Does Inclusion of Interactions Result in Higher Precision of Estimated Health State Values?

Nicolet, Anna; Groothuis-Oudshoorn, Catharina G. M.; Krabbe, Paul F. M.

Published in:
Value in Health

DOI:
10.1016/j.jval.2018.06.001

Publication date:
2018

Citation for published version (APA):
Nicolet, A., Groothuis-Oudshoorn, C. G. M., & Krabbe, P. F. M. (2018). Does Inclusion of Interactions Result in Higher Precision of Estimated Health State Values? Value in Health, 21(12), 1437-1444. https://doi.org/10.1016/j.jval.2018.06.001

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/jval

Preference-Based Assessments

Does Inclusion of Interactions Result in Higher Precision of Estimated Health State Values?

Anna Nicolet, MSc [1,*], Catharina G.M. Groothuis-Oudshoorn, PhD [2], Paul F.M. Krabbe, PhD [1]

[1]Department of Epidemiology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands;
[2]Department of Health Technology and Services Research, MIRA Institute, University of Twente, Enschede, The Netherlands

ABSTRACT

Background: Most preference-based instruments producing overall values for health states are devised on the simplifying assumption that the overall effect of distinct health-related quality of life domains (attributes) of the instrument equals the sum of the attributes. Nevertheless, health attributes are often inter-related and depend on each other. **Objectives:** To investigate whether inclusion of second-order interactions in the three-level EuroQol five-dimensional questionnaire (EQ-5D-3L) value function would result in better fit and lead to different health state values than a model with main effects only. **Methods:** Using an efficient design, 400 pairs of EQ-5D-3L health states were generated in a pairwise choice format. We analyzed responses of 4000 people from the general population using a conditional logit model, and we tested goodness of fit using pseudo R^2, Akaike information criterion, differences in log-likelihood, and likelihood ratio. We compared accuracies of models' predictions based on root mean square error and mean absolute error. **Results:** The interaction-effects model showed systematically lower values than the main-effects model. Inclusion of interactions resulted only in a slightly better model fit. Interactions comprising mobility and self-care were the most salient. **Conclusions:** For the EQ-5D-3L, a value function based on interactions produces systematically lower values than a main-effects model, meaning that the effect of two or more health problems combined is stronger than the sum of the individual main effects.

Keywords: discrete choice, EQ-5D-3L, main effects, second-order interactions, values

Introduction

A construct commonly used in health outcomes measurement is health-related quality of life (HRQOL), a subjective measure of perceived health status consisting of physical, mental, and social domains [1,2]. One common framework to measure HRQOL is the use of preference-based measurement methods. Instead of measuring the level of reported complaints (i.e., their frequency and intensity) for distinct health domains, these methods express the quality of a patient's health condition. Preference-based measures differ from other approaches that measure health condition in that they explicitly incorporate weights reflecting the importance attached to a set of specific health domains (technical term: attributes) that each capture a specific health aspect. The measures produced by these methods are expressed in a single metric number, which here we refer to as "value." The core of a preference-based measurement framework consists of a response task comparing at least two objects (in the present case health condition) and to express which object is preferred (is better). Often the structured description of a health condition is referred to as a health state: a small set of attributes each with a limited number of levels of severity. The respondents do not score the attributes one by one but consider the whole set of health attributes, which requires reading and mentally processing all the attributes in the set simultaneously [3]. The response task is to compare complete attribute sets, differing according to levels of severity, or to compare sets with a specified health outcome (e.g., immediate dead or living in full health for a specified number of years). By these comparisons a preference for one of the combinations of health states or health outcomes is evoked. There are several techniques allowing health state evaluation within a preference-based framework, but in the present study we chose the more recently introduced method of discrete

Conflicts of interest: The authors have indicated that they have no conflicts of interest with regard to the content of this article.

* *Address correspondence to:* Anna Nicolet, Department of Epidemiology, University Medical Center Groningen, University of Groningen, P.O. Box 30.001, Groningen 9700 RB, The Netherlands.
E-mail: a.selivanova@umcg.nl

https://doi.org/10.1016/j.jval.2018.06.001

choice modeling. Discrete choice modeling is widely used to elicit personal and societal preferences in health valuation studies [4]. Discrete choice is considered a relatively easy task for the respondents because it mimics individual everyday choices: Which of the available options is more preferable? (Fig. 1).

The total number of states to be valued is determined by the possible level combinations of the classification. If there are few states, it may even be feasible to value them all. If there are many, a well-chosen subset (constructed in such a way so as to maximize the information derived from a limited set of states out of all possible states) can be valued empirically, and the values for the remaining states can be estimated (usually by regression modeling). The values produced by these preference-based systems can be implemented in health outcomes research, disease-modeling studies, and economic evaluations to compare different health care interventions and in the planning and monitoring of health programs. The most common preference-based instruments (e.g., six-dimensional health state short form [SF-6D] or 15D) were developed using value functions comprising only main effects and ignoring the interactions between health attributes [5,6]. Main-effects functions rely on the simplifying assumption that the overall effect of all HRQOL attributes equals the sum of the attribute levels included in the function. Interactions play a role when the overall effect of two separate attributes is significantly more (or less) than their individual effects (e.g., reduction in perceived health status may intensify if two different health problems interact). Nevertheless, health attributes are often related and considered to depend on each other. Interactions were taken into account only for the Health Utility Index (seven attributes with five or six levels per attribute) and the Assessment of Quality of Life (which has versions with four, six, seven, or eight attributes with multiple attributes). Nevertheless, by using a multiplicative model the interactions among all attributes were forced to be the same [2,7]. Other explorative studies [4,8] demonstrated that the effect of health state attributes is not simply additive and that interactions may be important. This assumption, however, has not yet been tested thoroughly for preference-based instruments [9–11].

Using the three-level EuroQol five-dimensional questionnaire (EQ-5D-3L) instrument, this study investigates whether the inclusion of interaction terms leads to different estimated values for health states, and whether a model with interactions has better fit than a main-effects model.

Methods

EQ-5D-3L Instrument

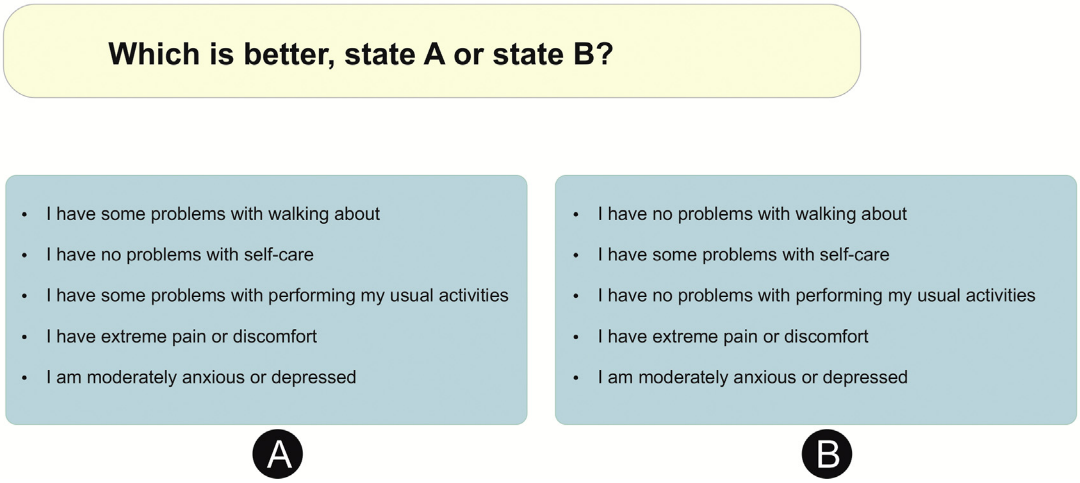
The EQ-5D instrument was developed by the EuroQol Group (www.euroqol.org) as a relatively simple generic preference-based instrument that could be used in clinical studies and would provide values of health states for use in economic evaluations [12]. The EQ-5D-3L descriptive system comprises five attributes: mobility (MO), self-care (SC), usual activities (UA), pain/discomfort (PD), and anxiety/depression (AD). Each attribute has three levels: no problems, some problems, and severe problems. EQ-5D-3L health states are defined by selecting one level from each attribute, with 11111 denoting perfect health (no problems in any attributes) and 33333 the worst possible health state (severe problems in all attributes). While developing the EQ-5D, researchers were experimenting with various valuation techniques and considered discrete choice modeling as a promising alternative to the conventional valuation techniques (time trade-off, standard gamble, and visual analogue scale). Nevertheless, the produced health values were based on value functions comprising only main effects and were produced by methods other than discrete choice methodology [13,14]. Simple additive value functions comprising main effects assumed that each of the five attributes was independent of others, ignoring the effects of any other attribute or their interactions [15].

Discrete Choice Modeling

Discrete choice modeling is a widely used technique to elicit personal and societal preferences in health valuation studies [4]. The statistical literature classifies it within the modern framework of probabilistic discrete choice models that are consistent with economic theory (i.e., the random utility model) [16–19]. Discrete choice modeling is based on probabilistic statistical routines (logit or probit regression models) and are used to establish the relative merit of one phenomenon relative to others [20,21]. Such choice models allow estimating the relative importance of health state attributes with certain levels, and overall values for health states with different combinations of attribute levels.

Health State Selection

The EQ-5D-3L contains five attributes with three levels each, yielding $3^5 = 243$ possible health states. Presenting health states as paired comparisons in the discrete choice task (two health states being assessed together) increases this number to 29,403 possible


Which is better, state A or state B?

- I have some problems with walking about
- I have no problems with self-care
- I have some problems with performing my usual activities
- I have extreme pain or discomfort
- I am moderately anxious or depressed

- I have no problems with walking about
- I have some problems with self-care
- I have no problems with performing my usual activities
- I have extreme pain or discomfort
- I am moderately anxious or depressed

Fig. 1 – Example of a discrete choice task for the EQ-5D-3L. EQ-5D-3L, three-level EuroQol five-dimensional questionnaire.

combinations. The evaluation of all possible combinations is known as a full factorial design, which allows the researcher to estimate all main effects and all possible interaction effects. In practice, this design is rarely used, because it is considered tedious and/or cost-prohibitive [22]. Another practical deterrent is that it usually entails very large sample sizes, a requirement that cannot always be met. These conditions explain why full factorial designs are almost never used for the valuation of health states, and even rarely in the field of marketing.

Fractional designs were developed to facilitate the careful selection of a subset of choice tasks out of all possible combinations. A carefully selected subset should be sufficient to reveal all important information for the investigated issue (in our case, attributes with their different levels in each of the two health state descriptions), while using only part of experimental efforts necessary for the full factorial design [23]. A fractional design was applied in the present study. The first step was to determine how many health state pairs to include in the design. This number should be sufficient for estimating all main effects and all second-order interaction effects for the EQ-5D-3L. In discrete choice models, the minimum criterion implies that the number of choice tasks is defined by the number of parameters. Specifically, the minimum number exceeds by one the number of parameters needed to estimate in the model. The attribute levels used for the present study are categorical variables, which are represented by dummy variables: MO1 (no problems with mobility), MO2 (some problems with mobility), MO3 (confined to bed), SC1 (no problems with self-care), SC2 (some problems with washing or dressing), SC3 (unable to wash or dress myself), UA1 (no problems with usual activities), UA2 (some problems with usual activities), UA3 (unable to perform usual activities), PD1 (no pain/discomfort), PD2 (moderate pain/discomfort), PD3 (extreme pain/discomfort), AD1 (no anxiety/depression), AD2 (moderate anxiety/depression), and AD3 (extreme anxiety/depression). Effects coding was used in the design of the study, whereby level 3 was chosen as reference (omitted). Therefore, the main-effects model included 11 parameters for all nonomitted attributes at levels 1 and 2 (no problems and some problems), summing up to 10 parameters to estimate Equation 1. Expressed as a formula, the model predicts latent values (V) of individuals choosing health states, where β_{1-10} represents unknown regression coefficients and MO1, MO2, SC1, SC2, …, AD2 are alternative-specific explanatory variables. In effects coding, the effects of the reference variable (level 3) can be derived as a negative summation of the effects of all nonomitted levels (levels 1 and 2). For example, the effect of level 3 mobility is calculated as $-(\beta_1\mathrm{MO1} + \beta_2\mathrm{MO2})$.

$$\begin{aligned}\mathrm{Vs} = \alpha + \beta_1\mathrm{MO1} + \beta_2\mathrm{MO2} + \beta_3\mathrm{SC1} + \beta_4\mathrm{SC2} + \beta_5\mathrm{UA1} + \beta_6\mathrm{UA2}\\ + \beta_7\mathrm{PD1} + \beta_8\mathrm{PD2} + \beta_9\mathrm{AD1} + \beta_{10}\mathrm{AD2}.\end{aligned} \tag{1}$$

The interaction model included the intercept, all main effects (10 parameters), and all second-order interactions between levels 1 and 2 (40 parameters), resulting in 51 parameters. This implies that at least 52 pairs of health states are needed to identify the model (Equation 2).

After consideration of the number of choice tasks used in earlier studies [24–26] and the criteria for the number of choice tasks to include in the design, it was decided to increase the number of pairs to 400. That would allow for a wider range of estimated health states with various severity levels.

Experimental Design

Interaction models are rarely applied because of their complexity due to the large number of health state pairs to be judged by respondents. Judging a large number of pairs by the same respondent can result in respondents' fatigue. To avoid this, researchers need to develop a design, optimal in terms of statistical and response efficiency, in which different blocks of pairs are offered to different sets of respondents. In our study, we used the following approach: the set of 400 health state pairs was divided into 25 blocks with 16 choice tasks each. Earlier studies suggested that 16 choice tasks would be acceptable to the respondents and would not affect their responses [24,27,28]. Reliability may be questionable if the respondents are bored or fatigued. Burden can be caused either by task complexity or by having a large number of tasks to carry out. The complexity of the tasks was reduced by implementation of two-level overlap for the health states, and the number was limited to 16 choice tasks per respondent. The two-level overlap implies fixing two out of five attributes at the same level while the other three attributes can vary.

A common problem in health state valuation exercises is dominance, because all attributes are ordered, and smaller health problems are always preferred to bigger ones. Dominant pairs do not offer additional information but instead reduce the design's statistical efficiency (variability of parameter estimates rises; standard errors are getting larger). Therefore, such combinations, where for one health state all the attributes were worse (or better) than those of its paired state, were removed from the candidate pairs for constructing the design. The set of possible pairs without dominant combinations and with two-level overlap was selected out of all possible 29,403 pairs. Out of the resulting set of 14,580 pairs, 400 health state pairs were selected using an efficient design (Ngene software, Choice Metrics, Sydney, Australia, the multinomial logit model, taking 500 Bayesian draws, Halton sequence, modified Fedorov algorithm). An experimental design is called statistically efficient if the parameters are estimated with the least possible standard errors. Additional to statistical efficiency, there is response efficiency. This means that respondents are offered tasks with reduced complexity to avoid attentional failures and failure in memory, thereby getting more reliable responses. The design was constructed using an iterative procedure, whereby designs were compared in terms of their D-error, which is the measure of statistical efficiency we decided to use. D-errors were computed on the basis of expected values of the model parameters. Generation of an efficient design in Ngene requires prior distributions of the parameters, which were derived from a previous EQ-5D-3L study [4]. Because that study was not aimed at

$$\begin{aligned}&\mathrm{Vs} = \alpha + \beta_1\mathrm{MO1} + \beta_2\mathrm{MO2} + \beta_3\mathrm{SC1} + \beta_4\mathrm{SC2} + \beta_5\mathrm{UA1} + \beta_6\mathrm{UA2} + \beta_7\mathrm{PD1} + \beta_8\mathrm{PD2} + \beta_9\mathrm{AD1} + \beta_{10}\mathrm{AD2}\\ &+\beta_{11}\mathrm{MO1} \times \mathrm{SC1} + \beta_{12}\mathrm{MO1} \times \mathrm{SC2} + \beta_{13}\mathrm{MO2} \times \mathrm{SC1} + \beta_{14}\mathrm{MO2} \times \mathrm{SC2} + \beta_{15}\mathrm{MO1} \times \mathrm{UA1}\\ &+\beta_{16}\mathrm{MO1} \times \mathrm{UA2} + \beta_{17}\mathrm{MO2} \times \mathrm{UA1} + \beta_{18}\mathrm{MO2} \times \mathrm{UA2} + \beta_{19}\mathrm{MO1} \times \mathrm{PD1} + \beta_{20}\mathrm{MO1} \times \mathrm{PD2}+\\ &\beta_{21}\mathrm{MO2} \times \mathrm{PD1} + \beta_{22}\mathrm{MO2} \times \mathrm{PD2}\varepsilon + \beta_{23}\mathrm{MO1} \times \mathrm{AD1} + \beta_{24}\mathrm{MO1} \times \mathrm{AD2} + \beta_{25}\mathrm{MO2} \times \mathrm{AD1}+\\ &\beta_{26}\mathrm{MO2} \times \mathrm{AD2} + \beta_{27}\mathrm{SC1} \times \mathrm{UA1} + \beta_{28}\mathrm{SC1} \times \mathrm{UA2} + \beta_{29}\mathrm{SC2} \times \mathrm{UA1} + \beta_{30}\mathrm{SC2} \times \mathrm{UA2} + \beta_{31}\mathrm{SC1}\times\\ &\mathrm{PD1} + \beta_{32}\mathrm{SC1} \times \mathrm{PD2} + \beta_{33}\mathrm{SC2} \times \mathrm{PD1} + \beta_{34}\mathrm{SC2} \times \mathrm{PD2} + \beta_{35}\mathrm{SC1} \times \mathrm{AD1} + \beta_{36}\mathrm{SC1} \times \mathrm{AD2}+\\ &\beta_{37}\mathrm{SC2} \times \mathrm{AD1} + \beta_{38}\mathrm{SC2} \times \mathrm{AD2} + \beta_{39}\mathrm{UA1} \times \mathrm{PD1} + \beta_{40}\mathrm{UA1} \times \mathrm{PD2} + \beta_{41}\mathrm{UA2} \times \mathrm{PD1} + \beta_{42}\mathrm{UA2}\times\\ &\mathrm{PD2} + \beta_{43}\mathrm{UA1} \times \mathrm{AD1} + \beta_{44}\mathrm{UA1} \times \mathrm{AD2} + \beta_{45}\mathrm{UA2} \times \mathrm{AD1} + \beta_{46}\mathrm{UA2} \times \mathrm{AD2} + \beta_{47}\mathrm{PD1} \times \mathrm{AD1}+\\ &\beta_{48}\mathrm{PD1} \times \mathrm{AD2} + \beta_{49}\mathrm{PD2} \times \mathrm{AD1} + \beta_{50}\mathrm{PD2} \times \mathrm{AD2}.\end{aligned} \tag{2}$$

interaction estimations, only priors for main effects were set accordingly, and the priors for interactions were set to 0.

Sample Recruitment

According to the golden rule formulated by Johnson and Orme [29], N > 500*c*/(t × *a*), where N is the minimum sample size per a block of a survey, *c* is the largest product of levels in interactions, *t* is the number of tasks, and *a* is the number of alternatives. In the model with second-order interactions for EQ-5D-3L in the present study, *c* = 9, *t* = 16, *a* = 2, the number of blocks is 25, and the calculated minimum sample size is N > 500 × 9/(32 × 25) =141 × 25 = 3525, although a more sophisticated calculation procedure may be found [30]. In discrete choice modeling, a total of 50 to 60 observations per response task would generally be considered sufficient. On the basis of this number of observations per choice set, the minimum sample size for 400 response tasks (400 pairs) is 1500. The final sample for the present study consisted of 4000 members of the Dutch general population of working age 18 to 65 years, representative on age and sex. The respondents were recruited using the panel of the marketing Survey Sampling International company SSI (Rotterdam, The Netherlands). Possible dropouts or insufficient quality of responses, which could diminish the size of the sample eligible for the final analysis, were accounted for. Responses were assumed to be of insufficient quality when the completion time fell below 2 minutes, which was considered too short to perform 16 choice tasks carefully.

Analysis

The conditional logit routine was used to obtain coefficients of the EQ-5D-3L attribute levels from both models of interest: the main-effects model and the model with all second-order interactions (Stata 14.0, StataCorp, College Station, TX). Because the research question of the present study focuses on overall values and not on heterogeneity among respondents, the basic conditional logit model was considered as sufficient [16]. In the latter model, the estimated coefficients represented the effects of attribute levels and the interactions between the separate levels of one attribute versus the separate levels of another attribute. The overall significance of the 10 second-order interactions (MO × SC, MO × UA, MO × PD, MO × AD, SC × UA, SC × PD, SC × AD, UA × PD, UA × AD, and PD × AD) was not estimated on the basis of coefficients. Rather, it was tested on the basis of the likelihood ratio to conclude whether adding the interactions improved the model fit. The likelihood ratio was calculated for the model with all second-order interactions and the model without one specified interaction (i.e., MO × SC). If the *P* value in the likelihood ratio test is low (<0.05), the goodness of fit of the model with specified interactions is deemed significantly better than the goodness of fit without the specified interactions.

The goodness of fit for the model with main effects only and the model with interaction effects was investigated using pseudo R^2 and Akaike information criterion (AIC). A higher pseudo R^2 and a lower AIC indicate better model fit. In addition, mean absolute error (MAE) and root mean square error (RMSE) were calculated to assess the accuracy of predictions of both models. MAE and RMSE present the differences between observed and predicted values from each model, therefore reflecting the accuracy of models' predictions.

Table 1 – Respondents' characteristics

Characteristics	Respondents (N = 3669)
Male, n (%)	1645 (45)
Age (y), mean ± SD	46.0 ± 13.4
Age group, n (%)	
18–24 y	145 (9)
25–34 y	219 (13)
35–44 y	316 (19)
45–54 y	426 (26)
>55 y	539 (33)
Female, n (%)	2024 (55)
Age (y), mean ± SD	42.5 ± 13.8
Age group, n (%)	
18–24 y	313 (15)
25–34 y	329 (16)
35–44 y	394 (20)
45–54 y	529 (26)
>55 y	459 (23)

To demonstrate the differences between the estimates for the main-effects model and those for the interaction-effects model, predicted values of 243 EQ-5D-3L health states were plotted against each other (SigmaPlot 13.0, Systat Software, Inc., San Jose, CA, USA). The value for the alternative in a choice task is modeled as the product of the health state characteristics (severity of an attribute, such as level 1 problems with mobility or level 2 problems with self-care) and the health state preference parameters (β). It needs to be noted that in the conditional logit model the constant term α was not shown because it does not vary across the alternatives. For instance, having parameter estimates for nonomitted levels 1 and 2 from the conditional logit model, and calculating estimates for level 3 as the negative summation for the effects of all nonomitted levels (levels 1 and 2), we can calculate the predicted value for the health state 23112 on the basis of the main-effects model (Equation 4) as follows:

$$\begin{aligned} U &= \beta\text{MO2} - (\beta\text{SC1} + \beta\text{SC2}) + \beta\text{UA1} + \beta\text{PD1} + \beta\text{AD2} \\ &= 0.351 - (0.488 + 0.084) + 0.393 + 0.563 + 0.205 = 0.94. \end{aligned} \tag{4}$$

For the interaction-effects model, the estimates for all 243 health states were calculated by the summation of main-effects and interaction-effects coefficients of levels comprising the health state. Consider, for example, the calculation for health state 23112 (Equation 5):

$$\begin{aligned} U = {} & \beta\text{MO2} - (\beta\text{SC1} + \beta\text{SC2}) + \beta\text{UA1} + \beta\text{PD1} + \beta\text{AD2} - (\beta\text{MO2} \times \text{SC1} + \beta\text{MO2} \times \text{SC2}) \\ & + \beta\text{MO2} \times \text{UA1} + \beta\text{MO2} \times \text{PD1} + \beta\text{MO2} \times \text{AD2} - (\beta\text{SC1} \times \text{UA1} + \beta\text{SC2} \times \text{UA1}) \\ & - (\beta\text{SC1} \times \text{PD1} + \beta\text{SC2} \times \text{PD1}) - (\beta\text{SC1} \times \text{AD2} + \beta\text{SC2} \times \text{AD2}) + \beta\text{UA1} \times \text{PD1} \\ & + \beta\text{UA1} \times \text{AD2} + \beta\text{PD1} \times \text{AD2} = 0.329 - 0.565 + 0.397 + 0.572 + 0.194 - 0.001 \\ & + 0.043 - 0.048 + 0.021 - 0.019 - 0.043 - 0.034 + 0.040 - 0.002 - 0.023 = 0.86. \end{aligned} \tag{5}$$

The given calculations of values (Equations 4 and 5) are based on unscaled model coefficients (i.e., values are not scaled from 0 to 1). To see whether the health state values in the main-effects model differ from the health state values in the model including second-order interactions, the values of all health states were rescaled from 0 (worst health state 33333) to 1 (best health state 11111) and then plotted.

Results

Sample

The survey was completed by 4000 respondents aged between 18 and 65 years. Nevertheless, 309 respondents were removed from the analysis because their responses were deemed unreliable because of the short amount of time spent on the survey (<2 minutes). Before the analysis, the responses of 22 respondents were discarded because of the observed pattern of choosing only the left or only the right alternative. Ultimately, 3669 respondents were included in the final analysis. The representative sample from the Dutch population was recruited in October 2016 (Table 1).

Main-Effects and Interaction-Effects Models

In the main-effects model for EQ-5D-3L, all estimates were logically ordered and statistically significant at the 95% level (Table 2). In the interaction-effects model, all main effects were statistically significant at the 95% level (Table 3). Inclusion of all second-order interactions simultaneously resulted in a statistically significant improvement of model fit (log-likelihood ratio test: LR χ^2 (40) = 289.74; $P = 0.00$). Moreover, all 10 pairwise interactions between attributes are significant. The interaction term consisting of mobility and self-care is the most salient one because its likelihood ratio test statistic is the highest (LR = 98.3) and the associated *P* value is very low. The lowest likelihood ratio test statistic (LR = 11.17) was shown for the interaction of mobility with anxiety/depression (Table 3). Nevertheless, inclusion of all second-order interactions improves the fit only slightly on the basis of the indicators of pseudo R^2 and AIC. The improvement of model fit by including interaction effects based on pseudo R^2 was modest (rise from 0.174 to 0.177). Similar results were found for the AIC, whereby the lower AIC indicated better model fit (67271.2 for the main-effects model and 67061.5 for the interaction-effects model). The measures of model accuracy, RMSE and MAE, indicated in favor of the model with interactions in terms of predicting accuracy. Health states and predicted values from the main-effects and interaction-effects models were plotted (Figs. 2 and 3), and it was demonstrated that the interaction-effects model had lower values than the main-effects model on the entire range of health states. The maximum difference between the values produced by a main-effects and an interaction-effects model is 0.129, whereas the average difference is 0.076.

Discussion

We have demonstrated the feasibility of deriving values for EQ-5D-3L states using a discrete choice model with all second-order interactions and efficient experimental design properties. It was shown that the effect of the health attributes is not simply additive. Interactions do contribute to the final estimated values for health states.

Most studies do not use all possible interaction terms but only those of interest [31]. For example, instead of including all second-order interactions, some studies [4,12,13] used one overall (omnibus) term (N3) to capture having severe (level 3) problems for at least one attribute. Other studies investigated the inclusion of a constant signifying any movement away from perfect health, or a D1 term (interaction term representing the number of movements away from perfect health because of having one or more attributes at level 2 or 3) [32,33]. These studies found little impact of interactions on the model fit. This is not surprising, because they were not designed to properly estimate all possible interactions between distinct health attributes. Intuitively, many combinations of health attributes are imaginable, in which case interactions would exist. For example, the ability to perform usual activities may depend on a person's mobility or feeling of pain/discomfort, because these attributes define and are integrated into usual activity.

Table 2 – Parameter estimates for main-effects model on the basis of discrete choice data, effects coding

Attribute	Main-effects estimates	
	β (SE)	P value
MO1	0.618 (0.01)	0.000
MO2	0.351 (0.01)	0.000
SC1	0.488 (0.01)	0.000
SC2	0.084 (0.01)	0.000
UA1	0.393 (0.01)	0.000
UA2	0.197 (0.01)	0.000
PD1	0.563 (0.02)	0.000
PD2	0.309 (0.01)	0.000
AD1	0.538 (0.01)	0.000
AD2	0.205 (0.01)	0.000
Pseudo R^2	0.1736	
AIC	67,271.21	
Log-likelihood	−33,625.61	
MAE	0.058	
RMSE	0.0745	

Note. The coefficients for omitted categories (level 3) can be calculated as the negative summation of nonomitted variables' coefficients. For example, β for MO3 = −(0.618 + 0.351) = −0.969.
AIC, Akaike information criterion; Attributes: MO, mobility; SC, self-care; UA, usual activities; PD, pain/discomfort; AD, anxiety/depression; MAE, mean absolute error; RMSE, root mean square error; SE, standard error.

The present study showed that although adding all possible second-order interactions improved the model fit, their inclusion improved the explained variance only slightly. The estimates were consistently lower moving downward from level 1 (having no problems), which suggested declining values for health states associated with incremental moves away from perfect health. The obtained estimates from the interaction-effects model were systematically lower than the estimates from the main-effects model. Moreover, estimates were consistently negative, which suggested a declining marginal utility associated with additional shifts away from perfect health. The results of the present study demonstrated presence of interactions among the attributes in the EQ-5D-3L, meaning that the effect of two or more health problems combined is stronger than the sum of the individual main effects. The same effect was investigated in the development of Health Utility Index 3 [7].

We found a number of quantitatively and statistically significant interactions among the attributes mobility, self-care, and pain/discomfort. The most salient one is between mobility and self-care; inclusion of this interaction term contributes more to model fit improvement than does the inclusion of others. In the study by Mulhern et al. [34] investigating the interactions between the attributes of EQ-5D health state and duration, the interaction between pain/discomfort and duration showed the largest effect on values of health states, whereas the effect of interaction between mobility and duration was the lowest. In the study by Viney et al. [24], the weights for the attributes pain/discomfort, mobility, and self-care were larger. They also found that the following two interactions had the largest effects on the values of the health states: the interaction between mobility and self-care and the interaction between mobility and pain/discomfort. These findings concur with the present study findings. In the study by Jelsma and Maart [35], severe problems with mobility and pain/discomfort

Table 3 – Parameter estimates for interaction-effects model based on discrete choice data

Attribute	Interaction-effects estimates	
	β (SE)	P value
MO1	0.636 (0.01)	0.000
MO2	0.329 (0.01)	0.000
SC1	0.489 (0.01)	0.000
SC2	0.077 (0.01)	0.000
UA1	0.397 (0.01)	0.000
UA2	0.187 (0.01)	0.000
PD1	0.572 (0.02)	0.000
PD2	0.291 (0.01)	0.000
AD1	0.550 (0.01)	0.000
AD2	0.194 (0.01)	0.000
MO × SC (likelihood value)	98.30	0.000
MO1 × SC1	0.104 (0.01)	0.000
MO1 × SC2	0.000 (0.01)	0.989
MO2 × SC1	−0.043 (0.01)	0.002
MO2 × SC2	0.045 (0.01)	0.001
MO × UA (likelihood value)	36.77	0.000
MO1 × UA1	0.008 (0.01)	0.566
MO1 × UA2	−0.003 (0.01)	0.831
MO2 × UA1	0.043 (0.01)	0.001
MO2 × UA2	0.028 (0.01)	0.019
MO × PD (likelihood value)	36.81	0.000
MO1 × PD1	0.083 (0.01)	0.000
MO1 × PD2	−0.038 (0.01)	0.002
MO2 × PD1	−0.048 (0.01)	0.001
MO2 × PD2	0.032 (0.01)	0.022
MO × AD (likelihood value)	11.17	0.025
MO1 × AD1	0.001 (0.01)	0.958
MO1 × AD2	−0.027 (0.01)	0.057
MO2 × AD1	0.017 (0.01)	0.207
MO2 × AD2	0.021 (0.01)	0.102
UA × SC (likelihood value)	29.60	0.000
UA1 × SC1	0.011 (0.01)	0.412
UA1 × SC2	0.008 (0.01)	0.563
UA2 × SC1	0.054 (0.01)	0.000
UA2 × SC2	−0.018 (0.01)	0.172
SC × PD (likelihood value)	24.74	0.000
SC1 × PD1	0.064 (0.01)	0.000
SC1 × PD2	−0.030 (0.01)	0.035
SC2 × PD1	−0.021 (0.01)	0.122
SC2 × PD2	0.027 (0.01)	0.036
SC × AD (likelihood value)	29.89	0.000
SC1 × AD1	0.053 (0.01)	0.000
SC1 × AD2	−0.022 (0.01)	0.083
SC2 × AD1	−0.032 (0.01)	0.023
SC2 × AD2	0.056 (0.01)	0.000
UA × PD (likelihood value)	65.43	0.000
UA1 × PD1	0.040 (0.01)	0.003
UA1 × PD2	−0.047 (0.01)	0.001
UA2 × PD1	0.055 (0.01)	0.000
UA2 × PD2	0.010 (0.01)	0.411
UA × AD (likelihood value)	12.87	0.012
UA1 × AD1	−0.010 (0.02)	0.536
UA1 × AD2	−0.002 (0.01)	0.864
UA2 × AD1	0.006 (0.01)	0.676
UA2 × AD2	0.035 (0.01)	0.005
PD × AD (likelihood value)	16.41	0.003
PD1 × AD1	0.052 (0.01)	0.000
PD1 × AD2	−0.023 (0.01)	0.084
PD2 × AD1	−0.009 (0.01)	0.501
PD2 × AD2	0.014 (0.01)	0.271
Pseudo R^2	0.1772	
AIC	67,061.48	
Log-likelihood	−33,480.74	
MAE	0.053	
RMSE	0.0673	

continued on next page

Table 3 – *continued*

Note. The coefficients for omitted categories (level 3) can be calculated as the negative summation of nonomitted variables' coefficients. β for MO3 = −(0.636 + 0.329) = −0.965; β for interaction MO3 × SC1 = −(βMO1 × SC1 + βMO2 × SC1) = −(0.104 − 0.043) = −0.061.

AIC, Akaike information criterion; Attributes: MO, mobility; SC, self-care; UA, usual activities; PD, pain/discomfort; AD, anxiety/depression; MAE, mean absolute error; RMSE, root mean square error; SE, standard error.

showed the largest significant effect on HRQOL as in the present study.

The present study has several strengths. An important one is the balance of design efficiency and response efficiency of our study. The design did not contain dominant pairs, and by implementation of two-level overlap, response efficiency was reached. This made the response tasks easier, thereby reducing respondent fatigue [36–40]. Furthermore, a large sample was obtained, which made it possible to estimate and investigate all possible second-order interaction terms for the EQ-5D-3L. Many health states were included in the study, which increased the accuracy of the

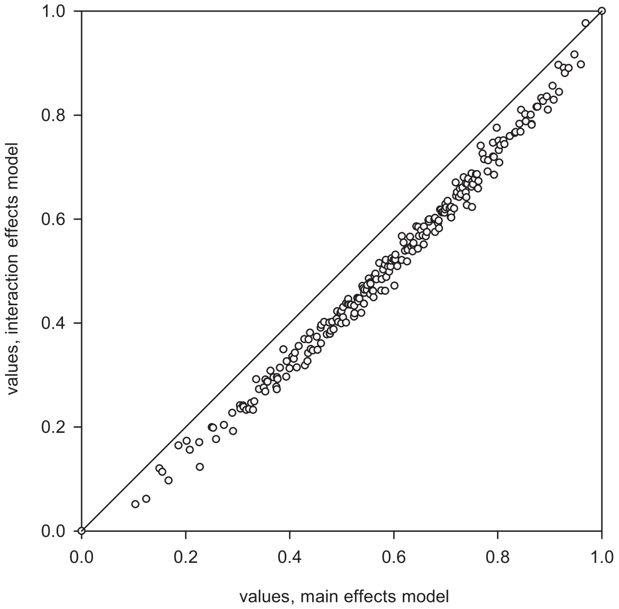


Fig. 2 – Predicted values (rescaled from 0 to 1) for 243 EQ-5D-3L health states based on the model with main effects and on the model including interactions. EQ-5D-3L, three-level EuroQol five-dimensional questionnaire.

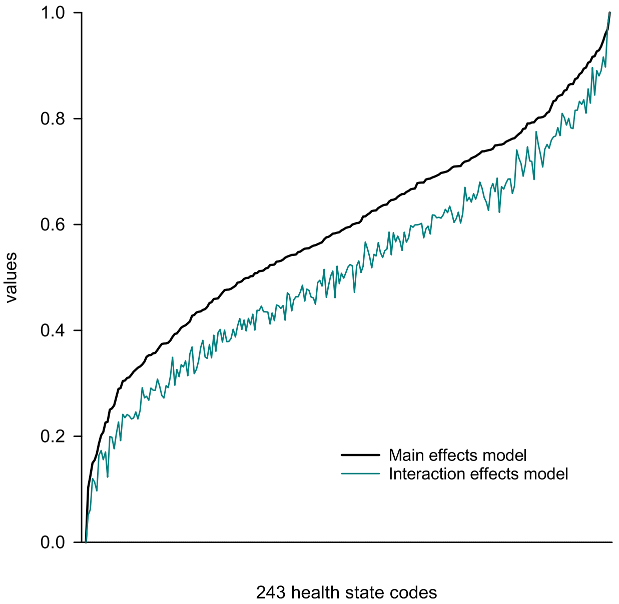


Fig. 3 – Predicted values (rescaled from 0 to 1) for 243 EQ-5D-3L health states based on the model with main effects and on the model including interactions, sorted by the values for the main-effects model. EQ-5D-3L, three-level EuroQol five-dimensional questionnaire.

results and aided in estimating all possible second-order interactions.

The study has some limitations too. The first being that no priors for interaction terms were used when constructing the experimental design. Priors were set to 0 because none of the previous studies had investigated all possible second-order interactions for the EQ-5D-3L jointly. It may be argued that priors for interactions could have been achieved with a pilot study. This, however, would have required redesigning 400 pairs of health states, terminating the sampling process, and rerunning the survey. Therefore, it was decided not to run a pilot, so the 0 priors were set for interaction terms. A second limitation is that the results may be affected by the fact that the assessment of the EQ-5D-3L health states was performed by a sample of the general population. Newly developed “experience-based” methods, which make use of patients who assess health state descriptions and compare these to their own health condition [41], might reveal larger interaction effects. Another limitation is the absence of theoretical hypothesis for testing specific interactions. Nevertheless, the aim of our study was to investigate whether adding all second-order interactions in a model results in different estimates for the health states and to test the feasibility of such a model, rather than testing specific interactions, such as the N3 term [14,15,42].

Testing specific interactions instead of all interactions could be beneficial to future research on the five-level EQ-5D (EQ-5D-5L), which has five instead of three levels for each of the five attributes, generating a much wider array of possible interactions. For this EQ-5D-5L version, testing all possible interactions could be troublesome because of the large number of parameters to be estimated and the very large sample size required. Therefore, theoretical knowledge and empirical evidence from the present study may be applied to select specific key interactions for further research. For example, the interactions among the attributes mobility and self-care, which appeared the most salient for the EQ-5D-3L, could be investigated in the EQ-5D-5L.

Conclusions

Estimation of EQ-5D-3L states using statistical models comprising all second-order interactions is feasible. Health attributes are related to and dependent on each other, an assumption that has been confirmed by the significance of the interactions between the five attributes of the EQ-5D-3L. For the EQ-5D-3L, a value function based on interactions produces systematically lower values than a main-effects model. It seems that the simple main-effects model for the EQ-5D-3L instruments may not be sufficiently accurate to produce credible health state values. Nevertheless, the practical implications of the differences between values generated with or without interactions may be small, because differences between values for various health states seem more comparable.

Source of financial support: Financial support for this study was provided entirely by a contract with EuroQol Group (EQ Project 2014150). The funding agreement ensured the authors’ independence in designing the study, interpreting the data, and writing and publishing the report.

REFERENCES

[1] World Health Organization. The First Ten Years of the World Health Organization. Geneva, Switzerland: World Health Organization; 1958.
[2] Krabbe PFM. The Measurement of Health and Health Status: Concepts, Methods and Applications from a Multidisciplinary Perspective. San Diego, CA: Elsevier/Academic Press; 2016.
[3] Selivanova A, Krabbe PFM. Eye tracking to explore attendance in health-state descriptions. PLoS One 2018;13:e0190111.
[4] Stolk EA, Oppe M, Scalone L, Krabbe PFM. Discrete choice modeling for the quantification of health states: the case of the EQ-5D. Value Health 2010;13:1005–13.
[5] Sintonen H. The 15D instrument of health-related quality of life: properties and applications. Ann Med 2001;33:328–36.
[6] Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. J Health Econ 2002;21:271–92.
[7] Feeny D, Furlong W, Torrance GW, et al. Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. Med Care 2002;40:113–28.
[8] Rowen D, Brazier J, Roberts J. Mapping SF-36 onto the EQ-5D index: How reliable is the relationship? Health Qual Life Outcomes 2009;7:27.
[9] Oppe M, Devlin NJ, van Hout B, et al. A program of methodological research to arrive at the new international EQ-5D-5L valuation protocol. Value Health 2014;17:445–53.
[10] Sullivan PW, Ghushchyan V. Mapping the EQ-5D index from the SF-12: US general population preferences in a nationally representative sample. Med Decis Making 2006;26:401–9.
[11] McDowell I, Newell C. Measuring Health: A Guide to Rating Scales and Questionnaires. 2nd ed. New York, NY: Oxford University Press; 1996.
[12] Bansback N, Brazier J, Tsuchiya A, Anis A. Using a discrete choice experiment to estimate societal health state utility values. J Health Econ 2012;31:306–18.
[13] Tsuchiya A, Ikeda S, Ikegami N, et al. Estimating an EQ-5D population value set: the case of Japan. Health Econ 2002;11:341–53.
[14] Lamers LM, McDonnell J, Stalmeier PF, et al. The Dutch tariff: results and arguments for an effective design for national EQ-5D valuation studies. Health Econ 2006;15:1121–32.
[15] Dolan P. Modeling valuations for EuroQol health states. Med Care 1997;35:1095–108.
[16] Arons MMA, Krabbe PFM. Probabilistic choice models in health-state valuation research: background, theories, assumptions and applications. Expert Rev Pharmacoecon Outcomes Res 2013;13:93–108.
[17] Krabbe PFM. Thurstone scaling as a measurement method to quantify subjective health outcomes. Med Care 2008;46:357–65.
[18] Louviere JJ, Lancsar E. Choice experiments in health: the good, the bad, the ugly and toward a brighter future. Health Econ Policy Law 2009;4:527–46.
[19] Thurstone LL. A law of comparative judgment. Psychol Rev 1927;4:273–86.
[20] McFadden D. Conditional logit analysis of qualitative choice behavior. In: Zarembka P, editor. Frontiers in Econometrics. New York, NY: Academic Press; 1974. p. 105–42.

[21] Krabbe PFM, Devlin NJ, Stolk EA, et al. Multinational evidence of the applicability and robustness of discrete choice modeling for deriving EQ-5D-5L health-state values. Med Care 2014;52:935–43.

[22] Kuhfeld WF. Marketing research methods in SAS: experimental design, choice, conjoint, and graphical techniques. 2005. Available from: http://support.sas.com/techsup/technote/ts723.html. [Accessed May 15, 2017].

[23] Box GE, Hunter JS, Hunter WG. Statistics for Experimenters: Design, Innovation, and Discovery. 2nd ed. Hoboken, NJ: Wiley; 2005.

[24] Viney R, Norman R, Brazier J, et al. An Australian choice experiment to value EQ-5D health states. J Health Econ 2014;23:729–42.

[25] Brazell JD, Louviere JJ. Length Effects in Conjoint Choice Experiments and Surveys: An Explanation Based on Cumulative Cognitive Burden. Sydney, Australia: Department of Marketing, University of Sydney; 1998.

[26] Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. J Health Econ 2002;21:271–92.

[27] Coast J, Flynn TN, Salisbury C, et al. Maximising responses to discrete choice experiments: a randomised trial. Appl Health Econ Health Policy 2006;5:249–60.

[28] Hall J, Fiebig DG, King MT, et al. What influences participation in genetic carrier testing? Results from a discrete choice experiment. J Health Econ 2006;25:520–37.

[29] Johnson R, Orme B. Getting the Most from CBC (Sawtooth Software Research Paper Series). Sequim, WA: Sawtooth Software; 2003.

[30] De Bekker-Grob EW, Donkers B, Jonker MF, Stolk EA. Sample size requirements for discrete-choice experiments in healthcare: a practical guide. Patient 2015;8:373–84.

[31] Norman R, Cronin P, Viney R, et al. International comparisons in valuing EQ-5D health states: a review and analysis. Value Health 2009;12:1194–200.

[32] Shaw JW, Johnson JA, Coons SJ. US valuation of the EQ-5D health states: development and testing of the D1 valuation model. Med Care 2005;43:203–20.

[33] Rand-Hendriksen K, Augestad LA, Dahl FA. A critical re-evaluation of the regression model specification in the US D1 EQ-5D value function. Popul Health Metr 2012;10:2.

[34] Mulhern B, Bansback N, Hole AR, Tsuchiya A. Using discrete choice experiments with duration to model EQ-5D-5L health state preferences: testing experimental design strategies. Med Decis Making 2016;37:285–97.

[35] Jelsma J, Maart S. Should additional attributes be added to the EQ-5D health-related quality of life instrument for community-based studies? An analytical descriptive study. Popul Health Metr 2015;13:13.

[36] Johnson FR, Lancsar E, Marshall D. Constructing experimental designs for discrete-choice experiments: report of the ISPOR Conjoint Analysis Experimental Design Good Research Practices Task Force. Value Health 2013;16:3–13.

[37] Flynn TN, Bilger M, Malhotra C, Finkelstein EA. Are efficient designs used in discrete choice experiments too difficult for some respondents? A case study eliciting preferences for end-of-life care. Pharmacoeconomics 2016;34:273–84.

[38] Jonker MF, Attema AE, Donkers B, et al. Are health state valuations from the general public biased? A test of health state preference dependency using self-assessed health and an efficient discrete choice experiment. J Health Econ 2017;26:1534–47.

[39] Louviere JJ, Islam T, Wasi N, et al. Designing discrete choice experiments: Do optimal designs come at a price? J Consum Res 2008;35:360–75.

[40] Maddala T, Phillips KA, Reed Johnson F. An experiment on simplifying conjoint analysis designs for measuring preferences. J Health Econ 2003;12:1035–47.

[41] Krabbe PFM. A generalized measurement model to quantify health: the multi-attribute preference response model. PLoS One 2013;8:e79494.

[42] Luo N, Johnson JA, Shaw JW, Coons SJ. A comparison of EQ-5D index scores derived from the US and UK population-based scoring functions. Med Decis Making 2007;27:321–6.