Less is more: cross-validation testing of simplified non-linear regression model specifications for EQ-5D-5L health state values

Kim Rand-Hendriksen, PhD\textsuperscript{1,2}; Juan Manuel Ramos-Goñi, MSc\textsuperscript{3,4}; Liv Ariane Augestad, PhD, MD\textsuperscript{1,2}; Nan Luo, PhD\textsuperscript{5}

1. Health Services Research Centre,
   Akershus University Hospital,
   Lørenskog,
   Norway

2. Dept. of Health Management and Health Economics,
   University of Oslo,
   Oslo,
   Norway

3. Executive Office. EuroQol Research Foundation,
   Rotterdam,
   The Netherlands

4. Red de Investigación en Servicios de Salud en Enfermedades Crónicas (REDISSEC),
   Spain

5. Saw Swee Hock School of Public Health,
   National University of Singapore,
   Singapore

Key words: EQ-5D, regression models, valuation, QALY, non-linear, cross-validation

Running title: Simplified EQ-5D-5L regression models
Funding:

This study was funded by the EuroQol Group. All four authors are members of the EuroQol Group, and receive reimbursement for travel to EuroQol meetings and compensation for work done for the Group. No other conflicts of interests.

Corresponding author

Kim Rand-Hendriksen, Cand.Psychol, PhD,

Health Services Research Centre,
Akershus University Hospital,
ørørenskog,
Norway

Dept. of Health Management and Health Economics,
University of Oslo,
Oslo,
Norway

Tel: +47 40220366
E-mail: kim.rand-hendriksen@medisin.uio.no

Complete author information:

Kim Rand-Hendriksen, Cand.Psychol, PhD,

Health Services Research Centre,
Akershus University Hospital,
ørenskog,
Norway

Dept. of Health Management and Health Economics,
University of Oslo,
Oslo,
Norway

Tel: +47 40220366
E-mail: kim.rand-hendriksen@medisin.uio.no

Juan Manuel Ramos-Goñi

Executive Office. EuroQol Research Foundation,
Rotterdam,
The Netherlands

Tel: +34 670757100
E-mail: jramos@euroqol.org

Liv Ariane Augestad, MD, PhD,
Dept. of Health Management and Health Economics,
University of Oslo,
Oslo,
Norway

Health Services Research Centre,
Akershus University Hospital,
Lørenskog,
Norway

Tel: +47 45285389
E-mail: l.a.augestad@medisin.uio.no

Nan Luo, PhD,

Saw Swee Hock School of Public Health,
National University of Singapore,
16 Medical Drive, Block MD3,
Singapore 117597
Tel: +65 6516 4966
Fax: (+65) 6779 1489
Email: nan_luo@nuhs.edu.sg

Acknowledgements:

The study received financial support from EuroQol Research Foundation, grant 2015070. The EuroQol Group had no involvement in the study. The views expressed in this manuscript are those of the authors, and should not be taken to reflect the views of the EuroQol Group.

Words in abstract: 237
Words in main document: 3874
Figures: 2
Tables: 3
Online appendix: 1
Less is more: cross-validation testing of simplified non-linear regression model specifications for EQ-5D-5L health state values

Words in abstract: 237

Words in main document: 3874

Figures: 2

Tables: 3

Online appendix: 1
Abstract

Background

The conventional method for modeling of EQ-5D-5L health state values in national valuation studies is an additive 20-parameter main-effect regression model. Statistical models with many parameters are at increased risk of overfitting; fitting to noise and measurement error, rather than the underlying relationship.

Objective

To compare the 20-parameter main-effect model to simplified, non-linear, multiplicative regression models in terms of how accurately they predict mean values of out-of-sample health states.

Methods

We used data from the Spanish, Singaporean, and Chinese EQ-5D-5L valuation studies. Four models were compared: an 8-parameter model with single parameters per dimension, multiplied by cross-dimensional parameters for levels 2, 3, and 4; 9 and 11-parameter extensions with handling of differences in the wording of level 5; and the “standard” additive 20-parameter model. Fixed and random intercept variants of all models were tested using two cross-validation methods: leave-one-out at the level of valued health states, and of health state blocks used in EQ-5D-5L valuation studies. Mean absolute error, Lin’s Concordance Correlation Coefficient, and Pearson’s R between observed health state means and out-of-sample predictions were compared.

Results

Predictive accuracy was generally best using random intercepts. The 8, 9, and 11-parameter models outperformed the 20-parameter model in predicting out-of-sample health states.

Discussion and conclusion

Simplified non-linear regression models look promising, and should be investigated further using other EQ-5D-5L datasets. To reduce the risk of overfitting, cross-validation is recommended to inform model selection in future EQ-5D valuation studies.

Key words: EQ-5D, regression models, valuation, QALY, non-linear, cross-validation
Background:

The EQ-5D is the most used instrument for measuring quality adjusted life years (QALYs) worldwide. The instrument has two components: a descriptive system used for self-rating of health status, and a value set/value algorithm that describes how many QALYs a person will accrue by being in any described health state for a year. The EQ-5D describes health along five dimensions: mobility (MO), self-care (SC), usual activities (UA), pain/discomfort (PD), and anxiety/depression (AD). Two versions exist: the original, in which each dimension is described at three levels (corresponding roughly to “no”, “moderate”, and “extreme” problems), now referred to as EQ-5D-3L; and a new version, in which each dimension is described at five levels (corresponding to “no”, “slight”, “moderate”, “severe”, and “extreme” problems) referred to as EQ-5D-5L. Value sets for EQ-5D-3L have typically been derived in national valuation studies, in which subsets of the possible combination health states were valued by general population samples using the time trade-off (TTO) valuation technique. Currently, the valuation studies for the EQ-5D-5L use a standardized data collection protocol including two valuation techniques: a TTO variant called composite TTO (C-TTO), plus a discrete choice experiment (DCE). In this study, we are exclusively concerned with the modeling of values derived from C-TTO valuation. A total of 86 EQ-5D-5L health states are valued using C-TTO, with each respondent valuing one block of 10 health states. This C-TTO design was built to predict a value set using a regression model containing a core of 20 dummy variables, representing levels 2, 3, 4, and 5 (“slight”, “moderate”, “severe”, and “extreme” problems) for each of the five dimensions. The intercept is typically tacit in the literature on EQ-5D modeling. In line with the convention in the literature, we do not count the intercept when referring to the number of parameters in regression models in this paper.

There are three possible drawbacks with the standard 20-parameter additive regression model: the relatively large number of fitted parameters (20 or more) comes with the risk of random variance and measurement error substantially influencing the fitted models, a phenomenon referred to as overfitting. Generally, the risk of overfitting increases with model complexity (i.e. greater number of parameters), and is associated with lack of power. Overfitting reduces predictive accuracy, and may compound when fitted models are applied beyond the scope of the observed data. Second, using a large set of independent predictors leaves the model susceptible to non-monotonicity; fitted coefficients that are logically inconsistent with the structural hierarchy of the descriptive
system, suggesting that specific impairments are improvements. Non-monotonicity has been an issue in several EQ-5D-5L valuation studies. While it has been shown that the non-monotonic value sets may be explained at least in part by interviewer effects and insufficient quality control, the susceptibility of different modeling approaches to non-monotonicity varies, and tends to increase with model complexity. Third, while a strength of the 20-parameter model is that it allows the relative distance between levels to vary over dimensions, this could be also a weakness, since it disregards the descriptive similarity across dimensions; all five dimensions have five levels, described in a similar fashion, suggesting that they could have similar relative severity distances.

The development of the labeled scales for the EQ-5D-5L involving a series of studies, initially in English and Spanish, and later also in French and Chinese. Through these studies, the final labels were selected from a larger initial pool of possible labels based on several criteria, including approximate equidistance in severity rated by lay persons in a visual analogue scale task (medians close to 25th, 50th, and 75th percentile for level labels 2, 3, and 4), consistency in relative severity across dimensions, consistency in ranking, availability in colloquial language, and subjective reporting of how easy they were to understand. Following the selection of labels, follow-up studies in English, Spanish, French, and Chinese indicate that the labels are perceived as describing relatively equal intervals across the five dimensions.

The question we ask in this study is whether the regression modeling of values for EQ-5D-5L health states could capitalize on the shared level and label structure across the 5 dimensions. If the general population perceives the 5 dimension scales to have close to proportional relative severity distances between levels, as indicated by the studies conducted to test the level labels, value algorithms could potentially be generated using substantially fewer parameters. In the simplest case, a single parameter per dimension, interpretable as the disutility of the dimension at level 5, could be multiplied by level parameters shared across all dimensions, for levels 2, 3, and 4 for a total of 8 parameters. Specified in this manner, the parameters for levels 2, 3, and 4 are interpretable as proportions of the disutility of level 5. The 9- and 11-parameter extensions of this model are described in the methods section. We will refer to this type of non-linear regression model, in which estimated parameters are multiplied with each other, as being multiplicative. The described 8-parameter model would have 4 to 5 times more observations per coefficient (dimension parameters and level parameters, respectively) than the 20-parameter model when applied to the same datasets.
This kind of model has been discussed as a concept for a number of years. The first written reference we know of is in a paper presented at the EuroQol plenary meeting in 2012, in which Oppe, van Hout, and Ramos-Goñi performed analyses using a 9-parameter model which is equivalent to the 8-parameter model presented here, and a 10-parameter model equivalent to the 9 parameter model presented later in this paper. Our models have one less parameter because the models presented by Oppe, van Hout and Ramos-Goñi are not uniquely identifiable. Details can be found in the online appendix document, section A.

Model selection is a complex issue, and multiple criteria may be relevant, including performance in terms of various goodness-of-fit measures, monotonicity, and parsimony. The assumption that the distance between levels is close to proportional across dimensions is, arguably, a strong one. Essentially, it presupposes that observed deviations from proportionality are spurious. This assumption excludes several regular fit statistics when reported for the full set of observations (i.e. internal validation), including the coefficient of determination ($r^2$), mean absolute error (MAE), and mean square error (MSE). When reported for the observed data, these statistics invariably favor increased model complexity, and are uninformative as to predictive ability beyond the observed data, and thus increase the risk of overfitting. Other fit statistics, most notably the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) have been developed in order to penalize model complexity, and thus reduce the risk of overfitting. However, AIC and BIC are directly based on log-likelihood estimates, which are only comparable within modeling framework (e.g. mixed effect models cannot be directly compared with models without mixed effects). Cross-validation methods can be tailored to compare models across frameworks, given stringent criteria for comparing predictions to left-out observations. Due to the limitations of AIC and BIC when it comes to comparing mixed-effect and fixed-effect models, we decided to use cross-validation to compare model performance in this study.

Given that the aim of the modeling exercise in valuation studies is to generate the best possible estimates of the population means of values assigned to all describable health states, we can compare the relative merits of the 20-parameter model and the simpler multiplicative specifications using cross-validation methods; sequentially splitting the data into two parts, fitting the models of interest on one part, predicting values in the second part, and repeating until the whole dataset has been left out and predicted. Added model complexity is only worthwhile insofar as it results in improved predictions for non-observed data. Further, model performance could be assessed in
terms of the ability to accurately predict mean values, rather than the ability to predict values assigned by individual valuation study participants. If the multiplicative models are too restrictive, and the assumption of similar relative severity of levels across dimensions does not hold, the constrained models would be underfitted, resulting in less accurate predictions for unobserved health state means than the 20-parameter model. Conversely, if the 20-parameter model suffers from overfitting, the multiplicative models should make better predictions over unobserved health state means. Importantly, these tests are suited to falsify the hypothesis that using constrained models is a good idea; if the assumption that level distances are proportional across dimensions does not hold, the 20-parameter model should produce better out-of-sample predictions.

The objective of this study was to use cross-validation to compare the additive 20-parameter model to 8, 9, and 11-parameter multiplicative models in terms of their ability to accurately predict the mean values for left-out EQ-5D-5L health states using C-TTO data from three existing EQ-5D-5L valuation studies. All four models were tested using both fixed and random intercepts at the level of individual study participants.

Methods:

Data:

We used C-TTO data from the Chinese, Spanish, and Singaporean EQ-5D-5L valuation studies performed between 2012 and 2015. In all three studies, respondents were randomized to one of 10 health state blocks, each consisting of 10 EQ-5D-5L health states. All blocks included the worst state (55555), one of the five mildest health states, e.g. 11112, and 8 unique health states, such that 86 unique health states were valued in total. The order of presentation of the states was individually randomized. The specifics of the standardized C-TTO procedure and interview protocol is described elsewhere.

Models:

We tested four different regression models, each with fixed and random intercepts at the level of individual study participants. To distinguish between fixed and random intercept variants, we append the letters “F” (fixed) and “R” (random) to the model names, e.g. ADD20R and MULT11F.

The standard, additive 20-parameter model, referred to as ADD20, has parameters representing levels 2, 3, 4, and 5 for each dimension. Let $\alpha$ represent the intercept and $x_{dl}$ represent a dummy variable indicating the presence of
problems on dimension \( d \) at level \( l \). \( \beta_{dl} \) is the coefficient representing the estimated disutility of having problems on dimension \( d \) at level \( l \). As an example, \( \beta_{MO3} \) is the coefficient representing the disutility of having moderate problems (level 3) on mobility.

\[
ADD20: y = \alpha + \sum_i \sum_d \beta_{dl} x_{dl} + e = \alpha + \\
\beta_{MO2} x_{MO2} + \beta_{SC2} x_{SC2} + \beta_{UA2} x_{UA2} + \beta_{PD2} x_{PD2} + \beta_{AD2} x_{AD2} + \\
\beta_{MO3} x_{MO3} + \beta_{SC3} x_{SC3} + \beta_{UA3} x_{UA3} + \beta_{PD3} x_{PD3} + \beta_{AD3} x_{AD3} + \\
\beta_{MO4} x_{MO4} + \beta_{SC4} x_{SC4} + \beta_{UA4} x_{UA4} + \beta_{PD4} x_{PD4} + \beta_{AD4} x_{AD4} + \\
\beta_{MO5} x_{MO5} + \beta_{SC5} x_{SC5} + \beta_{UA5} x_{UA5} + \beta_{PD5} x_{PD5} + \beta_{AD5} x_{AD5} + e
\]

The variant with random intercept at the level of individual participants (ADD20R) corresponds to the method described as random effects or generalized least squares (GLS) in several EQ-5D valuation studies e.g.,22.

The simplest multiplicative model is the 8-parameter model named MULT8. This is a constrained variant of the ADD20-model, in which five parameters representing the disutility of having problems at level 5 on each of the five dimensions (\( \beta_{MO5}, \beta_{SC5}, \beta_{UA5}, \beta_{PD5}, \beta_{AD5} \)) are multiplied by parameters for levels 2, 3, and 4 (\( L_2, L_3, L_4 \)). Thus, the disutility of being at level 3 on mobility is \( \beta_{MO3} \times L_3 \). Note that \( x_{dl} \) still represents the dummy variable representing the presence of problems on dimension \( d \) at level \( l \).

\[
MULT8: y = \alpha + \sum_i (\sum_d \beta_{dl} x_{dl}) L_i + e = \alpha + \\
(\beta_{MO2} x_{MO2} + \beta_{SC2} x_{SC2} + \beta_{UA2} x_{UA2} + \beta_{PD2} x_{PD2} + \beta_{AD2} x_{AD2})L_2 + \\
(\beta_{MO3} x_{MO3} + \beta_{SC3} x_{SC3} + \beta_{UA3} x_{UA3} + \beta_{PD3} x_{PD3} + \beta_{AD3} x_{AD3})L_3 + \\
(\beta_{MO4} x_{MO4} + \beta_{SC4} x_{SC4} + \beta_{UA4} x_{UA4} + \beta_{PD4} x_{PD4} + \beta_{AD4} x_{AD4})L_4 + \\
\beta_{MO5} x_{MO5} + \beta_{SC5} x_{SC5} + \beta_{UA5} x_{UA5} + \beta_{PD5} x_{PD5} + \beta_{AD5} x_{AD5} + e
\]

MULT9 extends MULT8, with an additional parameter \( L_5 \) to distinguish level 5 for PD and AD, described using the label “extreme”, from level 5 for MO, SC, and UA, described using “unable to”. Thus, MULT9 assumes that the
relative distance between the levels is shared across dimensions, with the exception of the distance between levels 4 and 5, which is shared across the three first and two last dimensions only.

\[ \text{MULT9: } y = \alpha + \]
\[ (\beta_{x_{M0}x_{M2}} + \beta_{x_{S}x_{C2}} + \beta_{U_{A}x_{U2}} + \beta_{P_{D}x_{P2}} + \beta_{A_{D}x_{A2}})L_{2} + \]
\[ (\beta_{x_{M0}x_{M3}} + \beta_{x_{S}x_{C3}} + \beta_{U_{A}x_{U3}} + \beta_{P_{D}x_{P3}} + \beta_{A_{D}x_{A3}})L_{3} + \]
\[ (\beta_{x_{M0}x_{M4}} + \beta_{x_{S}x_{C4}} + \beta_{U_{A}x_{U4}} + \beta_{P_{D}x_{P4}} + \beta_{A_{D}x_{A4}})L_{4} + \]
\[ (\beta_{x_{M0}x_{M5}} + \beta_{x_{S}x_{C5}} + \beta_{U_{A}x_{U5}}) + (\beta_{P_{D}x_{P5}} + \beta_{A_{D}x_{A5}})L_{5} + e \]

MULT11 assumes that the difference in description of level 5 (“unable” vs. “extreme”) means that the relative severity distances between all 5 levels are different for the two sets of dimensions. Therefore, two sets of parameters are needed for each of levels 2, 3, and 4; one for the first three dimensions \((L_{U_{2}}, L_{U_{3}}, L_{U_{4}})\), and one for the last two \((L_{E_{2}}, L_{E_{3}}, L_{E_{4}})\). Unlike MULT9, no parameters are shared between the first three and the last two dimensions. Therefore, separate level 5 parameters are not necessary.

\[ \text{MULT11: } y = \alpha + \]
\[ (\beta_{x_{M0}x_{M2}} + \beta_{x_{S}x_{C2}} + \beta_{U_{A}x_{U2}})L_{U_{2}} + (\beta_{P_{D}x_{P2}} + \beta_{A_{D}x_{A2}})L_{E_{2}} + \]
\[ (\beta_{x_{M0}x_{M3}} + \beta_{x_{S}x_{C3}} + \beta_{U_{A}x_{U3}})L_{U_{3}} + (\beta_{P_{D}x_{P3}} + \beta_{A_{D}x_{A3}})L_{E_{3}} + \]
\[ (\beta_{x_{M0}x_{M4}} + \beta_{x_{S}x_{C4}} + \beta_{U_{A}x_{U4}})L_{U_{4}} + (\beta_{P_{D}x_{P4}} + \beta_{A_{D}x_{A4}})L_{E_{4}} + \]
\[ \beta_{x_{M0}x_{M5}} + \beta_{x_{S}x_{C5}} + \beta_{U_{A}x_{U5}} + \beta_{P_{D}x_{P5}} + \beta_{A_{D}x_{A5}} + e \]

Note that all the multiplicative models are restricted variants of the 20-parameter model, with fewer degrees of freedom; it is possible to derive 20 values equivalent to the coefficients of the 20-parameter model from them. Since all the multiplicative models have fewer degrees of freedom than ADD20, model fit on the complete set of data should logically be best for ADD20.
Cross-validation to test prediction accuracy

In the absence of external data for comparison, we used cross-validation techniques; sequentially leaving out part of the data, using the remainder to fit a model, and making predictions for the left-out data.

We applied two cross-validation methods to each of the three national datasets. For both, we calculated the mean absolute error (MAE), Pearson’s product moment correlation, and Lin’s concordance correlation coefficient (CCC) between observed and predicted mean values. All models were also tested for the presence of non-monotonicity.

Cross-validation method 1 - Leave-one-out by health state:

1. We sequentially removed all observations for each of the 86 health states, fit all the models on the remaining 85 health states, and predicted the value for the left-out state. Leave-out predictions for the 86 health states were then compared to the observed mean values for the same states.

Cross-validation method 2 – Leave-out by blocks of health states:

Health states in a specific health state block were all valued by the same set of individuals, and were not fully independent; when removing all observations of a single health state, information on how the same set of individuals valued 9 other health states remains observed. Thus, prediction of the left out state may be better than if we were to consider values for a health state fully external to the dataset. Since a total of 3039 EQ-5D-5L health states are not directly observed in any of the 10 blocks, it may be more relevant to assess the predictive ability of the models using cross-validation in which complete health state blocks were sequentially left out, rather than individual health states.

2. Leave-out by health state blocks:

We sequentially removed all observations for each of the 10 health state blocks, fit the model on the remaining 9 blocks, and predicted values for all states in the left-out block.

In addition to these cross-validation methods, description and results of a more complex leave-out operation tailored to isolate the impact of removing 10% of the data (as in method 2), while performing predictions within observed blocks, and a split-half cross-validation test can be found in the online appendix document, section D.
Software and code

All tests were performed using the R statistical package, version 3.3.0\textsuperscript{26}. Detailed description of the data structure, and the code used to fit each model in R can be found in online appendix document, sections B and C, with equivalent STATA\textsuperscript{27} code added where possible. The full set of functions used to perform the analyses presented here is available upon request.

Results:

Table 1 presents demographic characteristics of the study participants in the three studies, as well as their self-ratings on a visual analogue scale. The ratio of male to female participants is relatively even in all three study populations, ages between 20 and 60-70 years are well represented. For further descriptive details, we refer to the published and upcoming papers on the individual valuation studies.

[TABLE 1 ABOUT HERE]

The differences between the models were relatively small, and the most different models, ADD20R and MULT8R produce very similar predictions (Figure 1).

[FIGURE 1 ABOUT HERE]

Figure 2 is a graphical representation of the relationship between the coefficients of the 20-parameter and 8-parameter models, illustrated with fitted parameters from ADD20F and MULT8R on the Chinese data. The figure shows how the 8-parameter models are constrained variants of the 20-parameter model: the intersections between the vertical lines, representing the dimension parameters of MULT8R, and the diagonal lines, representing the level parameters, correspond to the interpretation of the coefficients of the 20-parameter models.

[FIGURE 2 ABOUT HERE]

In terms of MAE, the models with random intercept resulted in more accurate predictions for health state means than their fixed-intercept counterparts, with three exceptions (Table 2). For the Spanish data, MULT9F and MULT9R were equally accurate using cross-validation method 1, and both MULT9F and MULT11F were more accurate than MULT9R and mult11R using cross-validation method 2.
ADD20F and ADD20R were outperformed or matched in terms of MAE by MULT8R, MULT9R, and MULT11R in all three countries and using both cross-validation methods. Pearson’s R and Lin’s CCC were very similar for the random intercept variants of ADD20 and MULT8, 9, and 11.

[TABLE 2 ABOUT HERE]

Of the multiplicative models, MULT8R tended to predict better in both cross-validation methods, with the exception being MULT9F and MULT11F for the Spanish dataset, using cross-validation method 2. In the Spanish and Singaporean dataset, ADD20F and ADD20R resulted in non-monotonic values, and MULT9F and MULT11F were non-monotonic when fitted to the Singaporean dataset (Table 3).

[TABLE 3 ABOUT HERE]

Discussion

The multiplicative models rest on the assumption that the relative distance between levels of functioning in one dimension is informative for estimating the distance between the levels on the other dimensions. Our analyses suggest that this assumption has merit, with the constrained multiplicative models generally outperforming the 20-parameter model with fixed and random intercepts in terms of predicting for out-of-sample health states. This indicates that, at least for the three datasets tested here, respondents appear to perceive the relative severity of the descriptive levels to be roughly the same across dimensions. This corresponds with previous studies in which general population respondents from several countries rated the descriptive levels using a numerical rating scale\textsuperscript{19,20}.

The improved performance associated with restricting the freedom of the 20-parameter model suggests that the number of parameters puts it at risk of overfitting to the observed data; modeling between-dimension variance in relative level distance did not improve predictions, suggesting that this is largely influenced by measurement error and random variance. The structure imposed by the multiplicative models makes them more robust to reductions in both number of observations and number of observed health states, i.e. less susceptible to overfitting, with the added benefit of reducing the risk of non-monotonicity.

Two competing explanations for the observed improvement in predictive validity of the constrained models can be suggested: The multiplicative models may constitute better representations of the general tendencies driving how respondents value hypothetical EQ-5D health states. Alternatively, the “true” underlying model is more complex,
but correct identification would require a greater number of observations, a wider range of observed health states, or both. One advantage of the multiplicative models over the 20-parameter model is that they can be estimated with smaller number of health states and observations for the health states, which could lower the costs of estimating an EQ-5D value set. More generally, the findings illustrate why model selection should not be informed by model fit on the observed data; model fit will always improve with added complexity, but may reduce the predictive accuracy. Cross-validation such as performed here reduces the risk of selecting models that suffer from overfitting to the observed data. If these findings are replicated in other EQ-5D-5L datasets, variants of these multiplicative models should be considered for future EQ-5D-5L value set generation.

Although MULT8R tended to perform best in our tests, we cannot conclusively point to any of the multiplicative models as being superior to the others. Given that the differences in performance between the predictive accuracy of the multiplicative models are relatively small, the choice boils down to three considerations: the importance attributed to parsimony, the confidence we have in the representativeness of sampled participants with regards to the target population, and our confidence in the representativeness of the selected health states with regards to the complete set of EQ-5D-5L health states. Further analyses (see the online appendix document, section D) suggest that most of the observed difference between methods 1 and 2 may be attributable to the difference between predicting values within and outside observed blocks of states. Thus, if we consider the remaining 3,038 EQ-5D-5L health states to be outside of the observed blocks, the simplest model, MULT8, may be the more attractive. Additionally, which of the multiplicative models works best may depend on language; the difference between “extreme problems” and “unable to perform” is likely to vary between different translations. For languages in which the difference is large, MULT9 or MULT11 may be better choices than MULT8. If the predictive accuracy of MULT8 is found to be similar to the accuracy of MULT9 or MULT11, it should be preferred based on parsimony.

**Weaknesses**

First, model selection by means of cross-validation rests on the assumption that mutual predictive ability (the ability to predict one part of the data using the remainder) is an appropriate proxy for predictive ability outside of the complete dataset. If the 86 directly observed health states do not adequately represent the full range of EQ-5D-5L health states, cross-validation is not informative. However, these assumptions are not specific to cross-validation, but are necessary for any modeling with the intent of extrapolation/interpolation. Second, the standard procedures
for data collection in EQ-5D-5L valuation studies are under continuous development. For example, it has been found that such valuation studies are susceptible to interviewer effects, and there is increasing awareness that continuous follow-up of interviewers is necessary to reduce the impact of idiosyncratic interviewer behaviour. As a consequence, recent developments of the EQ-VT system emphasize quality control more strongly. Therefore, it would be interesting to replicate the present analysis in datasets from more recently completed EQ-5D-5L valuation studies. Finally, the extent to which the findings presented here apply to DCE-based models, or hybrid models (using both C-TTO and DCE data) is not known. In addition to C-TTO, valuation studies for EQ-5D-5L collect preferences using discrete choice experiments (DCE). While the multiplicative models described here could be fitted to DCE data, cross-validation on DCE would require different leave-out procedures (e.g. leave-out of all observations of specific state pairs). In the case of hybrid models, cross-validation is complicated because the simultaneous fitting on two different kinds of data means that there is no apparent counterfactual that can easily serve as observed values in the leave-out procedure.

Conclusions

Random-intercept variants of the 8, 9, and 11-parameter models, assuming that the relative severity of the descriptive levels of the EQ-5D-5L is similar across dimensions, outperformed the 20-parameter model in terms of predicting the mean C-TTO values for out-of-sample health states. Modeling between-dimension variation in level distance in the manner of the 20-parameter model appears to cause overfitting on random variation. To reduce the risk of overfitting, cross-validation is recommended to inform model selection in future EQ-5D valuation studies. The multiplicative models were more robust to reductions in number of observations and number of observed health states, and were less prone to non-monotonicity over the EQ-5D-5L health state space. Replication using other EQ-5D-5L valuation study datasets is called for. Which of the multiplicative models is the best depends on how much we wish to penalize added complexity and the degree of confidence we have in the representativeness of the selected population sample and the EQ-5D-5L health states sample.
References:


Figure 1 legend:
Health states ordered by observed mean value. Predictions from MULT8R and ADD20RE in blue and red, respectively.

Figure 2 legend:
The elevation of the circles corresponds to the 20 parameters of the ADD20F model fit on the Chinese dataset. The vertical lines represent the 5 dimension parameters of MULT8R, and the dashed lines labelled L2, L3, and L4 represent the three level parameters. If the two models were identical, the circles would center on the intersection between the dashed lines and the vertical dimension parameter lines.
Dimension parameters

Level parameters

1 (fixed)

L4

L3

L2

UA2
UA3
UA4
UA5

AD2
AD3
AD4
AD5

SC2
SC3
SC4
SC5

PD2
PD3
PD4
PD5

MO2
MO3
MO4
MO5
### Table 1

**A) - number (%) of respondents by country, age, and sex**

| Age     | China        |          | China        |          | China        |          | China        |          | China        |          | China        |          | China        |          | China        |          | China        |          | China        |          | China        |          | China        |          | China        |          | China        |          | China        |          | China        |          | China        |          | China        |          | China        |          | China        |          | China        |          | China        |          | China        |          | China        |          | China        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain        |          | Spain four
Table 2: Cross-validation results and model fit.

<table>
<thead>
<tr>
<th></th>
<th>Fixed intercept</th>
<th></th>
<th>Random intercept</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ADD20 MULT8 MULT9 MULT11</td>
<td>ADD20R MULT8R MULT9R MULT11R</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>China</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All data</td>
<td>0.0339 0.0370 0.0362 0.0354</td>
<td>0.0349 0.0370 0.0365 0.0361</td>
<td></td>
</tr>
<tr>
<td>Method 1</td>
<td>0.0447 0.0411 0.0410 0.0410</td>
<td>0.0426 0.0398 0.0398 0.0404</td>
<td></td>
</tr>
<tr>
<td>Method 2</td>
<td>0.0437 0.0411 0.0417 0.0421</td>
<td>0.0420 0.0408 0.0409 0.0416</td>
<td></td>
</tr>
<tr>
<td>Spain</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All data</td>
<td>0.0473 0.0517 0.0497 0.0489</td>
<td>0.0497 0.0526 0.0524 0.0515</td>
<td></td>
</tr>
<tr>
<td>Method 1</td>
<td>0.0631 0.0589 0.0574 0.0579</td>
<td>0.0589 0.0569 0.0574 0.0576</td>
<td></td>
</tr>
<tr>
<td>Method 2</td>
<td>0.0601 0.0575 <strong>0.0565</strong> 0.0565</td>
<td>0.0581 0.0574 0.0579 0.0577</td>
<td></td>
</tr>
<tr>
<td>Singapore</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All data</td>
<td>0.0734 0.0771 0.0764 0.0757</td>
<td>0.0738 0.0769 0.0766 0.0758</td>
<td></td>
</tr>
<tr>
<td>Method 1</td>
<td>0.0981 0.0878 0.0883 0.0901</td>
<td><strong>0.0966</strong> 0.0863 0.0871 0.0886</td>
<td></td>
</tr>
<tr>
<td>Method 2</td>
<td>0.0917 0.0868 0.0876 0.0869</td>
<td>0.0888 0.0851 0.0857 0.0854</td>
<td></td>
</tr>
<tr>
<td>Lin's CCC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>China</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All data</td>
<td>0.9882 0.9855 0.9865 0.9869</td>
<td>0.9876 0.9854 0.9864 0.9867</td>
<td></td>
</tr>
<tr>
<td>Method 1</td>
<td>0.9796 0.9820 0.9825 0.9825</td>
<td>0.9820 0.9831 <strong>0.9838</strong> 0.9838</td>
<td></td>
</tr>
<tr>
<td>Method 2</td>
<td>0.9866 0.9878 0.9877 0.9875</td>
<td>0.9880 0.9882 0.9886 0.9883</td>
<td></td>
</tr>
<tr>
<td>Spain</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All data</td>
<td>0.9711 0.9650 0.9661 0.9681</td>
<td>0.9680 0.9634 0.9636 0.9653</td>
<td></td>
</tr>
<tr>
<td>Method 1</td>
<td>0.9493 0.9549 0.9553 0.9554</td>
<td><strong>0.9547</strong> 0.9578 0.9570 0.9572</td>
<td></td>
</tr>
<tr>
<td>Method 2</td>
<td>0.9651 <strong>0.9687</strong> 0.9688 <strong>0.9700</strong></td>
<td>0.9668 0.9681 0.9675 0.9683</td>
<td></td>
</tr>
<tr>
<td>Singapore</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All data</td>
<td>0.9581 0.9534 0.9537 0.9549</td>
<td>0.9562 0.9524 0.9526 0.9537</td>
<td></td>
</tr>
<tr>
<td>Method 1</td>
<td>0.9266 0.9385 0.9376 0.9356</td>
<td><strong>0.9277</strong> 0.9392 0.9383 0.9363</td>
<td></td>
</tr>
<tr>
<td>Method 2</td>
<td>0.9507 <strong>0.9542</strong> 0.9534 0.9540</td>
<td>0.9516 0.9540 0.9536 0.9540</td>
<td></td>
</tr>
<tr>
<td>Pearson's R</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>China</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All data</td>
<td>0.9882 0.9855 0.9865 0.9869</td>
<td>0.9876 0.9854 0.9864 0.9868</td>
<td></td>
</tr>
<tr>
<td>Method 1</td>
<td>0.9796 0.9820 0.9825 0.9825</td>
<td>0.9821 0.9832 <strong>0.9839</strong> 0.9838</td>
<td></td>
</tr>
<tr>
<td>Method 2</td>
<td>0.9938 0.9946 0.9948 0.9945</td>
<td>0.9943 0.9948 0.9950 0.9948</td>
<td></td>
</tr>
<tr>
<td>Spain</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All data</td>
<td>0.9722 0.9669 0.9676 0.9694</td>
<td>0.9692 0.9648 0.9650 0.9666</td>
<td></td>
</tr>
<tr>
<td>Method 1</td>
<td>0.9501 0.9560 0.9561 0.9562</td>
<td><strong>0.9556</strong> 0.9587 0.9579 0.9581</td>
<td></td>
</tr>
<tr>
<td>Method 2</td>
<td>0.9842 0.9884 0.9874 0.9869</td>
<td>0.9862 0.9885 0.9882 0.9877</td>
<td></td>
</tr>
<tr>
<td>Singapore</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All data</td>
<td>0.9592 0.9550 0.9551 0.9562</td>
<td>0.9571 0.9536 0.9538 0.9548</td>
<td></td>
</tr>
<tr>
<td>Method 1</td>
<td>0.9272 0.9396 0.9385 0.9364</td>
<td><strong>0.9282</strong> 0.9400 0.9390 0.9370</td>
<td></td>
</tr>
<tr>
<td>Method 2</td>
<td>0.9623 0.9662 0.9655 0.9658</td>
<td>0.9627 0.9663 0.9657 0.9660</td>
<td></td>
</tr>
</tbody>
</table>

Best result for each row in bold. "All data" included for reference: regular fit, not cross-validation. Black font indicates if the fixed- or random intercept variant was best for specific model in each row.
Table 3: Presence of non-monotonicity in models fitted on all data and in cross-validation analyses

<table>
<thead>
<tr>
<th>Non-monotonicity</th>
<th>Fixed intercept</th>
<th>Random intercept</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ADD20</td>
<td>MULT8</td>
</tr>
<tr>
<td>China</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All data</td>
<td>0 / 1</td>
<td>0 / 1</td>
</tr>
<tr>
<td>Method 1</td>
<td>0 / 86</td>
<td>0 / 86</td>
</tr>
<tr>
<td>Method 2</td>
<td>1 / 10</td>
<td>0 / 10</td>
</tr>
<tr>
<td>Spain</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All data</td>
<td>1 / 1</td>
<td>0 / 1</td>
</tr>
<tr>
<td>Method 1</td>
<td>86 / 86</td>
<td>0 / 86</td>
</tr>
<tr>
<td>Method 2</td>
<td>10 / 10</td>
<td>0 / 10</td>
</tr>
<tr>
<td>Singapore</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All data</td>
<td>1 / 1</td>
<td>0 / 1</td>
</tr>
<tr>
<td>Method 1</td>
<td>85 / 86</td>
<td>0 / 86</td>
</tr>
<tr>
<td>Method 2</td>
<td>10 / 10</td>
<td>0 / 10</td>
</tr>
</tbody>
</table>

The numerator represents the number of non-monotonic models in the test, while the denominator indicates the total number of fitted models. Each model was fitted once to the full set of data. With 86 observed health states, there were 86 fitted models using method 1. With 10 health state blocks, there were 10 fitted models using method 2.