MEDICAL CARE
Volume 35, Number 11, pp 1095–1108


# Modeling Valuations for EuroQol Health States

PAUL DOLAN, DPHIL

**OBJECTIVES. It has become increasingly common for preference-based measures of health-related quality of life to be used in the evaluation of different health-care interventions. For one such measure, The EuroQol, designed to be used for these purposes, it was necessary to derive a single index value for each of the 243 health states it generates. The problem was that it was virtually impossible to generate direct valuations for all of these states, and thus it was necessary to find a procedure that allows the valuations of all EuroQol states to be interpolated from direct valuations on a subset of these.**

**METHODS. In a recent study, direct valuations were elicited for 42 EuroQol health states (using the time trade-off method) from a representative sample of the UK population. This article reports on the methodology that was adopted to build up a "tariff" of EuroQol values from this data.**

**RESULTS. A parsimonious model that fits the data well was defined as one in which valuations were explained in terms of the level of severity associated with each dimension, an intercept associated with any move away from full health, and a term that picked up whether any dimension in the state was at its most severe level.**

**CONCLUSIONS. The model presented in this article appears to predict the values of the states for which there are direct observations and, thus, can be used to interpolate values for the states for which no direct observations exist.**

**Key words: health status measurement, time trade-off, EuroQol. (Med Care 1997;35:1095–1108)**

Because resources for the provision of health care are scarce, choices have to be made about how they are allocated. It has been recognized that the impact on the health-related quality of life (HRQoL) of the population served should be an important consideration when making these choices. Because ultimately this will affect everybody, given that we are all potential patients, it will be necessary to establish the relative valuations attached to different states of health (defined in HRQoL terms) by members of the general public.

In developing a measure of HRQoL that was intended for this purpose, it is customary to describe health status in terms of certain aspects or dimensions. Several instruments have been designed to measure HRQoL, but with very different objectives.[1] Health states in the present study were defined in terms of the EuroQol Descriptive System, the *raison d'etre* of which was to provide a simple "abstracting" device for use alongside other more detailed measures of HRQoL.[2]

The choice of descriptive content in the EuroQol arose from a review of existing instruments and from the results of a survey of lay concepts of health.[3] In the five-dimen-

From the Department of Economics, University of Newcastle, Newcastle-Upon-Tyne, United Kingdom.

Address correspondence to: Paul Dolan, DPhil, Department of Economics, University of Newcastle, Newcastle-Upon-Tyne, NE1 7RU, United Kingdom.

sional set being discussed in this article, each dimension had three levels of "severity," broadly speaking, corresponding to "No problems" (level 1), "Some problems" (level 2), and "Extreme problems" (level 3), thus generating $3^5 = 243$ theoretically possible health states (Figure 1). For completeness, to these have to be added "unconscious" and "dead."

For the EuroQol to be used in evaluating the (HRQoL) benefits associated with different health-care interventions, it was necessary to derive a single index value for each of these health states (referred to as the "tariff"). The problem was that it was virtually impossible to generate direct valuations for all 243 EuroQol states. Therefore, a choice had to be made about how best to interpolate some of the values.

There were essentially two different approaches that could be adopted here: the "decomposed" approach and the "composite" approach. The former involves asking the respondent to value each level within a particular dimension, assuming that the levels of all other dimensions are held constant. Thus, the decomposed approach required few (and in some cases no) valuations for composite health states, although most studies that have adopted this approach have elicited valuations for a small subset of composite states.[4] Valuations for composite health states then could be generated by specifying a multiattribute function. Restrictions have to be imposed on the multiattribute if valuations for composite health states are to be derived: the least-restrictive model (in which the multiattribute function is multilinear) requires utility independence, which means that preferences for various levels of each dimension does not depend on the particular levels at which the other dimensions were fixed.[5]

The composite approach, which was adopted in this study, required each respondent to value a subset of composite health states. In principle, this approach places fewer restrictions on the resultant model because many (first-order and higher-order) interactions between dimensions and levels within dimensions can be taken into account. In practice, the resultant model often provides a highly simplified account of the data and thus is itself rather restrictive (see below).

Previous piloting showed that no one respondent could be expected to value more than approximately 13 states, but because this number was deemed to be too small to be representative of all possible EuroQol states, 42 states were chosen in total, and each respondent was asked to value a subset of these. In choosing the states both for use in the study itself and for each respondent, the most important consideration was that they should be spread widely over the valuation space so as to include as many combinations of levels across the five dimensions as possible. This was subject to the constraint that the states were likely to be considered plausible by respondents. Therefore, level 1 on usual activities (no problems) was not combined with level 3 on mobility (confined to bed) or with level 3 on self-care (unable to wash or dress self). Figure 2 shows the set of states chosen for direct valuation and how a subset of these were chosen for each respondent.

It was necessary, therefore, to find a procedure that allowed the valuations of all 243 EuroQol states to be interpolated from direct valuations on 42 of these. This article reports on the methodology adopted to address this issue and presents the results generated by it.

## The Data

The data reported here came from a survey designed to elicit the preferences of a representative sample of the noninstitutionalized adult population of England, Scotland, and Wales. In determining the size of the sample, there was the need for enough observations to be able to detect differences between the valuations given to different states. Although there was little evidence in the literature

Mobility

1. No problems walking about
2. Some problems walking about
3. Confined to bed

Self-Care

1. No problems with self-care
2. Some problems washing or dressing self
3. Unable to wash or dress self

Usual Activities

1. No problems with performing usual activities (e.g. work, study, housework, family or leisure activities)
2. Some problems with performing usual activities
3. Unable to perform usual activities

Pain/Discomfort

1. No pain or discomfort
2. Moderate pain or discomfort
3. Extreme pain or discomfort

Anxiety/Depression

1. Not anxious or depressed
2. Moderately anxious or depressed
3. Extremely anxious or depressed

Note: For convenience each composite health state has a five digit code number relating to the relevant level of each dimension, with the dimensions always listed in the order given above. Thus 11223 means:

1 No problems walking about
1 No problems with self-care
2 Some problems with performing usual activities
2 Moderate pain or discomfort
3 Extremely anxious or depressed

FIG. 1. The EuroQuol descriptive system.

about what size difference was required to be considered meaningful, it was decided that a 0.05 difference between health states was likely to be considered important in many contexts.[6] A sample size of 3,235 enabled such a difference to be detected at the 0.05 level of significance with 80% power. This required the selection of 6,080 addresses. The sample was drawn up by Social and Community Planning Research (SCPR) using the postcode address file.

In the event, 3,395 respondents were interviewed by 92 trained interviewers in their own homes between August and December 1993.[7] To enable modeling of the data at the individual level, only those respondents with complete valuations data were included in the analysis. There were 2,997 such

Each respondent valued 33333 and unconscious

plus

2 from 5 "very mild" states:

11112 11121 11211 12111 21111

plus

3 from 12 "mild" states:

11122 11131 11113 21133 21222 21312 12211 11133 22121 12121 22112 11312

plus

3 from 12 "moderate" states:

13212 32331 13311 22122 12222 21323 32211 12223 22331 21232 32313 22222

plus

3 from 12 "severe" states:

33232 23232 23321 13332 22233 22323 32223 32232 33321 33323 23313 33212

FIG. 2. Health states valued in the study.

respondents. Excluding those respondents with incomplete data did not compromise the representativeness of the sample.[8]

Valuations for health states can be elicited by a number of different methods.[9] Two that have been widely used are the standard gamble and the time trade-off (TTO). A recent (within-respondent) comparison of the two methods suggested that the TTO performed slightly better in terms of the internal consistency of the answers given by respondents, the sensitivity of valuations to parameters known to influence them, and the reliability of the responses when the valuation task was repeated by the same respondents some weeks later.[10] Thus, analysis in this article is based on valuations generated by the TTO method.

To gather valuable background information and as a warm-up exercise, each re-

spondent first was asked to describe their own health using the EuroQol descriptive system. They then were asked to rate a predetermined set of 15 health states (the 13 to be used in the TTO plus full health (11111) and "Immediate Death") on a vertical VAS with endpoints of 100 (best imaginable health state) and 0 (worst imaginable health state). It was explained that each state was to be regarded as lasting for 10 years without change, followed by death.

The 13 states (12 EuroQol states plus unconscious) then were valued by the TTO method using a specially designed double-sided board. One side was relevant for states that were regarded by the respondent as better than being dead, and the other side for states that were regarded as worse than being dead. In the former case, respondents were led by a process of "bracketing" to select a length of time (x) in the 11111 state that they regarded as equivalent to 10 years in the target state; the shorter the "equivalent" length of time, the worse the target state. Respondents were given an opportunity to refuse to trade-off any length of life to improve its quality. In the case of states worse than dead, the choice was between dying immediately and spending a length of time (10 - x) in the target state followed by x years in the 11111 state; the more time required in the 11111 state to compensate for a shorter time in the target state, the worse the target state.

If full health and dead are assigned scores of 1 and 0, respectively, then for states that are rated as better than dead on the TTO, scores are given by the formula $x/10$ where x is the number of years spent in full health. For states that are rated as worse than dead, the score is given by the formula $-x/(10-x)$. Thus, negative scores lie on a ratio (not an interval) scale and, unlike the case for states rated better than dead, are theoretically unbounded (though in this study, given the response categories available to respondents, they were bounded by -39). Problems associated with the asymmetry between positive and negative values are discussed below.

## Methods

The modeling in this article used a generalized least-squares (GLS) regression technique in which the functional form was additive. The dependent variable was defined as 1 - S where S is the value given to a particular health state. Besides the intercept, the specification of the remaining independent variables were derived from the ordinal nature of the EuroQol descriptive system. In total, three sets of dummy variables were created:

1. Two dummy variables for each dimension; one to represent the (assumed equal) move between levels and one to represent the move from level 2 to level 3 (this allows the effect of the move from level 1 to level 2 to be different from the effect of the move from level 2 to level 3).

2. Dummies to allow for possible (first order) interactions between dimensions.

3. Dummies to count the number of times a health state contains dimension(s) that are at level 1 or at level 3.

Figure 3 shows the independent variables used in the modeling. Note one further dummy (N3) that represents whether *any* of the dimensions is at level 3, of which more below.

Because the objective of this exercise was to estimate *one* preference-based EuroQol tariff for the *whole* community, respondent characteristics such as age, gender, and illness experience were not entered into the model. (Differences according to these characteristics may be important in some contexts, but analysis along these lines will be reported elsewhere.) The models were tested for possible interviewer effects using the joint test of significance (ie, including dummy variables for each interviewer in the regression equation and using the F-statistic to test for the significance of an interviewer effect). In all models, the effect of the interviewer was insignificant. This appeared to justify the decision to use a large number of interviewers in the study.

| Variable | Definition |
|---|---|
| a | Constant: associated with any move away from full health |
| MO | 1 if mobility is level 2; 2 if mobility is level 3; 0 otherwise |
| SC | 1 if self-care is level 2; 2 if self-care is level 3; 0 otherwise |
| UA | 1 if usual activities is level 2; 2 if usual activities is level 3; 0 otherwise |
| PD | 1 if pain/discomfort is level 2; 2 if pain/discomfort is level 3; 0 otherwise |
| AD | 1 if anxiety/depression is level 2; 2 if anxiety/dep. is level 3; 0 otherwise |
| M2 | 1 if mobility is level 3; 0 otherwise |
| S2 | 1 if self-care is level 3; 0 otherwise |
| U2 | 1 if usual activities is level 3; 0 otherwise |
| P2 | 1 if pain/discomfort is level 3; 0 otherwise |
| A2 | 1 if anxiety/depression is level 3; 0 otherwise |
| MOSC | The product of MO and SC |
| MOUA | The product of MO and UA |
| MOPD | The product of MO and PD |
| MOAD | The product of MO and AD |
| SCUA | The product of SC and UA |
| SCPD | The oroduct of SC and PD |
| SCAD | The Product of SC and AD |
| UAPD | The product of UA and PD |
| UAAD | The product of UA and AD |
| PDAD | The product of PD and AD |
| F11 | 1 if the health state contains 1 dimension at level 1; 0 otherwise |
| F21 | 1 if the health state contains 2 dimensions at level 1; 0 otherwise |
| F31 | 1 if the health state contains 3 dimensions at level 1; 0 otherwise |
| F41 | 1 if the health state contains 4 dimensions at level 1; 0 otherwise |
| F13 | 1 if the health state contains 1 dimension at level 3; 0 otherwise |
| F23 | 1 if the health state contains 2 dimensions at level 3; 0 otherwise |
| F33 | 1 if the health state contains 3 dimensions at level 3; 0 otherwise |
| F43 | 1 if the health state contains 4 dimensions at level 3; 0 otherwise |
| F53 | 1 if the health state contains 5 dimensions at level 3; 0 otherwise |
| N3 | 1 if any dimension is level 3; 0 otherwise |

FIG. 3. Definition of variables used in the modeling.

The approach adopted to model estimation followed the specific-to-general formulation in which simple models initially are estimated and new variables are added if necessary. This was deemed the most appropriate methodology for this type of data because the alternative, the general-to-specific approach, was more suited to time series data, in which any known collinearity between regressors could be taken into ac-

count better. (Problems of multicollinearity are discussed further below).

It was decided that analysis should take place on *individual-level* rather than aggregate-level data because it made the maximum use of the available data. In addition, the results of aggregate-level analysis were likely to be uninformative in that it was possible to find different models that fit the data equally well, with no objective way of choosing between them.

The asymmetry between positive and negative values posed problems for individual-level analysis because those respondents rating a state as worse than death would have a much greater impact on the model predictions than those respondents rating it as better than death. Patrick et al[11] transformed their negative values so that scores for states rated as worse than dead were bounded by -1, ie, symmetrical to the upper bound of +1 for states that are rated as better than dead. This transformation was justified on statistical grounds, but there is possibly a psychometric justification as well: that respondents may treat the scale for states worse than dead in the same way as they are assumed to treat the scale for states better than dead, ie, as an interval (not a ratio) scale. For these reasons, then, valuations for states worse than dead were transformed using the formula (x/10) - 1, where again x represented the number of years spent in full health.

Analysis at the individual level was complicated further by the fact that each respondent valued 12 EuroQol states, and, thus, it was reasonable to assume that these 12 scores were related to one another. This means that if a respondent gave one valuation that was lower than the population mean, then they were more likely to give a value lower than the population mean to the other states that they valued. This meant that the variance of the error term was likely to be partly determined by the individuals who valued the health states and was therefore unlikely to be constant. This violated one of the key assumptions underlying ordinary least-squares regression (OLS) and, thus, made this estimation procedure inefficient for this data.

The type of (GLS) model that addresses this issue is known as the *random effects* (RE) model, in which there is an overall intercept and an error term with two components; $e_{it}$ + $u_i$. The $e_{it}$ is the traditional error term unique to each observation. The $u_i$ is an error term representing the extent to which the intercept of the *i*th respondent differs from the overall intercept. This model assumes that the "individual specific" error term is normally independently distributed, which, given the size of the sample, seemed a valid assumption to make. Using the RE specification will reduce the possibility of drawing erroneous conclusions; for example, from an OLS estimation, it may be concluded that a particular respondent characteristic is an important determinant of the value attached to a health state, but this simply may be picking up an effect that will be nested within the RE model.

Using the RE model meant that limited dependent variable models, such as Tobit, were impractical because the functional form of these models with a RE component had yet to be specified. Clearly, the data here *was* constrained by an upper bound of 1 and a lower bound of -1, but the fact that the variance of the error term was almost certainly a function of the relationship between groups of values was deemed to be a more important characteristic of this data set. In any case, with such a large data set, the predicted values were unlikely to fall outside the range of possible values because less weight was given to outliers. Indeed, whether such values emerge or not can be used as a test of the validity of the RE model.

As a stringent test of the robustness of the models, each model has been estimated on a subsample of respondents (ie, an internal sample) and a comparison was made between the predicted values from this sub-

sample with the actual values of the remaining respondents (ie, an external sample). In this article, the internal sample was a randomly selected sample of two thirds of respondents, and the external sample constituted the remaining one third of respondents. In addition, because the purpose of the modeling exercise is to predict valuations for health states that were not included in the model-building process, each state has been excluded in turn, and the model was estimated using the remaining 41 states. The predicted value of the excluded state then could be compared to its actual value.

The modeling has been carried out using the LIMDEP statistical package.[12] When estimating the RE model, LIMDEP produces the OLS equation by default. It automatically performs a Lagrange Multiplier (LM) test, which is appropriate for large data sets like this. The LM test assesses whether the unrestricted model (ie, the RE one) represents an improvement on the restricted model (ie, the OLS one). If the LM value is significant ($P < 0.05$), then the RE model represents an improvement on the OLS one. The $R^2$ value produced from the RE model may not be any greater than that produced by the OLS model, and in some cases may be lower. It must be remembered, however, that OLS is the *only* estimation procedure that attempts to minimize the residual sum of squares; all other GLS models (of which RE is one) have a different objective function. Therefore, the $R^2$ values from the different models are not strictly comparable.

The models were tested for misspecification in two ways: a Ramsey RESET test, and a test for general heteroskedasticity. The reset test was undertaken in a two-stage process. In the first stage, the model was estimated, and the linear function was calculated. In the second stage, the square of the linear function was added to the equation, and the *t* statistic on this variable could be used as a test of the functional form of the original model.[13] The test for heteroskedasticity was created in a similar two-stage process. From the first stage, a new variable was created by squaring the value of the residuals. In the second stage, this new variable was regressed on the predicted values. The significance of the squared residual term could be used as a test for heteroskedasticity.

To make a choice between different ways of representing the relationship between the valuations of EuroQol health states and the different dimensions and levels, the model that ultimately was chosen had to predict a higher score for one state, A, than for another, B, if A was logically better than B on at least one dimension and no worse on any other dimension. In choosing between the many models that satisfy this consistency condition, the one that best explained the differences in the valuations given to those states on which there was direct data was chosen. For models with comparable goodness-of-fit statistics, the ultimate choice was made according to parsimony, ie, the simplest model (both in terms of the number of independent variables and the ability to explain them) was chosen. The results presented below are from the "best" model according to these criteria.

## Results

After testing many different models, one that fits the data well (in terms of goodness-of-fit statistics) *and* that was readily interpretable was a *main effects* model, in which each of the five dimensions was independent of others. None of the models that allowed for interactions between different dimensions improved the model significantly, and many introduced inconsistencies into the estimated values. The model did contain one further variable, however: an intercept dummy for whether *any* of the dimensions was at level 3. Without this additional dummy, which can be interpreted as reflecting the much greater disutility associ-

ated with "extreme problems," the residuals were related systematically to the predicted values in that the model underestimated the values of less severe states and overestimated the values of more severe ones.

Thus, the regression equation is as follows:

$$Y = \alpha + \beta_1 MO + \beta_2 SC + \beta_3 UA + \beta_4 PD + \beta_5 AD + \beta_6 M2 + \beta_7 S2 + \beta_8 U2 + \beta_9 P2 + \beta_{10} A2 + \beta_{11} N3$$

ie, TTO scores were explained by 12 independent variables: two variables for each dimension (one to represent the move from level 1 to level 2 and one to represent the move from level 2 to level 3), a term that picks up whether any dimension is at level 3, and an intercept (the interpretation of which is discussed below).

The coefficients on these variables for the full and internal samples are shown in Table 1. The $R^2$ of 0.46 (in both cases) was very high given the type of (cross-sectional) data analyzed here, and the results of the LM test indicated the RE specification to be a substantial improvement over the OLS model. In addition, the remarkable similarity between the parameter estimates for the whole sample and those for the internal sample suggested that the model is robust.

This model (and all other models) failed the RESET test, however, and suffered from general heteroskedasticity. That the model suffered from problems of omitted variables and/or incorrect functional form was not surprising given that the power of the RESET test increases as the sample size increases. Thus, with 2,997 × 12 = 35,964 observations, *any* model with relatively few independent variables was likely to be misspecified. The problems associated with heteroskedasticity were also difficult to overcome because the conventional means of dealing with them (eg, transformation of one or more independent variables) were not feasible given the (categorical) nature of the independent variables. In any event, heteroskedasticity was likely to result in inefficient rather than biased parameter estimates. Therefore, although misspecification was a problem, it was a problem about which little can be done.

TABLE 1. Parameter Estimates[a]

| Variable | Whole Sample | Internal Sample |
|---|---|---|
| *a* | 0.081 (10.35) | 0.075 (8.64) |
| MO | 0.069 (13.44) | 0.071 (10.21) |
| SC | 0.104 (19.23) | 0.105 (17.45) |
| UA | 0.036 (5.85) | 0.036 (4.64) |
| PD | 0.123 (23.92) | 0.121 (18.26) |
| AD | 0.071 (13.42) | 0.071 (11.76) |
| M2 | 0.176 (19.40) | 0.177 (16.03) |
| S2 | 0.006 (0.68) | 0.008 (0.66) |
| U2 | 0.022 (2.33) | 0.023 (1.76) |
| P2 | 0.140 (14.55) | 0.141 (12.97) |
| A2 | 0.094 (9.78) | 0.091 (7.18) |
| N3 | 0.269 (38.12) | 0.272 (31.19) |
| $R_2$ | 0.46 | 0.46 |
| LM Test | $P < 0.0001$ | $P < 0.0001$ |

[a] *t* statistics in parentheses.

Because the analysis concerns cross-sectional data, all variables from the main effects model have been left in the final equations, even those that might be considered "insignificant" (ie, have a *t* statistic whose absolute value was less than 1.96). This was to avoid any pretest type problems where insignificant variables may become "significant" if sampling were to be repeated. In addition, dropping variables whose absolute *t* statistic was less than 1 was likely to result in the mean square errors being higher than they should have been. Moreover, given that the regressors in this modeling were collinear, the significance of parameter estimates would vary according to the other independent variables in the equation. Because not enough was known about the nature of the functional form to address problems of multicollinearity with any degree of confidence, it was considered appropriate to in-

clude even insignificant main effects variables.

In computing the tariff from the model output, there was an issue relating to how the intercept, *a*, be interpreted. The strict statistical interpretation of *a* was that it represented the estimated value for (one minus) full health (ie, when all dummies take a value of zero, we have the estimated value for 11111). Thus, all estimated values should be rescaled by dividing them by 1 - *a*. Alternatively, given that by definition the value of 11111 was 1, we could interpret the intercept as representing any move away from full health. Thus, *a* could represent a discontinuity in the model between level 1 and level 2 in much the same way as the "N3" term represented a discontinuity between level 2 and level 3. In other words, we could interpret the intercept as picking up whether any dimension is at level 2, just as N3 picks up whether any dimension is at level 3. When predicted and actual values were compared, the algorithm in which *a* was treated in this way performed much better than when all estimated values were adjusted by 1 - *a*.

Table 1 shows that the constant was highly significant, suggesting that any move away from full health was associated with a substantial loss of utility. For the full sample, it can be seen that the largest decrement for a move from level 1 to level 2 was associated with pain or discomfort, some four times greater than that for the corresponding move on the usual activities dimension. Pain or discomfort continued to dominate the weighting for level 3, although mobility level 3 (confined to bed) was given a somewhat similar decrement. For the mobility, pain or discomfort, and anxiety or depression dimensions, the move from level 2 to level 3 was seen to involve a much greater decrement than the move from level 1 to level 2.

As an example of how the tariff was generated, consider the state 11223 estimated for the whole sample, shown in Table 2.

The actual (mean) and predicted values for the 42 states directly valued in the study, together with the differences between them, are given in Table 3. For only three states (21312, 23313, and 13332) did the difference between the mean and predicted value exceed 0.1 and the mean absolute difference (of 0.039) was considered acceptable.

Table 4 compares the predicted values generated from the internal two thirds of respondents with the actual (mean) values of the remaining one third of external respondents. It can be seen that the predictive power of the model remained high; only five states had a predicted value that was more than 0.1 different from the actual value, and the mean absolute difference was again below 0.05. Table 5 shows the predicted value for each state when direct values for that state are excluded from the modeling. Only three states had a predicted value that was more than 0.1 different from its actual value: the biggest difference was for the most extreme state (ie, 33333), which has a predicted value that is 0.173 below its actual value. The mean absolute difference was once again below 0.05.

Given a dataset of the kind analyzed here, there is a degree of uncertainty about the precise value that should be attached to any particular health state. One way of expressing this uncertainty is to calculate confidence intervals around the predicted values. The 95% confidence intervals are approxi-

TABLE 2. How the Preference-Based EuroQol Tariff Was Generated (State 11223)

| |
|---|
| Full health = 1.000 |
| Constant term (for any dysfunctional state): -0.081 |
| Mobility: level 1 -0 |
| Self-care: level 1 -0 |
| Usual activities: level 2 (1 × UA) -0.036 |
| Pain or discomfort: level 2 (1 × PD) -0.123 |
| Anxiety or depression: level 3 (2 × AD + 1 × A2) -0.236 |
| N3 (level 3 occurs within at least one dimension) -0.269 |
| Therefore, the estimated value for 11223 = 0.255. |

TABLE 3. Comparison of Estimated with Actual Values

| State | Actual Mean | Estimated | Estimated Mean |
|---|---|---|---|
| 2 1 1 1 1 | 0.878 | 0.850 | 0.028 |
| 1 1 2 1 1 | 0.869 | 0.883 | -0.014 |
| 1 2 1 1 1 | 0.834 | 0.815 | 0.019 |
| 1 1 1 2 1 | 0.850 | 0.796 | 0.054 |
| 1 1 1 1 2 | 0.829 | 0.848 | -0.019 |
| 1 2 2 1 1 | 0.767 | 0.779 | -0.012 |
| 1 2 1 2 1 | 0.742 | 0.692 | 0.050 |
| 1 1 1 2 2 | 0.722 | 0.725 | -0.003 |
| 2 2 1 2 1 | 0.645 | 0.623 | 0.022 |
| 2 2 1 1 2 | 0.662 | 0.675 | -0.013 |
| 1 1 3 1 2 | 0.552 | 0.485 | 0.067 |
| 2 2 1 2 2 | 0.540 | 0.552 | -0.012 |
| 2 1 3 1 2 | 0.536 | 0.416 | 0.120 |
| 2 1 2 2 2 | 0.553 | 0.620 | -0.067 |
| 1 2 2 2 2 | 0.551 | 0.585 | -0.034 |
| 2 2 2 2 2 | 0.500 | 0.516 | -0.016 |
| 1 3 2 1 2 | 0.389 | 0.329 | 0.060 |
| 1 3 3 1 1 | 0.346 | 0.342 | 0.004 |
| 1 1 1 1 3 | 0.392 | 0.414 | -0.022 |
| 1 1 1 3 1 | 0.200 | 0.264 | -0.064 |
| 1 2 2 2 3 | 0.216 | 0.151 | 0.065 |
| 2 1 3 2 3 | 0.160 | 0.128 | 0.032 |
| 2 3 3 2 1 | 0.147 | 0.150 | -0.003 |
| 3 2 2 1 1 | 0.152 | 0.196 | -0.044 |
| 2 1 2 3 2 | 0.064 | 0.088 | -0.024 |
| 2 2 3 2 3 | 0.042 | 0.024 | 0.018 |
| 1 1 1 3 3 | -0.049 | 0.028 | -0.077 |
| 2 2 3 3 1 | -0.011 | -0.003 | -0.008 |
| 2 3 3 1 3 | -0.070 | 0.037 | -0.107 |
| 3 3 2 1 2 | -0.022 | 0.015 | -0.037 |
| 2 3 2 3 2 | -0.084 | -0.126 | 0.042 |
| 2 1 1 3 3 | -0.063 | -0.041 | -0.022 |
| 3 3 3 2 1 | -0.120 | -0.095 | -0.025 |
| 3 2 3 1 3 | -0.152 | -0.098 | -0.054 |
| 2 2 2 3 3 | -0.142 | -0.181 | 0.039 |
| 3 2 2 2 3 | -0.174 | -0.163 | -0.011 |
| 3 2 2 3 2 | -0.223 | -0.261 | 0.038 |
| 1 3 3 3 2 | -0.228 | -0.115 | -0.113 |
| 3 2 3 3 1 | -0.276 | -0.248 | -0.028 |
| 3 3 2 3 2 | -0.332 | -0.371 | 0.039 |
| 3 3 3 2 3 | -0.386 | -0.331 | -0.055 |
| 3 3 3 3 3 | -0.543 | -0.594 | 0.051 |
| Mean absolute difference | | | 0.039 |

TABLE 4. Predicting the Values of an External Sample

| State | Mean of External Sample | Estimated from Internal Sample | Estimated Mean |
|---|---|---|---|
| 2 1 1 1 1 | 0.878 | 0.854 | 0.024 |
| 1 1 2 1 1 | 0.860 | 0.889 | -0.029 |
| 1 2 1 1 1 | 0.821 | 0.820 | 0.001 |
| 1 1 1 2 1 | 0.850 | 0.804 | 0.046 |
| 1 1 1 1 2 | 0.805 | 0.854 | -0.049 |
| 1 2 2 1 1 | 0.739 | 0.784 | -0.045 |
| 1 2 1 2 1 | 0.736 | 0.699 | 0.037 |
| 1 1 1 2 2 | 0.717 | 0.733 | -0.016 |
| 2 2 1 2 1 | 0.654 | 0.628 | 0.026 |
| 2 2 1 1 2 | 0.650 | 0.678 | -0.028 |
| 1 1 3 1 2 | 0.527 | 0.487 | 0.040 |
| 2 2 1 2 2 | 0.501 | 0.557 | -0.056 |
| 2 1 3 1 2 | 0.523 | 0.416 | 0.107 |
| 2 1 2 2 2 | 0.545 | 0.626 | -0.081 |
| 1 2 2 2 2 | 0.528 | 0.592 | -0.064 |
| 2 2 2 2 2 | 0.523 | 0.521 | -0.002 |
| 1 3 2 1 2 | 0.412 | 0.328 | 0.084 |
| 1 3 3 1 1 | 0.404 | 0.340 | 0.064 |
| 1 1 1 1 3 | 0.383 | 0.420 | -0.037 |
| 1 1 1 3 1 | 0.169 | 0.270 | -0.101 |
| 1 2 2 2 3 | 0.204 | 0.158 | 0.046 |
| 2 1 3 2 3 | 0.189 | 0.133 | 0.056 |
| 2 3 3 2 1 | 0.133 | 0.148 | -0.015 |
| 3 2 2 1 1 | 0.135 | 0.193 | -0.058 |
| 2 1 2 3 2 | 0.086 | 0.092 | -0.006 |
| 2 2 3 2 3 | 0.073 | 0.028 | 0.045 |
| 1 1 1 3 3 | -0.106 | 0.037 | -0.143 |
| 2 2 3 3 1 | 0.010 | -0.001 | -0.011 |
| 2 3 3 1 3 | -0.038 | 0.036 | -0.074 |
| 3 3 2 1 2 | -0.005 | 0.009 | -0.014 |
| 2 3 2 3 2 | -0.085 | -0.126 | 0.041 |
| 2 1 1 3 3 | -0.047 | -0.034 | -0.013 |
| 3 3 3 2 1 | -0.099 | -0.100 | -0.001 |
| 3 2 3 1 3 | -0.149 | -0.099 | -0.050 |
| 2 2 2 3 3 | -0.185 | -0.175 | -0.010 |
| 3 2 2 2 3 | -0.164 | -0.161 | -0.003 |
| 3 2 2 3 2 | -0.129 | -0.261 | 0.132 |
| 1 3 3 3 2 | -0.219 | -0.114 | -0.105 |
| 3 2 3 3 1 | -0.235 | -0.249 | 0.014 |
| 3 3 2 3 2 | -0.322 | -0.374 | 0.052 |
| 3 3 3 2 3 | -0.375 | -0.333 | -0.042 |
| 3 3 3 3 3 | -0.520 | -0.595 | 0.075 |
| Mean absolute difference | | | 0.046 |

TABLE 5. Predicting the Value of Each State When Direct Observations on that State are Excluded

| State | Actual Mean | Estimated | Estimated Mean |
|---|---|---|---|
| 2 1 1 1 1 | 0.878 | 0.834 | 0.044 |
| 1 1 2 1 1 | 0.869 | 0.901 | -0.032 |
| 1 2 1 1 1 | 0.834 | 0.803 | 0.031 |
| 1 1 1 2 1 | 0.850 | 0.776 | 0.074 |
| 1 1 1 1 2 | 0.829 | 0.853 | -0.024 |
| 1 2 2 1 1 | 0.767 | 0.784 | -0.017 |
| 1 2 1 2 1 | 0.742 | 0.681 | 0.061 |
| 1 1 1 2 2 | 0.722 | 0.727 | 0.000 |
| 2 2 1 2 1 | 0.645 | 0.621 | 0.024 |
| 2 2 1 1 2 | 0.662 | 0.677 | -0.015 |
| 1 1 3 1 2 | 0.552 | 0.456 | 0.096 |
| 2 2 1 2 2 | 0.540 | 0.557 | -0.017 |
| 2 1 3 1 2 | 0.536 | 0.447 | 0.089 |
| 2 1 2 2 2 | 0.553 | 0.655 | -0.102 |
| 1 2 2 2 2 | 0.551 | 0.599 | -0.048 |
| 2 2 2 2 2 | 0.500 | 0.528 | -0.028 |
| 1 3 2 1 2 | 0.389 | 0.295 | 0.094 |
| 1 3 3 1 1 | 0.346 | 0.337 | 0.009 |
| 1 1 1 1 3 | 0.392 | 0.412 | -0.020 |
| 1 1 1 3 1 | 0.200 | 0.287 | -0.087 |
| 1 2 2 2 3 | 0.216 | 0.118 | 0.098 |
| 2 1 3 2 3 | 0.160 | 0.125 | 0.035 |
| 2 3 3 2 1 | 0.147 | 0.155 | -0.012 |
| 3 2 2 1 1 | 0.152 | 0.217 | -0.065 |
| 2 1 2 3 2 | 0.064 | 0.099 | -0.035 |
| 2 2 3 2 3 | 0.042 | 0.022 | 0.020 |
| 1 1 1 3 3 | -0.049 | 0.049 | 0.000 |
| 2 2 3 3 1 | -0.011 | 0.011 | 0.022 |
| 2 3 3 1 3 | -0.070 | 0.068 | -0.002 |
| 3 3 2 1 2 | -0.022 | 0.001 | -0.023 |
| 2 3 2 3 2 | -0.084 | -0.140 | 0.056 |
| 2 1 1 3 3 | -0.063 | -0.037 | -0.027 |
| 3 3 3 2 1 | -0.120 | -0.141 | 0.021 |
| 3 2 3 1 3 | -0.152 | -0.076 | -0.076 |
| 2 2 2 3 3 | -0.142 | -0.197 | 0.055 |
| 3 2 2 2 3 | -0.174 | -0.157 | -0.017 |
| 3 2 2 3 2 | -0.223 | -0.270 | 0.047 |
| 1 3 3 3 2 | -0.228 | -0.065 | -0.163 |
| 3 2 3 3 1 | -0.276 | -0.224 | -0.052 |
| 3 3 2 3 2 | -0.332 | -0.384 | 0.052 |
| 3 3 3 2 3 | -0.386 | -0.314 | -0.072 |
| 3 3 3 3 3 | -0.543 | -0.716 | 0.173 |
| Mean absolute difference | | | 0.048 |

mately 0.75 for all states, irrespective of their severity. Thus, the confidence intervals (even the 50% one, which is 0.26) are large, indicating (as the standard deviations around the mean values indicate) that different people attached very different valuations to the same health state.

Confidence intervals do not tell the whole story, however, because it is quite plausible that respondents ranked adjacent states in the same way, but some did so using high values, whereas others did so using low values. The significance of the LM test (which compares the RE specification with the OLS one) suggests that this was indeed the case. Moreover, previous analysis of pairwise relationships between states revealed that there were no more than four states adjacent to any particular state that were *not* significantly different from it at the 1% level.[8]

## Discussion

The statistical analysis used in this article to interpolate valuations for all 243 EuroQol health states from direct valuations on a subset of 42 states was based on regression analysis in which the dependent variable was (one minus) the score given to the health states. All independent variables were dummies that derive from the ordinal nature of the EuroQol descriptive system. The functional form was a linear additive one, which seems a valid approach given the assumption that valuations elicited from the TTO method for states rated as better than dead exhibit interval scale properties (ie, the difference between 0.2 and 0.4 is the same as the difference between 0.6 and 0.8). Besides, estimating and interpreting different functional forms would be difficult given the (categorical) nature of the independent variables.

Before estimation could take place, there was the question of how to score states worse than death. Given the standard health preference scale, states preferred to death

are limited by an upper bound of +1; however, there is no comparable lower bound for states worse than death, which in this study could take a value as low as -39. As Torrance[14] noted, "this issue of large negative values and what to do about them needs much more study." Partly to overcome this problem, but partly to account for the fact that respondents were unlikely to be interpreting the scale in the way implied by theory, negative scores were transformed such that they were bounded by -1.

Analysis was of data at the individual level to make full use of the data available and was based on a form of GLS known as the RE model. This specification accounted for the fact that groups of observations came from one individual. An alternative approach would have been a fixed effects (FE) model, in which a dummy variable would be created for each respondent.

Models based on RE rather than on fixed effects were deemed more appropriate for this data set for a number of reasons. First, fixed effect models, which produce results that are conditional on the units in the data set, are only reasonable if the data exhaust the population. If the data are a sample of a larger population (as was the case here), and if we wish to draw inferences about other members of that population (as also was the case here), then "the fixed effects model is no longer reasonable; in this context, use of the random effects model has the advantage that it saves a lot of degrees of freedom."[15] Second, given that we cannot fully account for how and why valuations differ between individuals, it was reasonable to treat this type of ignorance in a fashion similar to the general ignorance represented by the error term. Finally, there was the practical problem of estimating and interpreting 2,997 coefficients.

Given that each respondent in this data set valued a number of different health states, it was expected that the RE models would represent an improvement over OLS. The results of all the Lagrange Multiplier tests confirmed this, although it should be noted that the beta coefficients (though not the standard errors) from the OLS equations were similar to those estimated by the RE models. Therefore, it appears for this dataset that an OLS specification produced unbiased yet inefficient estimates.

Besides the random effects specification, the model is simple; the data is explained in terms of a main effects model with one additional term to account for the much greater disutility associated with having extreme problems. In other words, although a simple additive model was too restrictive a model to provide adequate explanation of the data reported here, a model with only one interaction term (which is a function of the level of each dimension rather than of the particular dimensions themselves) was adequate.

The results from this modeling appear encouraging. The $R^2$ value (of 0.46) can be considered very good given the type of data analyzed here. There is very little data with which a direct comparison of these results could be made because much of the analysis of health state valuations data had been performed on aggregate level data, but, in a wider context, a number of econometric models, notably those concerning labor supply functions, report "robust" findings with $R^2$ values as low as 0.1.

In addition, the predicted values from this model were very close to the actual ones for the majority of EuroQol states, and the mean absolute difference (of 0.039) is unlikely to be considered meaningful in many contexts.[6] When the values of a randomly chosen two thirds of respondents were used to estimate the values of the remaining one third and when the direct values of each state were excluded in turn and the value for that state estimated from the remaining data, the mean absolute difference was less than 0.05 in both cases.

The model presented in this article appears to predict the *mean* values of the EuroQol states for which there are direct observations and, thus, can be used to interpolate

*social* values for the states for which no direct observations exist. The large confidence intervals around the point estimates warn against the values being used at the individual level, although we can be reasonably confident that most individuals would have the same *ordering* of states as that implied by the tariff values.

## Acknowledgments

The author thanks colleagues Claire Gudex, Paul Kind, and Alan Williams for their help and support in producing this article. The author also thanks Stuart Peacock, Nigel Rice, and Matt Sutton for their statistical help, and John Brazier for comments made on earlier drafts of this paper. The financial support of the Department of Health is gratefully acknowledged.

## References

1. Bowling A. Measuring health: A review of quality of life measurement scales. Milton Keynes: Open University Press, 1991.

2. The Euroqol Group. Euroqol: A new facility for the measurement of health related quality of life. Health Policy 1990;16:199.

3. van Dalen H, Williams A, Gudex C. Lay people's evaluations of health: Are there variations between different subgroups? J Epidemiol Community Health 1994;48:249.

4. Torrance GW, Furlong W, Feeny D, Boyle M. Multi-attribute preference functions: Health utilities index. Pharmacoeconomics 1995;7:503.

5. Froberg DG, Kane RL. Methodology for measuring health state preferences-I: Measurement strategies. J Clin Epidemiol 1989;42:344.

6. O'Brien BJ, Drummond MF. Statistical versus quantitative significance in the socioeconomic evaluation of medicines. PharmacoEconomics 1994;5:389.

7. Erens R. Health-related quality of life: General population survey. Technical Report. London: SCPR, 1994.

8. Dolan P, Gudex C, Kind P, Williams A. The time trade-off: Results from a general population study. Health Economics 1996;5:141.

9. Torrance GW. Measurement of health state utilities for economic appraisal: A review. J Health Econ 1986;5:1.

10. Dolan P, Gudex C, Kind P, Williams A. Valuing health states: A comparison of methods. J Health Econ 1996;15:209.

11. Patrick DL, Starks HE, Cain KC, Uhlmann RF, Pearlman RA. Measuring preferences for health states worse than death. Med Decis Making 1994;14:9.

12. Greene WH. LIMDEP version 6.0: User's manual and reference guide. New York, NY: Econometric Software, Inc., 1992.

13. Ramsey JB. Tests for specification errors in classical linear least squares regression models. Journal of the Royal Statistical Society B 1969;31:350.

14. Torrance GW. Health states worse than death. Presented at the Third International Conference on System Science in Health Care; 1984; Berlin, Germany.

15. Kennedy P. A guide to econometrics. Oxford: Blackwell Scientific Publications, 1992.