# NATIONAL INSTITUTE FOR HEALTH AND CLINICAL EXCELLENCE

# Briefing paper for methods review workshop on Measuring and valuing health effects

The briefing paper is written by members of the Institute's Decision Support Unit. It is intended to provide a brief summary of the issues that are proposed for discussion at a workshop to inform an update to the Institute's Guide to Methods of Technology Appraisal. It is not intended to reflect a comprehensive or systematic review of the literature. The views presented in this paper are those of the authors and do not reflect the views of the Institute.

The briefing paper is circulated to people attending that workshop. It will also be circulated to the members of the Method's Review Working Party, the group responsible for updating the guide.

For further details regarding the update of the Guide to the Methods of Technology Appraisal please visit the NICE website at http://www.nice.org.uk/aboutnice/howwework/devnicetech/technologyappraisalprocessguides/GuideToMethodsTA201112.jsp

## 1 Review of the 'Guide to Methods of Technology Appraisal'

The Institute is reviewing the 'Guide to the methods of technology appraisal', which underpins the technology appraisal programme.

The original Methods Guide was published in February 2001, and a revised version was published in 2007. The Methods Guide provides an overview of the principles and methods used by the Institute in assessing health technologies. It is a guide for all organisations considering submitting evidence to the technology appraisal programme and describes appraisal methodology.

The current 'Guide to methods of technology appraisal' is available from the NICE website at http://www.nice.org.uk/aboutnice/howwework/devnicetech/technologyappraisalprocessguides/guidetothemethodsoftechnologyappraisal.jsp

The review of the Methods Guide will take place between October 2011 and April 2012. As part of the process, a number of workshops will be held to help identify those parts of the Guide that require updating. These workshops will involve a range of stakeholders, including methods experts, patient representatives, industry representatives, NHS staff and NICE technology appraisal committee members.

A summary of the discussion at the workshop will be provided to the Methods Review Working Party, the group responsible for preparing the draft update of the Methods Guide. Further details of the process and timelines of the review process are available from the NICE website.

The revised draft of the Methods Guide will be available for a 3-month public consultation, expected to begin in May 2012. We encourage all interested parties to take part in this consultation.

# 2 Background

## *2.1 Current NICE reference case on health effects*

The NICE reference case on measuring and valuing health effects contains the following key features:

*Quality Adjusted Life Years (QALYs) as the measure of health effects*

The QALY combines the effects of an intervention on survival and health related quality of life (HRQL) into a single measure, by placing HRQL onto a scale where full health is one and dead is zero. This allows all health outcomes to be expressed in a common metric that allows comparisons across interventions. This has been a cornerstone of cost effectiveness methods in NICE Technology Appraisals for many years. There have been concerns that this focus excluded potential impacts on non-health outcomes

and aspects of the processes of care. This issue was partly addressed in the workshop on perspectives and is considered briefly in this review.

*HRQL should be reported by patients*

The Methods Guide states clearly that it expects to see the HRQL data to come from patient self-report. This reflects the evidence that carers and professionals reporting on the health of patients is often in disagreement with those of the patient, particularly for the more subjective dimensions of pain or mood. However, the Guide recognises that there may be circumstances where patients are unable to report their own health (e.g. cases of severe cognitive problems) and in this case the Methods Guide specifies close carers as the source for proxy data. This issue is not considered further in this review.

*The health effects on carer givers is included*

It is sometimes forgotten that the NICE Methods Guidance does allow the impact on caregiver's health to be included in the QALY calculation (unless the caregiver is employed by the NHS). There is a question of whether the measurement of external effects should extend to other family members not involved in caring and there is some debate in the literature on this subject (to which we return later).

*Health effects valued using a choice-based method by the general public*

The use of choice-based methods has been stipulated in the last two versions of the NICE Methods Guide, which are those preference elicitation techniques that require respondents to explicitly consider a trade-off between HRQL and some other part of their utility function, such as longevity (i.e. time trade-off (TTO)) or risk of death (i.e. standard gamble (SG)), rather than rating scales that ask for an assessment of feeling about a state. EQ-5D is the preferred instrument and it uses TTO to value health states (see below). Where other preference-based measures are used the Methods Guide requires the use of comparable valuation methods to EQ-5D, in other words TTO. This review examines this issue in as much as the EQ-5D will be re-valued in the near future probably using a different variant of TTO.

NICE decided on using general public values rather than those of patients or others due to the perspective of the decisions being taken. However, this continues to be a subject of considerable debate in the literature and policy circles, but not for this review

*EQ-5D as a preferred measure of HRQL*

The EQ-5D is a generic preference-based measure of health and has been validated in many conditions. The version of EQ-5D currently in use consists of 5 dimensions (mobility, ability to self-care, ability to perform usual activities, pain and discomfort, and anxiety and depression) and each dimension is described by a single 3 level item. Patients complete a 5 item one page questionnaire in order to assign them to one of the 243 states this descriptive system defines. There are a set of estimated preference-based health state values for each of the 243 states using values elicited from the UK general population by TTO. Recently the EuroQol Group has produced a 5 level version and is currently embarking on a programme of work to produce a new UK value set. An important consideration is whether these developments should be incorporated into the NICE reference case and if so, when.

NICE prefers a single measure of HRQL to be used in cost effectiveness models to promote consistency across appraisals. There is substantial evidence that other preference-based measures of health, be they generic (e.g. HUI3 or SF-6D) or condition specific, produce different values for the same patient. In order to compare across studies it is important to use the same measures. However, this raises the issues of what to do when EQ-5D data are not available and when EQ-5D is not appropriate in the patient groups.

*EQ-5D is appropriate but unavailable: use of mapping*

The amount and coverage (e.g. by medical condition) of EQ-5D data are increasing all the time. However, the NICE methods Guide recognised that sometimes there may not be sufficient relevant EQ-5D data and so they recommended the use of mapping (or cross-walking) methods in order to generate EQ-5D values from other measures of HRQL or even other clinical measures. This raises two important questions: when is it appropriate to use

mapping and how should mapping be undertaken? Nothing more specific was said in the last Methods Guide about when mapping should be undertaken. The Guide specifies that mapping should be based on empirical data (i.e. rather than judgement), it should have clearly described statistical properties and it should be validated. The NICE DSU Technology Support Document on Mapping (TSD 10) provides useful advice on how to undertake mapping studies. An issue to be addressed at this workshop is whether any of this advice should be incorporated into the NICE Methods Guide.

*Appropriateness of EQ-5D and the alternatives*

NICE recognised that there may be conditions or treatment effects that will not be adequately captured by the 5 dimensional 3 level EQ-5D. However, this was anticipated to be the exception rather than the rule. The inappropriateness of EQ-5D needs to be demonstrated with evidence on the properties of content validity, construct validity, responsiveness and reliability in the relevant patient population. Where an alternative measure is used, then the submission should give the reasons supported by evidence on these same properties.

Guidance on alternatives to EQ-5D was that these should be based on the direct valuation of a standardised and validated HRQL measure. This would seem to suggest that those states developed by experts, sometimes known as vignettes, should not be used in the reference case because they do not relate to patients reporting on their HRQL and so have little empirical basis. Another generic measure or a condition specific measure may be considered. However the NICE Methods Guide states that '…*the valuation of the descriptions should use the TTO method in a representative sample of the UK population, with 'full health' as the upper anchor, to retain methodological consistency with the methods used to value the EQ-5D*'. Of course, those submitting evidence are allowed to use other methods in any sensitivity analyses. This continues to be a contentious topic and there has been substantial research in the use of condition specific measures, so it seems right to re-consider it at this workshop.

*Use of measures in children*

It is recognised in HTA that there are important conceptual differences between children and adults in terms of the dimensions of HRQL as well as linguistic differences. It was recognised that there was not an obvious candidate measure for the reference case measure of HRQL in children. At the time of preparing the last Method Guide there was just the HUI2, but this was not felt to have the same status and uptake as EQ-5D to justify making it the preferred measure. The Guide asks those submitting evidence to consider the use of standardised and validated preference-based measures of HRQL, including the HUI2. Since that time a number of measures have been developed for use in children and so this review will consider this question further to see whether there is a clear candidate measure and the issue of whose values.

*Use of the literature and other secondary sources*

It is recognised that for populating cost effectiveness models, EQ-5D data may come from a number of different sources. Clinical trials have the attraction of internal validity, but they may not be generalisable to the populations being modelled, they may not follow-up outcomes for long enough and may not have sufficient data on key events (e.g. adverse effects). For this reason, it will often be appropriate to use other sources of data, such as observational studies, routine data sets (e.g. UK PROMS) or values published in the literature.

The NICE Methods Guide requires that estimates for the utility values for health states from published literature must be shown to have been identified and selected systematically. Where there is more than one plausible source of health state values sensitivity analyses are recommended. The ever growing published literature makes this an increasingly important source of values. This review does not propose to look at this issue any further, but readers are recommended to consult TSD 9 which provides detailed recommendations on how to conduct reviews of health state utility values (Papaioannou et al, 2011).

### *2.2 Relevance of health effects to the Appraisal Committee*

The measurement and valuation of the health effects of technologies is a fundamental component of the assessment of the cost effectiveness of health care interventions for the NICE Technology Appraisal. The previous review of NICE methods published in 2008 provided an overview of the core issues in measuring and valuing health including what is to be measured (e.g. should it be the QALY?), how it is to be described (e.g. should it be generic or specific to the condition?), how it is to be valued (e.g. time trade-off or standard gamble), whose values (e.g. general population or patient) and how should it be aggregated (e.g. is a QALY is a QALY regardless of who gets it or should QALYs be weighted in some way) (Brazier, 2007). The purpose of this briefing paper is not to revisit these core issues in general but to address a number of specific questions that have emerged since the publication of the last review of methods where it is deemed that there needs to be firmer guidance or where there have been important developments or research that have implications for the existing reference case methods in this area. Some of these have arisen from the development of the NICE DSU Technical Support Document (TSD) series of 5 on utilities (for further information see www.nicedsu.org.uk). The TSD series provides a review of the state of the art across a number of important issues in this topic to assist those making Technology Appraisal submission to NICE, but it is not a formal part of the NICE Methods Guide.

## 3 Proposed issues for discussion

After consideration of the developments in this methodological area, the current Methods Guide and the requirements of the Institute's Technology Appraisal Programme, it is proposed that the following key areas are discussed at the workshop:

1. When is the EQ-5D not an appropriate measure of health-related quality of life?

2. What are the alternative instruments and when are they more appropriate?

3. When is mapping the preferred approach? What principles underpin good mapping analysis?

4. Should NICE adopt the new 5 level version of EQ-5D and its associated value set?

5. What preference-based measure of HRQL should be used in children?

6. Measurement and valuation of health effects on people other than the recipient of the intervention. How should ‘related individuals’ be defined, measured and aggregated?

## *3.1 When is the EQ-5D not an appropriate measure of health related quality of life?*

The NICE reference case expressed a preference for the EQ-5D to measure HRQL in adults. It permits the use of other measures where it can be demonstrated that EQ-5D is not appropriate and provide reasons for the alternative supported by evidence. The Guide specifies the properties of reliability, content validity, construct validity, and responsiveness for assessing appropriateness.

Reliability takes two forms. One is random variation between assessments, and this has implications for sample size and precision of estimates for any given sample size. Where sample sizes are small, then this may be a cause for concern and there could be a case for using estimates from another instrument prone to less variation. However, in most cases a larger sample size will be the solution. Of more concern is unreliability from variation between methods of assessment. There is little evidence on this issue but what there is suggests there may be little difference between pencil and paper and computer completion of EQ-5D (Lloyd et al, 2011). However, there is evidence of significant differences between patient self-report and carer proxy report.

The assessment of validity is far more problematic due the lack of a gold standard measure of HRQL. While some health economists have been sceptical as to whether it is possible to assess the validity of preference-based measures, it is an important challenge facing the measurement of all psychological phenomena. The methods Guide recognised that the EQ-5D is not appropriate in all populations (Brazier and Longworth, 2011; Wailoo et al, 2010).

Content validity is concerned with whether the instrument covers all the dimensions of HRQL of importance to patients. This can be assessed through qualitative work with patients to identify ways in which their health status impacts on their physical, psychological and social functioning and wellbeing. Construct validity requires quantitative evidence on whether the measure reflects known differences between groups or converges with other relevant measures. Responsiveness is the extent to which the measure reflects changes in HRQL overtime. These criteria would preferably be assessed across the 5 dimensions of the EQ-5D as well as the overall index, though this is rarely done. Careful consideration must be given to the relevance of the variables used to test validity, which are often clinical assessments of symptoms, since these may not be important for patient's HRQL. Furthermore some of the conventional psychometric criteria may not apply to preference-based measures. In conventional psychometric analyses it is the instrument with the largest difference that is deemed best, for example as assessed using a standardized effect size (mean difference divided by standard deviation of the difference). However, bigger is not necessarily right since a highly focused measure of symptoms may achieve the highest effect size and yet not reflect the impact on overall HRQL and not be valued by the general public. The other extreme would be to argue that there must be no differences between the known groups or changes over time for the EQ-5D to be judged invalid. The truth may lie somewhere between the two and there will always remain a considerable degree of judgment in deciding on the validity of EQ-5D in any one patient group and whether another measure is more appropriate (as a measure of HRQL that matters to the general population).

An empirical literature on the validity of EQ-5D and other preference-based measures has begun to emerge over recent years. The standards of testing are often not high and are prone to the problems of interpretation highlighted above. While this is not the place to present a detailed review of the evidence, there have been a number of reviews recently conducted of the literature that give some idea of the extent of the problem. Evidence from recent reviews on construct validity and responsiveness suggests the EQ-5D is probably not appropriate for assessing the impact of hearing loss, some specific forms of visual impairment and schizophrenia (TSD 8). On the other hand it would seem that the EQ-5D is more appropriate in areas including depression and anxiety, a number of key cancers and skin conditions. However the evidence is at best patchy and often poor quality, with little evidence on content validity. In many cases, there is simply not sufficient evidence one way or the other to make definitive judgements about the suitability of EQ-5D for a given condition and there is often even less evidence on other generic or the condition specific preference-based measures.

Where alternative measures are used, those submitting evidence are required to demonstrate the likely impact on the cost effectiveness of the intervention (i.e. through sensitivity analyses). In many cases, it may not impact on the decision. However, where there is a potential impact on the decision, this still leaves the Appraisal Committee with a judgement about which should be used.

*Discussion points:*

- Should NICE penalise products that don't have EQ-5D data?
- How strong an evidence base is required to decide a measure is inappropriate?
- Should NICE provide more guidance on what evidence is required, how it should be reviewed and how it should be presented?
- Should NICE stipulate in advance, such as in the scoping stage, whether other measures are deemed more appropriate than EQ-5D (e.g. HUI3 in hearing loss)

## *3.2 What are the alternative instruments and when are they more appropriate?*

The reference case argues strongly that alternatives to EQ-5D should be based on a validated patient reported outcome measures rather than vignettes based on expert opinion and they should be valued using methods comparable to those for the EQ-5D. While the range of alternative generic measures in adults has not changed, there has been a large increase in the number of condition specific preference-based measures covering diseases such as asthma, cancer, dementia, sexual functioning, Parkinson's disease, visual function, urinary incontinence, mental health and many more with 28 identified in a recent review (Brazier et al, 2011). The concern about condition specific measures comes from the lack of comparability across them and so limits their use in making decisions across programmes. This problem may arise even where the valuation methods are the same (i.e. same upper anchor, same valuation method and same source of values) due to focusing effects (whereby respondents overemphasise those specific dimensions mentioned in the state since they are not placed in the context of overall HRQL), use of disease labels (that may lead to respondents in valuation surveys bringing irrelevant prior beliefs about the condition into their responses), and problems capturing important side effects and comorbidities (that may interact with condition specific dimensions).

There has been little work comparing these new measures to existing generic measures like EQ-5D in terms of their validity. What there is suggests that the condition specific measures do not tend to produce larger differences in utility values, though there are cases of that, but rather they provide more precise estimates because they are associated with smaller standard deviations. This is important for reducing the uncertainties around specific estimates. However, the evidence on whether they are more sensitive to particular differences is mixed, with some evidence that they are better at reflecting differences at the upper end of HRQL (Brazier et al, 2011). This does not suggest that focusing effects are unimportant, but rather that comorbidities

seem to be more important. However, this experience is likely to vary between conditions and measures.

More recently researchers have begun to examine the potential for adding on extra dimensions to EQ-5D as another means of overcoming the apparent lack of sensitivity or relevance in some conditions. Research has examined 'bolt-on' dimensions for vision, hearing, sleep and cognition. This research is at an early stage, but it has the potential of improving EQ-5D in some key conditions while at the same time overcoming some of the limitations with condition specific measures.

Other alternatives continue to be used, such as vignettes and patient's own valuations (where they value their own state using TTO or SG). Should these data continue to be admissible as evidence in submissions to NICE technology Appraisals? If so, should this be agreed at the scoping?

Finally, there is a concern that important elements of patient's experience of the processes of care are excluded from outcomes measures like EQ-5D, such as regular hospital attendance, oral versus insulin medication for diabetes and dignity of care. These have been dealt with using vignettes in some submission that brings concerns about having a poor evidential basis. There is a growing literature using techniques such as DCE to combine process and outcome attributes. This is promising where patient experience is being assessed using validated patient reported measure, but it raises two concerns. One is that this extends the scope of the appraisal of benefits beyond health and hence beyond the current reference case. Secondly, even where it is decided such benefits should be taken into account they are often on a different scale to the health effects. There has been work attempting to treat the process attribute as a bolt on to the EQ-5D, but this is at a very early stage of development.

*Discussion points:*

- When should alternative measures be used? When EQ-5D is appropriate (instead of mapping) or only where EQ-5D is inappropriate?

- Should alternatives be presented in the main analyses where EQ-5D has been shown to be inappropriate?
- For those conditions where EQ-5D is shown to be inadequate, should NICE express a preference for an alternative measure (e.g. preference-based VFQ-25 in visual functioning).
- Does there need to be evidence demonstrating how use of an alternative has impacted on QALY estimates?
- What should be the role of other (i.e. non-reference case) alternatives such as vignettes, patient values and should the experience of process benefits be taken into account?

### *3.3 When is mapping the preferred approach? What principles underpin good mapping analysis?*

Mapping (or cross walking) involves the development and use of an algorithm to predict EQ-5D values using data on other measures or indicators of health. The mapping algorithm should be based on statistical association and not expert judgement. The estimation of mapping functions requires an estimation sample containing the target variable (i.e. EQ-5D) and the source variable (e.g. another measure of HRQL). A statistical model is then estimated mapping the source onto the target using a range of possible specifications and estimation techniques and then it is applied. It can be used to predict EQ-5D values from data sets where it was not used, such as clinical trials or observational studies that are being used to populate an economic model. A recent review found that one quarter of submissions to the TA programme had used health state values from mapping algorithms (Tosh et al, 2011).

Mapping is usually a second best solution to using the EQ-5D in the population of interest (but there may be some exceptions to this such as where the sample in the trial is too small to obtain sufficiently precise estimates). As described below, there are known errors in mapping models that are best avoided. So an important question is when should it be the preferred approach? It should only be used when there are insufficient

relevant EQ-5D data. For some health states in a model, relevant EQ-5D values may already exist in the literature and so predictions based on mapping functions would be inferior. To ascertain the existence of relevant EQ-5D values requires a systematic search and review of existing literature. (For advice on how to do this see TSD 9). It might also be necessary to adjust published values to make them suitable for the population in the economic model, such as age or the existence of comorbidities. There are methods for making such adjustments and these are described in TSD 12 (Ara and Wailoo, 2011). The choice of using health state values from mapping, literature sources or EQ-5D data from specific trials depends on context. Mapping may be preferable to the literature where the latter does not cover the right population or misses important side effects of treatment, on the other hand literature values may be based on direct use of EQ-5D and better reflect the population of interest in the model than a pivotal trial.

Details concerning the methods of mapping are provided in TSD 10 (Longworth and Rowen, 2011). In summary the key concerns in mapping cover the estimation sample, model type, the model specification, uncertainty and validation. The characteristics of the estimation sample should be similar to the sample to which the mapping function will be applied. The choice of model should depend on the nature of the data and the expected relationships. EQ-5D data are not easy to model due to the skewed, censored and multi-modal nature of the distribution of the values. Ordinary Least Squares (OLS) regression models tend to be the most widely used, but this has theoretical limitations though often performs better than the alternatives. Attempts to improve on OLS include the Tobit, CLAD (but this provides median estimates), two part models, splining or mixture models. Some have modelled the responses to the classification rather than the EQ-5D index, which involves a two stage procedure of modelling onto the 5 dimension responses and then applying the EQ-5D value set. It is not possible to recommend any one method in all cases at this stage.

There should be clear reporting of the model and its performance. This should include statistical properties such as coefficients (e.g. size, significance),

mean absolute and root mean squared error; error reported across the EQ-5D score range and plots of observed to predicted values. Mapping functions should ideally be validated using external datasets. Mapping functions are often poorly reported in the literature with little attention given to such things. One solution is to be more prescriptive about reporting standards for mapping functions used to populate models and even require the data sets on which they are based to be made available to the Technology Assessment Groups where it has not been published.

A common finding is that mapping functions overestimate at the lower end and under-estimate at the upper end, and this can result in a reduction in the size of differences between health states based on severity or changes overtime. On the other hand mapping functions can result in less variability than the original EQ-5D. There is a tendency to ignore the uncertainties underlying the statistical model in the sensitivity analyses. NICE and other using mapping functions need to better understand the impact of using values estimated by mapping functions than observed EQ-5D values in cost effectiveness models.

*Discussion points:*

- When should mapping be used compared to using original EQ-5D data or literature values? When should mapping be used rather than alternatives (see above)?
- Should NICE recommend specific systematic reviews or databases of HSUVs for those submitting evidence and reduce the need for mapping?
- Does NICE need to be more prescriptive about the principles or methods used to estimate mapping functions and how they are presented? Or is the advice in TSD 10 sufficient?
- Should NICE recommend stand mapping functions or agree on one to be used at the scoping stage?

- How should the uncertainties underlying mapping functions be reflected in the cost effectiveness model?

## *3.4 Should the NICE Methods Guide adopt the new 5 level version of EQ-5D and its associated value set?*

### *3.4.1 The 5 level EQ-5D*

While the 3 level version EQ-5D has been shown to be valid and responsive in many conditions it has been criticised for the crudeness of having just 3 levels. With just 3 levels there are large proportion of patients at the ceiling (i.e. many respondents with health problems are allocated to state 11111) and a general insensitivity to change when the response categories involve such large steps. The EuroQol Group has been developing a five level version that retains the 5 dimensions with the descriptors adapted to a 5 level version as follows: no problem, slight problems, moderate problems, severe problems and unable to or extreme problems. The worst level of mobility has been changed from ‘confined to bed’ to ‘unable to walk’ and usual activities from performance to doing. Papers are starting to emerge using theEQ-5D- 5L and an important question is whether and how this version should be incorporated into NICE Guidance.

The argument for using the 5L for collecting data is that it would provide a more sensitive instrument. The evidence to date that this is the case is quite limited. There is evidence of a reduction in the numbers at level 1, a more even distribution across the levels and a modest increase in the correlations with related measures of health (Bas et al, 2011). There are only a couple of studies and these have been conducted by members of the EuroQol Group and not by independent researchers (Pickard et al, 2007, Janssen MF et al, 2011). There would be an intuitive case for saying 5 levels is an improvement, but the size and extent of the improvement across conditions is not known. Furthermore, there is no published evidence on the extent to which general population respondents in a valuation survey are able to distinguish between the 5 levels.

Another limitation is that no UK tariff for the 5L currently exists. There are plans to produce one in the UK (to be funded by DH), and these are discussed below. Mapping functions have been estimating for scoring the 5L from the 3L tariff. A number of methods have been explored for estimating mapping functions including OLS, non-parametric models, ordered logistic regression and item response theory (Janssen et al, 2011). These seem to achieve a similar fit with RMSEs of around 0.12. These functions suffer from familiar problems with a reduced range (since it does not predict one in many cases) and slightly flatter gradient to the predictions than would be expected from an exact fit. The implication of this error for differences between key states has not been explored (e.g. between grades of severity of different conditions).

There is a major cost to NICE in recommending an instrument that will in the end produce different values for the same patients as the EQ-5D-3L. Possible recommendations to NICE include: never adopt, adopt after further evidence and the value set become available or recommend it is used now and in the interim use the value set from the 3L. Assuming the 5L brings advantages and there will be a re-valuation of the EQ-5D in the UK in case (as planned), then to never adopt may become an untenable position. To delay recommending the use of the 5L until the next review would delay any benefits by four years. To recommend its use now and suggest the mapping function will bring some of the problems associated with mapping and indeed further statistical complications for mapping from other measures onto EQ-5D (i.e. the double mapping problem).

### *3.4.2 The new value set for EQ-5D*

The reference case tariff of values was obtained from members of the UK general population more than 15 years ago using TTO. The version of TTO was the MVH protocol where for states better than dead respondents are asked to compare living in health state h for 10 years and x years in full health (where x<10). At the point of indifference the value of h is x/10. For states worse than dead the choice is between (a) health state h for y years followed by full health for x years, after which they will die, or (b) immediate death.

Years in the health state, y (=10-x), and years in full health (x), are varied to determine the point where the respondent is indifferent between the two options. The value of h that is consistent with the theory is –x/y (i.e. x/(10-x)) . However when using the TTO protocol where t=10 this produces values bounded at -39 for the minimum possible value for any health state, where x=9.75 (i.e. 3 months followed by full health for 9 years and 9 months). State worse than dead responses have a larger impact on the model predictions than better than dead responses. For this reason the TTO data for states worse than dead were rescaled to onto 0 to -1 using for formula –x/10 (Dolan, 1997). The values for states better than dead and the transformed values for states worse than dead are pooled and modelled using regression techniques to estimate the tariff.

This review does not address the more general concerns with using TTO (such as the assumption of constant proportional trade-off) or the use of preferences rather than patient experience to value states). An important criticism of the value set, aside from its age, is the handling of states worse than dead. This is important since a third of mean EQ-5D health state values are negative and so worse than being dead and all other states have some negative responses. It currently uses a different valuation procedure for states worse than dead and respondents may view the prospect of returning to full health following a severe health state as unrealistic. The rescaling is arbitrary and it has been argued that the values can no longer be interpreted as utility values. The values produced by the two procedures are arguably not on the same scale.

One approach to deal with the latter problem is to incorporate the correct formulae for states better and worse than dead into the econometric model via an ‘episodic random utility model (Craig et al, 2009). The main contribution of the episodic RUM model is that all TTO responses are treated identically in the model specification. Yet this does not resolve the problems outlined earlier that the TTO choice tasks are different for states valued as better than or worse than dead. This approach is not being used by the EuroQol Group.

Another proposal is to use a different TTO procedure, such as one that introduces a ‘lead time’ whereby a period in full health is added to the start of the usual TTO, meaning that states worse than dead can be valued by cutting in to the lead time (Devlin et al, 2011). The ‘lead time’ TTO task provides respondents with a choice between (a) full health for f years followed by health state h for 10 years, after which they will die, or (b) full health for f+x years, after which they will die. Years in full health, x, is varied to determine the point where the respondent is indifferent between the two options where x can be negative where the lead time is exhausted. The utility for health state h is calculated using x/10. This approach has the advantage that it does not draw attention to the fact that respondents are valuing a state as worse than dead, yet this may mean that respondents are not fully aware of what their responses indicate. The lead time can be exhausted and so respondents may have to revert to a different procedure in some cases. This method also makes a strong assumption of additive separability where the value of state h should not be affected if it is preceded by full health for period f, and may suffer from the problem of ordering effects in moving from full health to a poor health state. This new procedure and others (including a ‘lag’ time TTO) is the subject of further methodological research being undertaken by members of the EuroQol Group and elsewhere.

Finally there has been research looking into the use of ordinal methods for valuing EQ-5D states. Initially this looked at the use of rank data (Solomon, 200X), but more recently the research has begun to look at discrete choice tasks. Asking respondents to compare EQ-5D states will provide values for those states on a latent scale, but leaves the problem of how to anchor onto the full health-dead scale required for calculating QALYs. One solution to this problem is a hybrid approach, whereby some states are valued by TTO and then the DCE and TTO data combined through anchoring or mapping to produce a value set (Rowen et al, 2011)). Lastly, there is a DCE task where survival is added as a sixth dimension and in effect providing a new TTO task where the pairs of scenarios are determined by a statistical design rather than a standard elicitation procedure. This does not get away from some of the concerns with TTO, such as the assumption of constant proportional trade-off

(although this can be tested within this approach), but it avoids the need for a different task for states worse than dead. Initial testing of this approach suggests it has promise (Bansback et al, 2010) and is currently being examined by the EuroQol Group.

The EQ group has been testing the various alternatives and currently has not decided on the best approach. However, it intends to make a decision in the near future.

- Does the 5L represent a sufficient improvement for NICE to recommend it is used: 1) as a reference case or 2) to collect data for the time being, and be adopted as the reference case at a later point in time?
- Will the 5L compromise NICE's need to be consistent in decision making? How will submissions using the 5L be compared to previous 3L ones?
- Should NICE adopt the EuroQol Groups final decision regarding the method of valuation?
- If yes, when should a new tariff be adopted as the reference case by NICE?
- What should be the transition arrangements for moving from 3L to 5L??

## *3.5 What preference-based measure of HRQL should be used in children?*

There are now three preference-based measures for children or adolescents (HUI2, AQOL-6D or AQOL-8D, and CHU-9D) and one in development (EQ-5D-Y). The HUI2 has 6 dimensions (sensation, mobility, emotion, cognition, self-care and pain) and comes with a UK SG value set elicited from adults (in addition to the Canadian used in most published studies. It was developed by experts based on a survey of parents in Canada. The AQOL-6D has six dimensions (independent living, mental health, coping, relationships, pain and senses) that were adapted from the adult instrument (ref) and there is an

Australian value set obtained using TTO elicited from adults. The AQOL-6D also has a valuation tariff from adolescents which was developed using a transformation of the adult values from a sample of states valued by adolescents. The AQOL-8D contains 2 additional dimensions to the AQoL6D and has a valuation tariff from adults. The CHU9D is the only instrument where the content of the descriptive system was developed from interviews with children about the way their health impacts on their HRQL. It was developed in children aged between 7-11, but has been used in adolescent children up to 17 years. Finally there is the EQ-5D-Y whose descriptive system has been developed from the adult EQ-5D without any alteration of the conceptual dimensions, just a change in language to make it understandable by young people. This continues to be under development and currently does not come with a value set. While these measures are starting to be used more in research, particularly in their self-reported form, there is no single measure that stands out in terms of being more widely adopted or performing notably better.

The measurement and valuation of HRQL is more complicated in children and raises important practical problems and normative issues. While self-report is being increasingly used, there are difficulties in younger age groups. There is little experience in younger children (e.g. <7), where measures of health tend to be confounded by childhood development (e.g. scores can improve simply because the child gets older). Indeed the relevance of any of these measures in the under 5 population is questionable. It is also not clear where the boundaries are between childhood, adolescence and adulthood, and how the transition between instruments should be handled when calculating QALYs or trying to make cross programme comparisons. All existing instruments use adults to value the states, but there has been interesting work in trying to elicit preferences form older children using ordinal methods that is showing promise (Ratcliffe et al, 2011), though problem of anchoring onto the full health-dead scale remains. The question of whose values presents an important normative dilemma and one that will vary by age (if for no other reason than younger children may not understand the task).

Research into measuring and valuing HRQL in children is on-going and many of these issues cannot be resolved at this workshop.

*Discussion points:*

- Are separate measures required for children, adolescents and adults, if so, what should be the ages of transition?
- Should NICE be encouraging self-report (at least in older children)?
- Should one instrument be preferred over the rest for certain age group?
- Are adult values acceptable or should NICE be encouraging the development of values sets based on the values of children and adolescents?
- How should the transition between instruments be dealt with in a cost effectiveness analyses of interventions with impacts across age groups and should comparison be made across programmes by age?

## *3.6 Measurement and valuation of health effects on people other than the recipient of the intervention. How should 'related' individuals be defined and how should the effects be measured and aggregated?*

With an ageing population, the health system increasingly relies on close family and friends to provide informal care. This may impact on the caregiver's health, and the current NICE reference case allows for the incorporation of these health effects in the calculation of the overall QALY impact. Within the reference case this would normally be measured using the EQ-5D. The time of carers is not currently included within the NHS and social care perspective taken in the reference case.

There are important questions regarding who should be counted as a caregiver. It does not include professional caregivers who are already compensated for their time and effort and will be included in the staff cost in an economic evaluation. Providing informal care has been shown to impact on physical and/or psychological health, and has even been associated with a

higher risk of morality (Brouwer, 2006). This is taken into account in the current NICE reference case. However, there may be significant others who do not provide care (e.g. the children of ill parents) whose health will be affected by having members of their family who are unwell, particularly through their emotional well-being (Bobinac et al, 2010). To exclude such 'family' effects' requires the separation of family members not only into two groups, those who give care and those who do not. It also means having to net out the family effect in those who give care. To include them substantially increases the data requirements of economic models.

Another consideration is whether carer and/or family effects are already proxied by the EQ-5D. In which case, there is little need to add it into the QALY estimate since it will impact on all interventions equally. However, it is suspected that for a given EQ-5D score the impact will vary by condition, severity of condition, age (e.g. children), type of treatment (e.g. at home or in hospital), type of care being provided and the nature of the relationship. However little is understood about these relationships at present.

Finally there is the question of how to measure the impact on carers and family beyond health effects. Recent years has seen the development of quality of life scales for use with carer. However, these scales are not anchored on the full health-dead scale and even if they could it raises an important problem of how to aggregate broader measures of quality of life in carers with the health effects of the patients. If the measure for carers uses a broader notion of quality of life then why should the measure for patients be limited to HRQL?

*Discussion points:*

- Should the impact on significant others be broadened out to include other members of the family who are not directly involved in care?
- Should the impact on carers and significant others be limited health effects or extended to quality of life more generally?
- How should impacts on carers, significant others and parents be aggregated?

# 4 References


Ara R, Wailoo A. NICE DSU Technical Support Document 12: The use of health state utility values in decision models. 2011. Available from http://www.nicedsu.org.uk

Bobinac A, Van Exel NJ, Rutten FF, Brouwer WB. Caring for and caring about: disentangling the caregiver effect and the family effect. J Health Econ 2010 Jul;29(4):549-56.

Brazier J. Valuing health states for use in economic evaluation. *Pharmacoeconomics* 2007, 26(9):769-779.

Brazier J, Rowen D. NICE DSU Technical Support Document 11: Alternatives to EQ-5D for generating health state utility values. 2011. Available fromhttp://www.nicedsu.org.uk

Brazier J, Longworth L. NICE DSU Technical Support Document 8: Applying the NICE reference case to the measurement and valuation of health. 2011. Available from http://www.nicedsu.org.uk

Brazier J, Rowen D et al. Developing and testing methods for deriving preference-based measures of health from condition specific measures (and other patient based measures of outcome) Health Technology Assessment (forthcoming)

Craig BM, Busschbach JJ. (2009) The episodic random utility model unifies time trade-off and discrete choice approaches in health state valuation. Population Health Metrics. 13 no. 7: 3.

Devlin, N., Tsuchiya, A., Buckingham, K.J., Tilling, C. A Uniform Time Trade Off Method for States Better and Worse than Dead: Feasibility Study of the 'Lead Time' Approach. *Health Economics* 2011; Forthcoming.

Dolan, P. 1997. Modeling valuations for EuroQol health states. *Medical Care* 1095-1108

Janssen MF, Birnie E, Haagsma JA et al. Comparing the standard EQ-5D three level system wuth a five level version. Value in Health (in press).

Janssen MF et al. Values sets for the EQ-5D-5L. EuroQoL Paper.

Lloyd, A.J., Kind, P., Thompson, T., Leese, B., Nixon, A., Quadri, N. Paper to web: equivalence testing of EQ-5D report to the Department of Health. 2011.

Tosh JC, Longworth LJ, George E. Utility values in National Institute for Health and Clinical Excellence (NICE) Technology Appraisals. *Value in Health* 2011;14(1):102-9.

Longworth L, Rowen D. DSU Technical Support Document 10:The use of mapping mehtods to estimate health state utlity values. Available from http://www.nicedsu.org.uk

McCabe CJ, Stevens KJ, Brazier JE. Utility scores for the Health Utilities Index Mark 2: an empirical assessment of alternative mapping functions. Med Care 2005 Jun;43(6):627-35.

National Institute of Health and Clinical Excellence (NICE). Guide to the methods of technology appraisal. London: NICE; 2008.

Papaioannou D, Brazier J, Paisley S. (2011) NICE DSU Technical Support Document 9:The identification, review and synthesis of health state utility values from the literature. Available from http://www.nicedsu.org.uk

Pickard AS, De leon MCV, Kohlmann T et al. Psychometric comparison of the standard EQ-5D to a 5 level version in cacner patients. Med care 2007; 45(3):259-263.

Salomon, J.A. 2003. Reconsidering the use of rankings in the valuation of health states: a model for estimating cardinal values from ordinal data. *Popul.Health Metr.*, 1, (1) 12 available from: PM:14687419

Stevens K. Developing a descriptive system for a new preference-based measure of health-related quality of life for children. Qual Life Res 2009 Oct;18(8):1105-13.

Wailoo A, Davis S, Tosh. The incorporation of health benefits in CUA using the EQ-5D. NICE DRU Report, 2010. http://www.nicedsu.org.uk/PDFs%20of%20reports/DSU%20EQ5D%20final%20report%20-%20submitted.pdf

# 5 Author/s

Prepared John Brazier (Health Economics and Decision Science, School of Health and Related Research, University of Sheffield) on behalf of the Institute's Decision Support Unit, October 2011.

# 6 Acknowledgements

The author is grateful for comments on earlier drafts from Meindert Boysen, Carole Longson, Louise Longworth, Donna Rowen, Andrew Stevens, Paul Tappenden and Allan Wailoo.