). An instrument with poor construct validity may simultaneously correlate highly with another (criterion) instrument. Two cases are illustrated below.

Insensitivity: In the first case, insensitivity (strictly a subset of content invalidity) may occur when there are insufficient items in an instrument or too few categories to fully capture a dimension affect. In Figure 2 true utility (True U) on the vertical axis will be recorded by instrument Z on the horizontal axis as 0.0 until it reaches a 'switch point' of 'a' when the average recorded response becomes 'a¹'. True utility must rise to 'b' before recorded utility switches to b¹, etc. As a result, a program which increases true utility by an amount 'A' will record no change on instrument Z. Conversely, a program increasing utility by a smaller amount, B, will result in a large recorded increase in utility from 'b¹' to 'c¹'.

Content invalidity: Figure 3 again depicts real utility on the vertical and measured utility on the horizontal axes. Points shown would result in a high positive correlation between them. However the omission of items or dimensions from instrument Z may result in a cluster of points, A, where

the dimensions of instrument Z improve but the effect is more than offset by the negative effect from the omitted items/dimensions.

To illustrate these figures consider a mobility instrument consisting of three questions.

- (i) I have no pain;
- (ii) I have mild pain; and
- (iii) I have extreme pain.

This instrument would produce a significant correlation with a validated pain instrument as the two responses would each attract the patients closest to it. Despite the high correlation the item clearly lacks sensitivity to gradations and types of pain (content invalidity).

The instrument could be 'validated' against a more detailed instrument such as the McGill Pain Questionnaire; that is, it would correlate highly. Suppose, however, this instrument was used to measure the QoL of two people, one of whom was receiving pain reducing medication with side effects omitted for the overall instrument (eg loss of vitality). As in Figure 2 a measured increase in utility in the mid range attributable to the medication might correspond with a decrease in true utility due to a loss of vitality. The 'validated' instrument, however, would produce results suggesting the opposite conclusion as a result of its content invalidity in the context of this evaluation.

Figure 2 Insensitivity/content invalidity

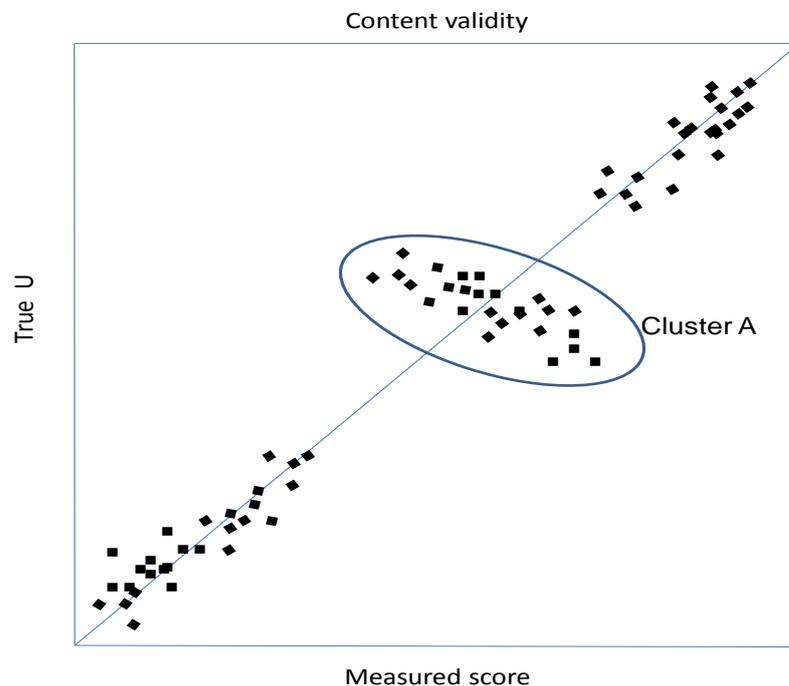
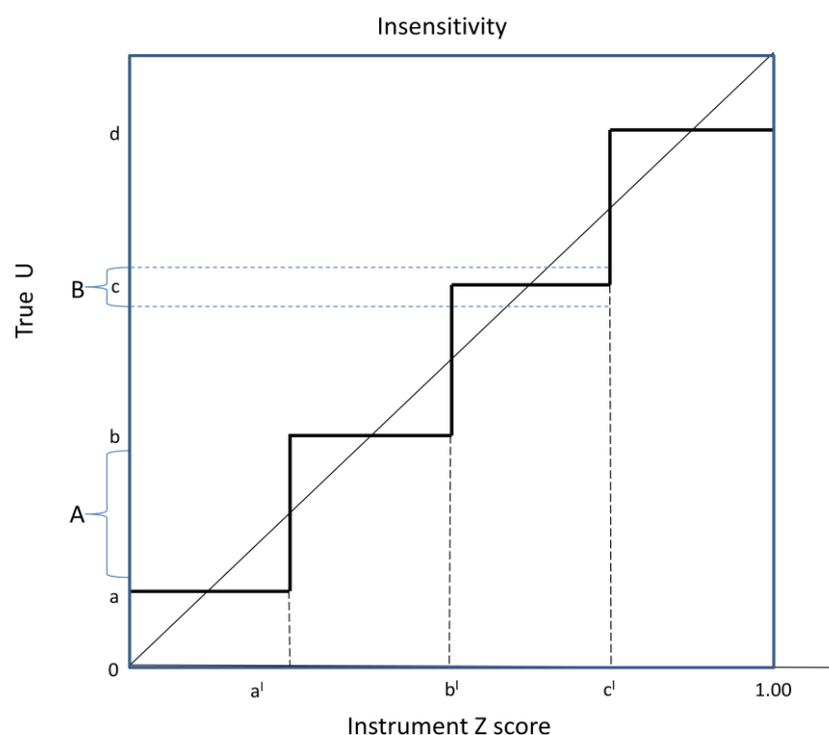


Figure 3 Construct validity



Economics versus Psychometrics: The GDP dilemma

(Formative versus Reflective Modelling)

Economics and (much of) psychology follow different measurement traditions reflecting the different problems they tackle and the different types of data available. The two disciplines overlap in the area of MAU instrument construction and this is potentially a source of confusion.

Economic methods are typified by the measurement of GDP. The plethora of goods and services which are observed flowing through the economy have little in common, ranging from services by insurance companies, hotels, doctors and governments to consumer durables such as refrigerators, TVs and cars. There is no common concept behind these except their purpose which is to satisfy people's wants. This might suggest that these goods and services could not be aggregated, a view captured in the adage 'you can't add apples and oranges'. But the goods and services are not added. Rather it is the *value of the goods and services* that is added where value is conceptually how much people would be willing to pay for the goods and services or, in practice, the market price they actually pay. (The two are not the same as the market price does not include the so called 'consumer surplus'.)

There is a direct parallel between this and the valuation of health states using MAU instruments. Different attributes of QoL – physical, social and mental wellbeing – are analogous to the goods and services of GDP. They are the source of QoL. To aggregate these dissimilar attributes requires something analogous to the market price and this is provided by the utility weights obtained by the scaling instruments (TTO, SG, etc). Just as the value of GDP is the sum of price times volume of products, so the value or total utility of a health state is the product of attributes and the utility of the attributes. Valid MAU instruments however have an extra layer of complexity, discussed below.

Psychology: In contrast with the methods described above, psychometrics commences with the fact that the attributes of QoL are not directly observed. They are 'constructs' which we infer from other behaviours. Since behaviours vary enormously and our vocabulary has adapted to this, a single construct subsumes a very wide range of behaviours and a correspondingly wide range of linguistic descriptions. For example, the 'simple' construct of 'happiness' can subsume the 'sub-constructs' of delight, comfort, exuberance, joy, ecstasy, pleasure, zest, excitement, optimism, hope, faith, trust, confidence, satisfaction, fulfilment, pride and serenity. To complicate matters further, these sub-constructs are not generally separate from one another as is the case with the goods and services of the GDP. People who are experiencing joy may be experiencing some excitement. They may or may not be feeling confident and serene. See Figure 1 in 'What is validation'.

Over many years psychometricians have developed methods for bringing coherence into this confusion of overlapping concepts. After a concept has been postulated (theory) a large array of items are collected describing its many manifestations. Each of these allows a person to indicate the extent to which the item applies to themselves (using, for example, a scale from 'strongly agree' to 'strongly disagree'). Statistical analysis allows the items to be grouped into clusters of items – where people give very similar answers (eg cluster 1 may be joy, ecstasy, zest, excitement; cluster 2 may be optimism, hope, confidence). The analysis identifies items which best describe the cluster and these become the basis for scoring the cluster and construct. If the process is carried out well the result will give confidence that that description of the concept has 'construct validity'.

The 'GDP dilemma' is that there appears to be two ways of deriving an instrument. The first follows the classical psychometric procedures evolved by psychologists for instrument construction. The second follows the (decision analytic) tradition incorporated in GDP measurement in which importance weights are external.

Formative and Reflective Models: The two traditions are closely related to the distinction between *formative* and *reflective* modelling. In the formative model the elements define the concept. The goods and services produced (or their value) are combined to *form* the concept of GDP. Similarly, people's education, income and occupation are combined to *form* the concept of social economic status. In contrast, in a reflective model the elements *reflect* the underlying concept. A person laughing, joking and smiling *reflect* the concept or trait of happiness. Problem solving, memory and verbal skills *reflect* the concept or trait of intelligence. In principle, in a formative model causation runs from the elements to the concept they create. In a reflective model causation runs from the underlying concept or trait to the elements which reflect the concept or trait.

MAU modelling: In the context of MAU instrument development both traditions and models discussed above have shortcomings. The implicit assumption in the economics tradition that the contents of constructs are visible or at least unproblematical like the goods and services of the economy is clearly wrong. Multi-clustered (multi-nuanced) constructs (concepts) cannot generally be described by a single item and the manner in which items are selected cannot be reliably determined by logical (ie non-empirical methods).

Reflective modelling is important to measure the concepts of mental, physical and social health.

However the classical psychometrics approach similarly has shortcomings. If a statistical analysis does not show some relationship (correlation) between a cluster of items and other clusters then it will be excluded from the concept. But this may result in a concept of 'health' which, for

example, excludes oral health despite the fact that oral health is part of what we may wish to include in our measurement of 'health'. The economics perspective has the attraction that our purpose or objective will 'trump' the statistics, not vice versa. If our purpose is to obtain the value of 'quality of life' and we wish this to include oral health then the GDP approach has the attraction that it would ignore the classical reflective results and simply include and weight oral health (independently measured) in the same way as other dimensions were weighted and included; that is, formative modelling would be used.

However this latter approach conflates the questions of content validity of a particular dimension with the overall objective (in the latter case, the measurement of a broad concept which includes oral health). The issue of content validity cannot be circumvented and it is always pertinent to ask whether any given instrument 'fully captures the content of the concept'. If it does not, as described earlier, the data produced by the instrument may sometimes correlate with a sub-set of other relevant instruments (convergent validity) but not with others which are sensitive to the omitted elements of the construct.

In very broad terms, the construction of dimensions for an MAU instrument should use the reflective modelling of classical psychometrics and the combination of dimensions into an overall concept of QoL should use the formative modelling of economics/decision analysis. There is a *relatively* clear concept of the dimensions of mental, physical health, etc and a less clear concept of overall quality of life. Should it, for example, include oral and social health, spirituality, etc?

However the issues are blurred. To an extent, each of the dimensions is also formative: core behaviours give rise to the concepts of mental and physical health and cannot be omitted. Conversely, formative models will not bring together dimensions which have no common concept (*albeit* a vague one). Behind the concept of GDP is the concept of a net increase in material wellbeing. The questions 'what is quality of life' and 'what is health' may invite vague and varying answers but they are not meaningless as would be, for example, a constructed concept composed of shoes and ships and ceiling wax, cabbages and kings.

In practise, this implies a degree of flexibility in the modelling of both dimensions and the overall concept. On the basis of professional advice, for example, a particular element may be inserted in the mental health dimension (or sub-dimension) because it is essential to the concept irrespective of its correlation with other elements. Similarly, the choice of dimensions for combining into an overall concept of QoL will be influenced by the result of focus groups and other input which reflects the common conception of the dimensions included. Fortunately, this analytical flexibility is facilitated by Structural Equation Modelling (SEM).

Despite this, there is an unavoidable (and largely ignored) issue of what should be the scope of the content of the concept 'quality of life'. The answer necessarily depends upon the context: what is the question and what is the problem to be solved? In economic evaluation we wish to create quality adjusted life years (QALYs) and, as usually conceived, it is the individual who decides for themselves what is important for themselves. This implies the widest possible concept and the inclusion of any element affected by a health program. More restrictive concepts are, however, possible.

In summary, content validity is an important part of the 'nomological web' of evidence which should be the basis for selecting an instrument. It remains true that a purely numbers driven psychometric instrument may omit a desired dimension (oral health in the example above). But this does not justify the neglect of content validity for the included dimensions. Rather, the desired dimensions must be forced into the instrument while seeking content validity for each dimension.

With at least some correlation between dimensions such a reflective/formative model may be achieved with SEM – the restriction is that the higher-order latent variable must also be ‘causing’ at least two outcomes in the model. If there is no correlation then the final instrument must employ the methods of economics to quantify and ‘add on’ to the remainder of the construct what is wanted. In this case there will be no coherent psychological construct corresponding with the numbers produced by the instrument but – as in the case of GDP – it may be quite meaningful.

Decision analytic versus psychometric valuation

A third tradition – decision analysis – is also closely related to MAU instrument construction and sits between the economics and psychometrics traditions.

The decision analytic (DA) and psychometric approaches commonly apply in different contexts but have parallel structures. Multi attribute utility modelling can (should) bring these together but, to date, this has only occurred with the AQoL suite of instruments.

Objectives

DA: The objective of DA is to guide decision makers by determining the value (utility) of each of the multiple options available. It does so by breaking down the final decision into the multiple attributes (dimensions) which make up the decision. The choice of a location for a factory, for example, depends upon the local cost of labour and capital and the distance to the market.

PI: Psychometric instruments are used to place a value (‘utility’) upon psychological states. They do so by breaking down and valuing the dimensions/elements of the state. The utility of different states guides decision making.

Elements

DA: The elements which combine to determine the options are often objective and generally orthogonal – unrelated to each other. The options for the location of a car factory, in the example above, depend upon largely independent attributes – distance to the market, unit cost and productivity of capital, unit cost and productivity of labour, local entrepreneurial skill, industrial relations, political conditions, etc. Decision analysts strive to achieve a decision model with such orthogonal elements.

PI: The concept to be measured is a ‘construct’ or ‘concept’ not an objective, tangible entity. Sub-dimensions of the concept may similarly be constructs. They are abstractions from multiple observations or ‘manifest items’. ‘Arithmetic skill’, for example, cannot be accurately determined by a single question but is inferred from a variety of questions relating to different arithmetic procedures (see ‘Face Validity’ and ‘The Need for Multiple Items’).

Combining Elements

DA: Importance – ‘utility’ – weights are employed to model the overall score. In the simplest case an additive model uses weights summing to unity.

PI: In principle, variable weights could be used to sum elements as in DA. In practice, they are seldom employed and scores are calculated using the same ‘unitary’ weights for all responses (which is described as ‘unweighted’). For example, with 5 ranked response

categories where best = 5 and worst = 1, a person ticking the third response category would have '3' added to their 'utility' (score).

Combining DA and PI in MAU instruments

MAU instruments are, in effect, a special application of DA methods aimed at guiding decisions by determining the utility of each of the multiple health states. The distinguishing feature of MAU instruments is that the dimensions of health are generally not independent and the dimensions are not observed and objective, but are constructs. This has three implications:

- (i) The dimensions should have construct validity;
- (ii) The overlapping dimensions – correlation – implies there is an underlying concept which may be labelled HR-QoL but the dimension structure which best defines this must be determined psychometrically;
- (iii) When dimensions overlap the achievement of content validity through the inclusion of multiple items is likely to also result in 'redundancy' – double counting of some elements of health. To be valid utility scores for an economic evaluation the effects of redundancy must be removed.

C. Validating MAU Instruments

How is an MAU instrument validated?

Validation is a complex issue (see 'What is Validation'). The validity of an MAU instrument requires four types of *achievable* evidence. These relate to:

- (i) The instrument 'measurement model' (or descriptive system) . Importantly, content validity may be context specific;
- (ii) The instrument model: does the model which combines items achieve criterion validity: does it result in the same utility prediction as the holistic measurement of the same health state using the same utility scaling instrument;
- (iii) The measurement of utility: does the scaling instrument (SG, TTO, etc) measure what we want to measure (see 'Validating utility');
- (iv) The instrument utility scores: does the instrument produce scores with predictive and convergent validity (correlate with other 'validated' scales)? Again, the answer may be context specific as with the scores from other QoL instruments.

Necessary and sufficient conditions for QALY validation

To achieve gold standard validation of an MAU instrument each of the necessary conditions above must be met. However, to interpret the final numbers as utilities which are suitable for calculating QALYs also requires the achievement of the 2 conditions below. Each is an arithmetic consequence of the definition of a QALY as 'QALYs = (length of life)x(utility of life) where length may be discounted for time preference.

- (v) Evidence of a scale 'interval property' *with respect to preferences*. This means for example that any 0.2 interval along the scale (eg 0.3-0.5; or 0.6-0.8) would have the same meaning with respect to the preferred trade-off between life and death;

-
- (vi) The strong interval property. Percentage changes in utilities from an MAU instrument should correspond with the percentage changes in the expected length of life in decisions actually made when a trade-off is possible between the length and quality of life.

For a discussion of the strong interval property see Richardson, Working Paper 5 (1990) *Cost Utility Analysis: What should be measured* (also published in *Social Science and Medicine* (1994)).

For practical reasons evidence with respect to the last two criteria would be hard, though not impossible to obtain. The criterion remains virtually unmentioned in the literature. A review of this reveals that, to date, the majority of validation studies are concerned with convergent validity, ie the correlation of instrument scores with non utility, disease specific instruments or correlation with other utility instruments which have been similarly validated! There has been little attention given to the other requirements above. The AQL is the only instrument whose descriptive system was constructed using the psychometric principles of instrument construction designed to obtain content validity.

Have MAU instruments been satisfactorily validated?

See 'How is an MAU instrument validated'.

In sum, this involves 4 separate issues:

- The validity of the utility measurement or scaling technique (TTO, SG, PTO, RS) ; and
- The validity of questionnaires or descriptive systems. See 'Validation'
- The validity of the model which combines items
- Evidence that final instrument scores have an interval and strong interval property.

The claim that an instrument 'has been validated' generally rests upon context specific tests of the second property. However correlations found in these studies do not represent strong evidence of general validity (see 'Does correlation demonstrate validity'). The low correlation between the major MAU scales in use demonstrates that some or most or all lack content validity across all contexts.

Do MAU instruments correlate highly?

MAU instruments have been validated primarily by correlating them with other disease specific or generic QoL instruments (convergent validity) or demonstrating that they discriminate between populations of well and ill persons. These are weak forms of content validation except in the specific context for which the other instrument has been validated (see 'Does correlation demonstrate validity'). Even in this context the test is necessary but not sufficient for validation as a QALY (see 'Necessary and sufficient conditions for QALY validation').

Evidence indicates that the necessary condition of convergent validity is not well met by existing MAU scales. Results in Box 3 indicate that, in the only two 5 instrument studies conducted to date, instruments explained only 41 to 64 percent of other instruments variance (R^2). This is much less than would be expected for instruments measuring the same property. Two scales measuring weight would not be considered valid if the correlation between them was as low as indicated in these tables.

Box 3 Empirical results from two 5 instrument comparative studies

Hawthorne, Richardson and Day (2001)

Five instruments were administered to 396 members of the community, 334 hospital outpatients and 226 hospital inpatients. Response rates were 58, 43 and 68 percent respectively. The percentage of each instrument's variation explained by each of the other instruments (R^2) is shown in Table 1.

Proportion of variance explained by another instrument (R^2): Australia

	AQoL-4D	EQ5D	HUI 3	15D	SF-6D
AQoL-4D	1.00				
EQ5D	0.53	1.00			
HUI 3	0.55	0.41	1.00		
15D	0.64	0.58	0.55	1.00	
SF6D	0.55	0.56	0.44	0.59	1.00
MEAN	0.57	0.52	0.49	0.59	0.53

R^2 = correlation coefficient squared

Source: Hawthorne et al. (2001) p369

Fryback, Palta, et al (2010)

Data were obtained from the National Health Measurement Study from 3,844 US adults for 5 instruments. Results below were derived from reported correlations. The HUI 2 is omitted as the HUI 3 obtained similar or higher correlations with other instruments.

Proportion of variance explained by another instrument (R^2): USA

	EQ5D	HUI 3	QWB SA	SF6D
EQ5D	1.00			
HUI 3	0.49	1.00		
QWB SA	0.41	0.45	1.00	
SF6D	0.50	0.52	0.43	1.00
MEAN	0.47	0.49	0.43	0.48

Source: Fryback, Palta et al. (2010) p4

Why do MAU instruments have a low correlation?

The construction of existing MAU instruments differs in virtually all respects. The economics literature has focused upon the importance of scaling instruments – the choice of TTO, SG, PTO or VAS for assigning utilities. However it has never been shown that these would cause the discrepancies in utility scores observed between instruments. Indeed, preliminary work with AQoL-8D demonstrates a much higher correlation between scores from the unweighted AQoL and its utility scores than have been obtained by correlating different MAU instruments. This suggests that differences are attributable more to the content of the descriptive systems (content validity) than to the method of evaluating utilities. Two supporting examples are given in Case Studies 1 and 2 below.

Box 4 Case studies of content invalidity (see text for explanation)

Case Study 1

Health dimension	AQoL-4D	15D
Physical health and mobility	<ul style="list-style-type: none"> • Gets around home/community without difficulty • Has some difficulty focussing • Hears normally 	<ul style="list-style-type: none"> • Walks normally, slight difficulty • Cannot read text; can see to walk • Hears normally • Shortness of breath on exertion • Eats normally • Serious bowel/bladder problems
Activities of daily living	<ul style="list-style-type: none"> • Needs no help with household tasks or personal care 	<ul style="list-style-type: none"> • Performs usual activities without difficulty
Bodily pain, general health	<ul style="list-style-type: none"> • Moderate pain, occasionally disturbing activities 	<ul style="list-style-type: none"> • Severe physical discomfort/pain • Has great problems with sleeping • Feels very weary
Social function	<ul style="list-style-type: none"> • Has no close warm relationships • Has friends and is not lonely • Some parts of the family role affected by health • No difficulty communicating 	<ul style="list-style-type: none"> • Speaks normally • Sexual activity almost impossible
Emotional and mental health	<ul style="list-style-type: none"> • Moderately anxious worried or depressed 	<ul style="list-style-type: none"> • Feels extremely sad and anxious • Slight difficulties with thanking and memory
UTILITY	0.49 (->0.14 when family and relationships included)	0.55 (-> no change because family and relationships not included)

Case Study 2

Health dimension	HUI-3	EQ5D
Physical health and mobility	<ul style="list-style-type: none"> • Walks without difficulty • Full use of hands and fingers • Unable to see well even with glasses • Some hearing difficulty 	<ul style="list-style-type: none"> • No problems walking around
Activities of daily living	<ul style="list-style-type: none"> • Bathes, eats and dresses normally 	<ul style="list-style-type: none"> • No problems with personal care • No problems performing usual activities
Bodily pain, general health	<ul style="list-style-type: none"> • Moderate pain, occasionally disturbing activities • Health rated as fair 	<ul style="list-style-type: none"> • Moderate pain or discomfort
Social function	<ul style="list-style-type: none"> • No problems with communicating 	
Emotional and mental health	<ul style="list-style-type: none"> • Occasionally fretful, angry or depressed • Somewhat forgetful, but able to think clearly 	<ul style="list-style-type: none"> • Not anxious or depressed
UTILITY	0.74 (->0.14 when senses included)	0.80 (-> no change because senses not included)

Source: Hawthorne Richardson et al. (2001) p 368

In each case study, a score for a single individual was obtained from each of two instruments whose content is described in the figures. The number without brackets in the final row reports the scores obtained from the two instruments. In Case study 1 the same individual obtained a score of 0.14 and 0.55 on the AQoL and 15D respectively. In Case study 2 one individual obtained a score of 0.14 and 0.80 on the HUI 3 and EQ5D respectively. The number in brackets reports the score which would have been obtained if the instrument yielding the lower score omitted the items which had no corresponding items in the other instruments (ie the utility of these items were set equal to 1.00 = best outcome). These items are shown by the arrows. In both case studies 'removing' these items largely bridges the difference between the instrument scores, with AQoL rising from 0.14 to 0.49 in Case study 1 and HUI 3 rising from 0.14 to 0.74 in Case study 2. In these cases content validity, not the method of scoring or choice of scaling instrument, explained most of the difference in the results.

Why is modelling necessary: have utility models been validated?

The AQoL-4D measures 1.07 billion health states which is a very small subset of the number of health states defined by AQoL-8D! (Many are improbable, eg being blind, deaf, bedridden, full of energy and in control of your life!) These health states cannot all be measured individually and MAU instruments (except for the Rosser-Kind Index) model utility scores from a limited number of observations. To date, most instruments have adopted an additive model in which the disutility associated with each response from each item is independently measured, and the overall disutility estimated or modelled as a weighted average of these disutilities, where the weights are also obtained empirically during the scaling survey. This additive model is probably invalid (at least in theory as distinct from its practical application over a limited range of observations). For example a person encountering problems with independent living 'IL', but no other significant problems might experience a significant loss of utility – say 0.2 or a quarter of their utility. However a person who is bedridden, with a score of only 0.2 is unlikely to lose the full 0.20 when the effects of IL are included in the calculation. This may plausibly reduce the net utility by 20 percent of the previous score of 0.2 (ie to 0.16) but not to a utility of equivalent to death. This suggests the superiority of the multiplicative model employed by the HUI and AQoL instruments which reduces utilities (more or less) in percentage terms (Richardson, Hawthorne et al. 1998). However there is no certainty that even this more flexible model does not introduce significant estimation bias.

In sum, the validation of MAU instruments should be regarded as work in progress. At present, progress is exceedingly slow.

Validating utility (scaling instruments)

Validating scaling techniques such as the TTO or SG is problematical as it is difficult to observe actual trade-offs between the quality and length of life which correspond with the trade-offs implied by the various scaling instruments. Some have argued that the standard gamble should be regarded as the gold standard for utility measurement as its use assumes the axioms of von Neumann and Morgenstern. This appears to make the standard gamble results consistent with mainstream economic theory. However, the axioms have been shown to be empirically incorrect and theoretically defective (Schoemaker 1982; Richardson and Pope 2009). Because of this history there has been little discussion of the question 'how should we evaluate utilities' or, more generally, 'how should we decide upon the measurement units used for QoL'.

As used in CUA, 'utility' needs two interval properties:

-
- a) The weak (conventional) 'interval property' is that an interval or scale (eg an increase of 0.2 from 0.1 to 0.3 or from 0.8 to 1.0) has the same meaning with respect to preferences;
- b) The 'strong interval property' is that the preference for a 10 percent increase in the utility index would be the same as the preference for a 10 percent increase in the number of life years. This follows from the definition that $QALY = utility \times (life\ years)$. The left hand side of this equation is equally affected by a 10 percent increase in either of the right hand side variables. For a discussion see Richardson (1994), 'Cost Utility Analysis: What should be measured'.

Can there be negative utilities?

Yes. For very poor health states it is appropriate to ask whether the person would prefer death than live any time in the health state. If so a worse-than-death TTO question is asked: 'would you choose (i) death; or (ii) n years in the health state and (10-n) years full health'. If 'indifference' – indecision – occurs when $n = 10$ then death is equivalent to the health state. If $n = 2$ then 8 years of full health is needed (as compensation) to make 2 years in the health state equivalent to death, ie the state is worse than death. As n decreases – more compensation is needed – the worse-than-death state gets worse.

Placing a numerical value on these states, however, is difficult. If a person refused even one day in the health state followed by 10 years of full health the implied numerical value of the health state is minus infinity. This problem is discussed at length in Richardson and Hawthorne (2001) and various options are discussed and their numerical implications demonstrated. The use of lower value of -1.00 on utilities to achieve 'symmetry' with the upper value of 1.00 (the rationale offered in the EQ5D) is logically invalid as the meaning of the units changes at 0.00. The final algorithm suggested for the calculation of utilities transforms negative scores in such a way that the lower boundary is $U = -0.25$; that is, there is a disutility of 1.25.

Interview methodology is presented in detail in Iezzoni and Richardson (2009) 'Measuring Quality of Life at the Centre for Health Economics'.

D. Validity of the AQoL Instruments

Why is another instrument necessary?

Existing instruments do not correlate highly with each other (see 'Do MAU instruments correlate'). This is indicative of the fact that the content validity of the major instruments has not been satisfactorily researched and different instruments are measuring differing QoL domains. Most rely upon face validity which is an unsatisfactory basis for measurement (see 'Face validity') and convergent validity which is a weak and context specific test for the validity of an MAU instrument (see 'Validating MAU instruments'). This implies that existing instruments are being used in a way which systematically favours some interventions (where instruments have higher content validity) over other interventions where they have lower content validity.

The extent of the differences in the performance of instruments in various disease areas is unknown as researchers have used the weak evidence of convergent validity in one context to assert that instruments are 'validated', implying a generality of validity for use in economic analysis which is not justified (see 'Does correlation demonstrate validity').

Have AQL utility scores been validated?

This includes two separate issues.

- The validity of the TTO utility measurement or scaling technique ; and
- The validity of AQL questionnaires or descriptive systems (see 'Have the AQL instruments been validated' below and 'Have MAU instruments been validated').

Validating scaling techniques such as the TTO or SG is problematical as it is difficult to observe actual trade-offs between the quality and length of life which correspond with the trade-offs measured by the various scaling instruments. Consequently, validity has been determined primarily by face validity. Some have argued that the standard gamble should be regarded as the gold standard for utility measurement as its use assumes the axioms of von Neumann and Morgenstern. This appears to make the standard gamble results consistent with mainstream economic theory. However, as the axioms have been shown to be empirically incorrect and theoretically defective (Schoemaker 1982; Richardson and Pope 2009), rather, for the reasons outlined by Richardson (1994) and Dolan et al (1996), we have accepted the time trade-off as having the greatest prima facie validity.

Have the AQL instruments been validated?

How is an MAU validated?

This question cannot be answered with a simple 'yes' or 'no' and this is also true for other instruments despite the common and misleading statement that 'instrument x has been validated'. Validity is a complex issue and the validation of MAU instruments has been very limited in scope.

Four types of evidence increase confidence in the validity of an instrument, ie in confidence that it measures what we want it to measure (see 'Have MAU instruments been validated satisfactorily').

- Evidence of construct validity:** The AQL suite of instruments were the only ones to be constructed using psychometric methods developed by psychologists for achieving content and construct validity. *Other MAU instrument descriptive systems have been based upon 'logical' considerations* (face validity) or in the case of the SF6D upon another instrument, the SF36. The need for construct validity was the motivating reason for the commencement of the AQL program.
- Evidence of criterion validity for the combination model:** That is, a demonstration that the combination model predicts the scores of multi attribute health states when they are independently measured.

AQL-6D, 7D and 8D have demonstrated this property. To date other tests of this property do not appear to have been widely conducted for other instruments.
- Evidence of the validity of utility (scaling) instruments (TTO, SG, etc):** Like other MAU instruments, the AQL suite of instruments have assumed the validity of a particular scaling instrument, viz, the TTO. However the issue is problematic (and largely ignored in the literature) (see 'Have AQL utility scores been validated').
- Evidence of correlation between MAU utility scores and other instrument scores:** This constitutes the overwhelming majority of studies 'validating' other instruments. It is necessary but far from sufficient for demonstrating validity. As the newest of the MAU

instruments, the AQoL suite of instruments has not been included in many such studies (and results from the 100+ studies using an AQoL-4D instrument have not been completed and compiled). However AQoL-4D and 8D have been included in five instrument studies which indicate a sufficiently high correlation between different instruments to confidently predict that this type of evidence will be obtained. As noted elsewhere, however, this type of evidence is 'soft' in the sense that it is easily achieved as even instruments with overall low content validity are likely to correlate with other QoL instruments (see 'Correlation and Validity').

- (v) **Evidence of overall criterion validity:** No instrument has shown that the percentage change in predicted utility from an MAU instrument corresponds with the same percentage change in real trade-off decisions between the quantity and quality of life. The overall status of MAU instruments, qua predictors of utility, therefore remains problematical.

E. Conclusion

The need for caution

In sum, correlation in one context is an insufficient test of validity in another context. It cannot relieve the researcher of the responsibility for ensuring that an instrument is capable of measuring the outcomes of relevance.

References

- Dolan, P., C. Gudex, et al. (1996). 'Valuing health states: A comparison of methods.' *Journal of Health Economics* **15**: 209-231.
- Fryback, D.G., M. Palta, et al. (2010). 'Comparison of 5 Health Related QoL indexes using Item Response Theory analysis.' *Medical Decision Making* **30**(1): 5-15.
- Hawthorne, G., J. Richardson, et al. (2001). 'A comparison of the Assessment of Quality of Life (AQoL) with four other generic utility instruments.' *Annals of Medicine* **33**: 358-370.
- Iezzi, A., and J. Richardson. (2009). *Measuring Quality of Life at the Centre for Health Economics*, Research Paper 41, Centre for Health Economics, Melbourne: Monash University.
- Richardson, J. (1994). 'Cost utility analysis: What should be measured.' *Social Science & Medicine* **39**(1): 7-21.
- Richardson, J., G. Hawthorne, et al. (1998). *Difficulty with Life and Death: Methodological Issues and Results from Utility Scaling of the Assessment of Quality of Life (AQoL) Instrument*, Working Paper 70, Centre for Health Program Evaluation, Melbourne: Monash University.
- Richardson, J., G. Hawthorne, et al. (2001). *Negative utilities and the evaluation of complex health states: Issues arising from the scaling of a Multi Attribute Utility Instrument*, Working Paper 113, Centre for Health Program Evaluation, Melbourne: Monash University.
- Richardson, J. and R. Pope. (2009). *The Invalidity of Expected Utility Theory and its misuse in the Economic Evaluation of Health and Safety*, Research Paper 36, Centre for Health Economics, Monash University, Melbourne. (Forthcoming in (ed) M Schlander, *Economic Evaluation of Health Care Programs: Current Concepts, Controversies and International Experience*, Springer: New York).
- Schoemaker, P. (1982). 'The expected utility model: Its variants, purposes, evidence and limitations.' *Journal of Economic Literature* **XX**: 529-563.
- Streiner, D. and G.R. Norman. (2003). *Health Measurement Scales: A Practical Guide to their Development and Use*. Oxford: Oxford University Press.