Elicitation and modelling of imprecise utility of health states

Michał Jakubczyk¹ · Dominik Golicki²

Published online: 18 May 2019
© The Author(s) 2019

Abstract

Utilities of health states are often estimated to support public decisions in health care. People's preferences may be imprecise, for lack of actual trade-off experience. We show how to elicit the utilities accounting for imprecision (using fuzzy sets), discover the main drivers of imprecision, and compare several approaches to modelling health state utility data in the fuzzy setting. We extended the time trade-off (TTO) questionnaire, to elicit utilities of states defined in the EQ-5D-3L descriptive system (health described by five dimensions) in 184 respondents. Our study demonstrates that respondents are capable of assessing their own imprecision and rigorous mathematical modelling is possible. The imprecision is larger than as inferred from the standard TTO method and is larger than estimation error, even in our smallish sample. Non-trading in TTO often results from imprecision, rather than lexicographic preferences for longevity over quality. People are especially imprecise in assessing the impact of usual activities (one of the dimensions) on utility; also, the internal inconsistency of a health state increases the imprecision. Fuzzy least squares method seems best suited to assign disutilities to individual dimensions, while separately modelling the location of utility and amount of imprecision seems best to produce value sets. If crisp parameters are estimated, accounting for imprecision changes the results little.

Keywords Quality-adjusted life years · Preference elicitation · Utility · Imprecise preferences · Time trade-off · Fuzzy modelling

Mathematics Subject Classification 91B08 · 91B16 · 03E72

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s11238-019-09707-6>) contains supplementary material, which is available to authorized users.

✉ Michał Jakubczyk
michal.jakubczyk@sggw.waw.pl

¹ Decision Analysis and Support Unit, SGH Warsaw School of Economics, Warsaw, Poland

² Department of Experimental and Clinical Pharmacology, Medical University of Warsaw, Warsaw, Poland

1 Introduction

Imagine you may live for 10 years in a wheelchair or T years in full health (both followed by sudden and painless death). What T would make you exactly indifferent? Would the choice be obvious if one extra day was added to such T , or a month, or a year?

In ordinary life, people make no explicit choices between health states or between health-related quality and longevity of life. Hence, their preferences may not be precisely established. Still, in thought experiments, trade-offs are used to assign utilities to health states, to subsequently support the appraisal of health technologies (e.g. Xie et al. 2016; Shirowa et al. 2016). In time trade-off (TTO) method, we elicit time, T , that makes the respondent indifferent between T years in full health vs 10 years in a hypothetical state; the first paragraph demonstrated how difficult that may be.

Imprecision of preferences is present in many decision contexts and can explain observed behaviours, e.g. some violations of the standard decision theory (Cubitt et al. 2015; Butler and Loomes 2007, 2011). Forcing respondents to answer may increase the dropout rate or reduce data quality (Stieger et al. 2007; Décieux et al. 2015). Additionally, when the actual value is not known precisely, how the question is framed may change the outcome.¹ In TTO, a pre-defined sequence of T s is used; imprecise preferences may cause the respondent to accept the first in a range of indistinguishably good answers, not the single best T . Thus, what sequence of T s is used may change the outcome. If it is imprecision that is responsible for this à la satisficing behaviour (see Simon 1956), then simply increasing the motivation of the respondents may not suffice; rather, they must be helped to understand their own preferences, e.g. via focus groups discussion prior to elicitation task. Also, the effect of imprecision may not decrease with sample size: if the sequence of T s results in lower value of the range being selected more often (e.g. in a bottom-up approach), then using more respondents will not help. Hence, getting respondents to understand own preferences better may outweigh getting more respondents.

That satisficing prevails in TTO can be seen from the indifference being found at all: a non-obvious fact, given a limited number of T s offered and the continuum of possible utility values. Whether the rejected T s provide useful information (on what was not satisfactory enough) was tested by Ramos-Goñi et al. (2017) in the interval regression. Still, the authors use a standard TTO protocol and assume that the disutilities of health state characteristics are regular numbers, implicitly assuming the preferences are precise.

In the present paper, we address the imprecision of health preferences and contribute to the literature in two ways. Firstly, we allow the respondents to directly report in TTO their answers as imprecise, by asking them to indicate (1) the range of values they consider as equally plausible as the standard, single-number outcome T and (2) the range of values they consider somewhat plausible (i.e. not entirely precluded). To define health states, we use the EQ-5D-3L descriptive system (Brooks and De Charro 1996): a health state is described by five dimensions (hence, 5D)—mobility (MO),

¹ For example, large differences were observed between the willingness to pay and willingness to accept, especially for non-market goods (Horowitz and McConnell 2002; Tunçel and Hammitt 2014), with which the respondents have little trade-off experience.

self-care (SC), usual activities (UA), pain/discomfort (PD), and anxiety/depression (AD)—each set on one of three levels (hence, 3L, 1 = no problems, 3 = extreme problems, see Table 1 in Supplementary Material for exact wording). Then, we may see how the characteristics of a health state are associated with the amount of imprecision, how capable of explaining the inconsistencies in TTO answers the imprecision is, and how much a path of rejected *T*s reveals about the degree of imprecision.

Secondly, we test several modelling approaches to assign disutilities to individual dimensions, accounting for the fact that TTO answers are imprecise, but also that the disutilities should be treated as imprecise numbers. To enable mathematical rigour in modelling, we treat utilities as fuzzy sets (notion introduced by Zadeh 1965). Our results show, e.g. that UA adds most imprecision to the valuation results—a factor that could be considered in working on a descriptive system.

Representing the imprecision as fuzzy sets/numbers has been used before in health preference context. Jakubczyk and Kamiński (2017) showed how to treat the willingness to pay/accept for health as fuzzy. Understanding the imprecision of health state utilities could be yet another factor to consider in sensitivity analysis in health technology assessment. Modelling these two forms of imprecision mathematically may make it easier to combine them in a meaningful way.

Regarding health state utilities, Jakubczyk et al. (2018b) estimated the disutilities of the EQ-5D-5L system (five levels per dimension) as fuzzy based on standard discrete choice experiment (DCE) data. In their study, the data came as choices between paired health states with duration; hence, the disutility of a state was not directly observed, contrary to the present study. Jakubczyk and Golicki (2018) estimated the disutilities as imprecise based on a path of answers in standard TTO. In accounting for the path, their study was similar to Ramos-Goñi et al. (2017), but the novelty lay in estimating the disutilities as imprecise. The limited amount of information derived from the path resulted in a very large error of estimates. As we show in the present paper, using a direct assessment of the imprecision from the respondents results in much narrower ranges, even in a smaller sample.

The structure of the paper is as follows. Because multiple analyses were performed, we did not distinguish the methods and results sections; instead, we split the paper by the area of analysis. In the next section, we present the questionnaire design and the characteristics of the sample used. Then (Sect. 3), we discuss the raw results of the questionnaire. In Sect. 4, we analyse the determinants of the amount of imprecision. Section 5 is central to the paper: we discuss how the disutility is shaped by the worsening in individual dimensions in the EQ-5D-3L descriptive system. The main results are further discussed in the last section.

2 Sample and questionnaire design

The experimental questionnaire consisted of three parts: (1) background questions, (2) own health valuation, (3) extended TTO experiment. In the first part, the respondents stated their gender (56% men), age (22.3 years on average, range 19–28), university and field/year of study (a convenience sample of 184 Polish students was recruited), and past experience with an illness (own or in a family; 51.7% had one).

In the second part, the respondents self-rated their own health with EQ-5D-3L (see Table 1 in Supplementary Material). Only nine EQ-5D-3L states occurred: 11111 (63.6%), 11112 (16.3%), 11122 (8.7%), 11121 (7.1%), and five more (no level 3). The respondents also rated their own health on a visual analogue scale (VAS), as a number between 0 and 100 (worst and best imaginable state, respectively). The mean VAS of 84.6 was obtained (median 85; standard deviation, SD, 8.61; range 60–100). The respondents rated being dead on the same scale (allowing the worst state to be worse than dead), resulting in mean 12 (median 10, SD 12.96, range 0–60).

To familiarize the respondents with assessing imprecision, they were asked to evaluate how convinced they were about selecting a given level in individual dimensions (from 0 to 100).² The respondent also provided a range of values representing equally plausible answers (henceforth, *epas*) and somewhat plausibly answers (henceforth, *spas*) as the single VAS score. Specifically, for *epas* the respondents were asked for a range of answers that seem just as good as the single one given (i.e. the range within which they cannot single out the most appropriate answer). For *spas*, the respondents were asked for a range of answers that are not precluded (the range outside of which the answer surely is not located). In each of both, the respondents were required to provide two numbers defining the ends of the range (without any particular elicitation protocol). The detailed results for *epas/spas* for VAS are in the Supplementary Material.

In the third and main part of the questionnaire, the respondents answered 13 TTO questions. The first three were warm-up tasks (living in a wheelchair, a mild 22111, and a severe 33333), to introduce to the TTO; hence, no ranges were asked for. A typically used procedure was used.³ If for state Q the procedure ended with T , we inferred the utility amounts to $u(Q) = T/10$. If a state was found worse than dead (WTD), then lead time version of TTO (LT-TTO) was used (Devlin et al. 2011): 10 years of full health with 10 years in Q were compared with T years of full health; then, for the result T we concluded $u(Q) = T/10 - 1$.

The remaining ten TTO experiments constituted the main part of a study. It differed from the standard TTO: when T was identified,⁴ the respondent was asked to provide a range of *epas* and *spas*, i.e. of values of T that seemed equally or somewhat plausible, respectively. The respondent was asked for ranges also in LT-TTO; however, the ranges were confined to either TTO or LT-TTO. The bounds of the *epas/spas* ranges can be used to calculate the analogous ranges for the utility, with formulas as in the previous paragraph.

In the main part, 17 health states were used and grouped in two blocks of 10 states (3 states in both blocks). The first block contained 11112, 11211, 12111, 21121, 11312, 32211, and 32313; the second block contained 11121, 21111, 11113, 11131, 11133,

² Other versions were tested: e.g. a Likert scale was used to measure the conviction, but respondents frowned at the option 'I disagree' with the just-selected option.

³ First using 10 years, then 0 and 5 years, and then 6, 7, ... or 4, 3, Half a year was added if needed.

⁴ If the respondents kept switching between neighbouring allowed T or insisted to, they were allowed to provide their own T , i.e. also not using half-a-year granularity. In 215 TTO experiments (11.7%), the respondents provided their answer not rounding to half a year. Even though this violated the original protocol, the answers were included as they came directly from the respondents.

13311, and 32223; health states 11311, 22222, and 23232 were in both blocks. Exactly half of the interviews used each block.

In design, milder states were preferred, due to foreseen difficulties with assigning ranges for WTD states. The health states were selected starting from a set of 17 states suggested by Macran and Kind (2000) (see also Lamers et al. 2006). We removed 33323 and 33333 (the latter was used in warm-up), for the milder states preference. Instead, 21121 and 11311 were introduced, to test how imprecision is aggregated when health problems in different dimensions are combined (e.g. how do the ranges for 11112, 11311, and 11312 relate). All the states were split by trial and error, so that the blocks are similar in average misery index (MI, the sum of levels, 8.3 and 8.2) and every level of every dimension is represented in each block.

Nine interviewers (graduate students) were recruited and were first subject to the test questionnaire. This testing helped to frame the questions and showed that ten states should not be exceeded due to fatigue. A convenience sample of interviewers' student colleagues was used. Before the interview, the respondent gave written consent to participate. The interviews were face to face in a quiet environment, with no computer assistance, but using printed cards to illustrate the TTO and printed health state descriptions. The interviewers were forbidden to go back (e.g. changing previous valuations, once when seeing consecutive state), to make the process more tractable.⁵

3 Data cleaning and raw results

The data cleaning and quality checks are described in the Supplementary Material. What is important for the present study is that in just three cases the respondents explicitly mentioned after the interview that they had found it difficult to provide answers for ranges, and in seven more the interviewers had the impression that the respondents had found it difficult. This problem was no more prevalent than claiming that state descriptions are unrealistic/internally inconsistent or not precise enough to imagine (10 interviewed).

The internal consistency was verified using the Pareto dominance: whether a given respondent assigned greater utility to an objectively worse state (i.e. having at least one dimension worsened, and no dimension improved). A serious violation was considered when the utility differed by at least 0.5 (as in EQ-5D-3L valuation in Poland by Golicki et al. 2010); in total, 129 such Pareto violations were identified (out of 184 respondents, all of whom valued 10 states, with a total of 3114 pairs of states Pareto comparable), i.e. 4.1%. Only 15 serious violations were present, and only two respondents had two of them (i.e. two pairs of states wrongly ordered). As Golicki et al. (2010) only removed observations if ten or more serious inconsistencies were present (out of 23 states per interviewed), we did not remove any data in the present study.

In 51 cases, the (non-serious) logical inconsistencies were explained away by the fuzzy approach in the sense that one of epas contained the other utility. In 65 cases (50.4%), the epas had a non-empty common part. In 83 cases (64.3%), the spas ranges

⁵ Just in one case, the respondent wanted to change VAS after having to value dead state on VAS (and was allowed).

Table 1 Raw results for point valuation

State	N	% NT	% WTD	Mean (95%CI)	SD	Median	Skew
11112	91	31.9	0.0	0.894 (0.866; 0.921)	0.133	0.95	−1.728
11113	94	10.6	8.5	0.596 (0.510; 0.682)	0.426	0.70	−2.134
11121	92	34.8	0.0	0.914 (0.888; 0.941)	0.128	0.95	−2.864
11131	91	11.0	17.6	0.436 (0.319; 0.553)	0.569	0.60	−1.316
11133	91	6.6	23.1	0.257 (0.134; 0.380)	0.597	0.50	−0.943
11211	92	22.8	0.0	0.864 (0.832; 0.895)	0.154	0.90	−1.522
11311	180	11.7	7.2	0.581 (0.514; 0.648)	0.459	0.70	−2.104
11312	92	4.3	8.7	0.500 (0.409; 0.590)	0.443	0.65	−1.995
12111	92	26.1	0.0	0.873 (0.842; 0.905)	0.153	0.90	−1.637
13311	89	3.4	21.3	0.282 (0.168; 0.397)	0.552	0.45	−1.097
21111	92	31.5	0.0	0.911 (0.887; 0.936)	0.120	0.95	−2.171
21121	92	16.3	0.0	0.810 (0.774; 0.846)	0.175	0.85	−1.176
22222	183	6.0	8.7	0.517 (0.452; 0.582)	0.447	0.60	−1.859
23232	91	1.1	42.9	−0.057 (−0.182; 0.068)	0.608	0.20	−0.283
32211	183	4.4%	14.8	0.356 (0.285; 0.427)	0.490	0.45	−1.360
32223	93	0.0%	34.4	0.020 (−0.087; 0.128)	0.528	0.15	−0.551
32313	92	1.1%	50.0	−0.122 (−0.244; 0.001)	0.598	0.02	−0.039

NT non-traders, WTD worse than dead, CI confidence interval, SD standard deviation

overlapped; hence, there was no Pareto violation in the interval order sense. Notice that due to the methodology, the ranges could not overlap if baseline utilities differed in sign, allowing for the epas/spas to cross zero, which might further reduce the number of inconsistencies.

Tables 1 and 2 present the descriptive statistics for the tested health states with respect to the point valuation and epas (see Table 2 in Supplementary Material for spas). As can be seen in Table 2, the average length of epas is typically two to five times larger than the standard error of the mean (SEM) of the utility. Only in about 20% the epas had zero length, suggesting there usually is some imprecision regarding utility. In about 30% of cases, the ranges (epas and spas) were equal, suggesting that the respondents in most cases were capable of differentiating between the two.

The ranges only included zero in 73 cases (4%) for spas and 35 cases (1.9%) for epas, which suggests that forcing epas/spas on one side of 0 was not a major limitation. More frequently (77 cases), the baseline utility amounted to -1 , the minimal LT-TTO-allowed value, possibly leading to censored observations.

Thanks to the pre-defined protocol, we know what values respondents rejected during TTO task (e.g. if the respondent settled for $T = 5.5$ years, they must have previously rejected 10, 0, 5, and 6 years). In as many as 56% of the valuations, at least one the rejected values belonged to the epas range, and in 15.6% of all valuations it was in the interior of epas. The proportions amounted to 74.7% and 42.8% for spas, respectively. Thus, rejecting a value in the course of TTO is not a valid indicator of clear preference.

Table 2 Raw results for equally plausible answers (SEM = standard error of the mean)

State	Range of means	Mean length ($\times$ SEM)	% Zero length	% Including 1 (0)
11112	(0.842; 0.923)	0.081 (5.8)	24.2	53.8% (0.0%)
11113	(0.530; 0.666)	0.137 (3.1)	10.6	21.3% (3.2%)
11121	(0.873; 0.943)	0.070 (5.2)	33.7	51.1% (0.0%)
11131	(0.387; 0.493)	0.106 (1.8)	17.6	18.7% (0.0%)
11133	(0.208; 0.327)	0.119 (1.9)	17.6	13.2% (1.1%)
11211	(0.808; 0.907)	0.099 (6.1)	18.5	43.5% (0.0%)
11311	(0.528; 0.638)	0.110 (2.3)	13.9	21.7% (1.1%)
11312	(0.444; 0.562)	0.118 (2.5)	15.2	10.9% (2.2%)
12111	(0.819; 0.916)	0.097 (6.1)	23.9	50.0% (1.1%)
13311	(0.234; 0.356)	0.122 (2.1)	18.0	6.7% (3.4%)
21111	(0.876; 0.939)	0.063 (5.0)	27.2	50.0% (0.0%)
21121	(0.761; 0.860)	0.099 (5.4)	12.0	35.9% (0.0%)
22222	(0.466; 0.578)	0.112 (2.4)	10.9	11.5% (0.5%)
23232	(− 0.100; 0.010)	0.109 (1.7)	14.3	4.4% (2.2%)
32211	(0.299; 0.423)	0.124 (2.4)	16.4	9.3% (4.4%)
32223	(− 0.019; 0.088)	0.107 (1.9)	18.3	1.1% (7.5%)
32313	(− 0.162; − 0.048)	0.114 (1.8)	14.1	4.3% (5.4%)

We checked how certain the non-traders (not trading off any life years in TTO, i.e. finishing with utility equal to 1) about this assigning maximal utility. Out of tasks resulting in non-trading, we measured the percentage of situations where the lower end of spas/epas did differ from 1, i.e. allowed some trading-off: it happened in 34.7% and 46.2%, respectively.

4 What drives the imprecision

The lengths of epas/spas differ between health states (see Table 2 and Table 2 in Supplementary Material). In Supplementary Material, we show that the imprecision is subadditive, i.e. the length of epas/spas for the compound state (e.g. 11133) is smaller than the sum of the lengths of the building states (i.e. 11131 and 11113). Therefore, for epas and spas, separately, we built three types of models explaining their length. One used only the dummy variables for dimensions/levels as in standard disutility modelling that served as a benchmark model. The other model also used five derived variables, to account for subadditivity: MI minus 5 (to measure the departure from 11111), the maximal level across dimensions (maxLvl) minus 1, the number of dimensions at maxLvl (maxCnt) minus 1, the minimal level (minLvl), and the count of dimensions at minLvl (minCnt). The third model also used the estimated point disutility, $1 - \hat{u}$ (panel random effects model, A1 in Table 4).

In the first type of model, the statistically insignificant variables were removed one by one (including the intercept, if needed). Then, in the other types, the remaining variables were added to the pool of regressors and removed if insignificant. The pre-

Table 3 Explaining the length of equally/somewhat plausible answers (epas/spas)

Parameter	Model				
	epas			spas	
	Type 1	Type 2	Type 3	Type 1	Type 2
Const	0.094	0.058	0.051	0.157	0.065
UA2	0.020	0.017	0.015	0.036	0.022
UA3	0.018			0.047	
PD2	− 0.010				
PD3				0.028	
AD3	0.017			0.027	
MI−5					− 0.012
maxLvl−1		0.037	0.049		0.088
maxCnt−1		0.010	0.012		0.022
minLvl		− 0.029	− 0.017		
minCnt−1		0.005			
$1 - \hat{u}$			− 0.036		
R^2	0.023	0.040	0.042	0.035	0.060

viously removed dummies for dimensions/levels were tested again, in the widened set of regressors. Panel random effects models were applied; robust variance–covariance matrices were used to determine significance.

Table 3 shows the results. For epas, the third type results in $1 - \hat{u}$ being insignificant; hence, this model was not constructed. UA2 is present in all models with positive coefficient and UA3 in both models using only dimension/level dummies. This fact suggests that imprecision is especially driven by the UA dimension. UA is followed by AD in driving the imprecision in valuation. We find it difficult to interpret the negative sign by PD2. Perhaps, this result is an artefact of the selection of states and the subadditivity (this term disappears in models with derived variables).

Increasing the maximal level increases the imprecision; so does adding dimensions at this maximal level: there is more imprecision for severe states. The negative sign by minLev can be interpreted: the increase in minLvl decreases the spread between the dimensions and apparently makes it easier to assign a utility. At the same time, the positive sign by minCnt confirms: the more dimensions are at the minimal level, the more internally diverse the health state is; hence, there is larger imprecision. That it is more difficult to precisely value an internally inconsistent state may explain why the imprecision is reduced when 11133 is considered as compared to 11131 and 11113.

Adding $\hat{u}$ yields a less interpretable model: severity of a health state increases imprecision via maxLvl but decreases it via disutility, $1 - \hat{u}$.

5 Estimating dimensions impact on (fuzzy) utility

In this central section of the paper, we check how individual EQ-5D-3L dimensions contribute to the overall disutility, when accounting for the imprecision of preferences.

To this goal, we build three classes of models (in each, several specifications are used). First, in the crisp–crisp models we neglect the imprecision altogether: both the dependent variable and the parameters are regular (crisp) numbers (the independent variables are necessarily crisp in all the approaches). This approach serves as a benchmark.

In the fuzzy–crisp models, we account for the imprecision of the TTO outcome, but estimate crisp model parameters. This approach serves to build a standard value set, making it easier to support decision making. Finally, in the fuzzy–fuzzy models, we estimate the model parameters (i.e. the disutilities of individual dimensions) as imprecise.

The crisp–crisp models use standard econometric methods and are only briefly presented in Sect. 5.1. Introducing the other classes requires some information on how imprecision is modelled with fuzzy sets, as done in Sect. 5.2. The multitude of approaches used serves to see which modelling technique is most suitable for health state utility data.

5.1 Crisp–crisp models

As usually done for convenience, we model as the dependent variable the disutility of a health state: $1 - u(Q)$. We use as independent variables the dummies, $d_{i,j}$, denoting if dimension i is at level j ; the corresponding parameters, $\alpha_{i,j}$ are disutilities of worsening dimension i to level j (from level 1). To make interpretation easier, in tables the parameters are denoted with dimension abbreviation and level number, e.g. MO2 instead of $\alpha_{1,2}$. Also a constant term, α_0 , is used in some specifications.

The results are presented in Table 4. First, we quote the valuation study Golicki et al. (2010) in the Polish population, to inspect possible differences due to the specific target group in the present study. The remaining models are based on the current sample. In model A1, we used the random effects panel model (with robust variance–covariance matrix), one of the standard approaches in this type of models.⁶

In A2, we estimated an analogous model using ordinary least squares (OLS) to test the impact of neglecting the panel structure, as done in the fuzzy–crisp/fuzzy approaches. In A3, we dropped α_0 , again, to match some fuzzy–fuzzy specifications that assign the imprecision entirely to individual dimensions, for a more convenient interpretation. Finally, in A4 we used median regression (Koenker 2005), for reasons explained in the next subsection.

Firstly, in the present sample, the respondents attach greater disutility to most of the dimensions/levels than the general population (except for PD3, where the estimation error is largest). This may be due to the sample consisting of young, healthy individuals, scared of health problems due to lack of such experience. This is particularly true for UA, suggesting that students value their free time a lot. Fortunately, the representativeness of the sample is not essential in the present study, as we do not focus on the exact values of utilities but rather on mechanisms present and methodological approaches to detect them. Still, our results are not drastically different from those of Golicki et al. (2010).

⁶ We deem it beyond the scope of the present paper to discuss how optimally to model TTO data, e.g. how to account for respondents' heterogeneity.

Table 4 Coefficients of the crisp–crisp models (with standard errors)

Variable	Golicki et al.	A1	A2	A3	A4
Const.	0.049 (0.018)	0.049 (0.014)	0.051 (0.016)	–	– 0.012 (0.011)
MO2	0.052 (0.011)	0.061 (0.022)	0.060 (0.022)	0.077 (0.020)	0.062 (0.013)
MO3	0.331 (0.014)	0.364 (0.036)	0.360 (0.041)	0.358 (0.042)	0.338 (0.030)
SC2	0.054 (0.012)	0.095 (0.020)	0.095 (0.024)	0.108 (0.023)	0.112 (0.017)
SC3	0.235 (0.015)	0.290 (0.038)	0.295 (0.051)	0.284 (0.051)	0.250 (0.039)
UA2	0.046 (0.014)	0.125 (0.017)	0.125 (0.023)	0.146 (0.024)	0.112 (0.015)
UA3	0.212 (0.014)	0.367 (0.029)	0.363 (0.029)	0.411 (0.026)	0.312 (0.024)
PD2	0.057 (0.011)	0.063 (0.016)	0.067 (0.023)	0.084 (0.022)	0.062 (0.011)
PD3	0.489 (0.012)	0.456 (0.043)	0.458 (0.043)	0.494 (0.042)	0.350 (0.043)
AD2	0.026 (0.013)	0.082 (0.018)	0.076 (0.023)	0.077 (0.022)	0.062 (0.016)
AD3	0.207 (0.012)	0.278 (0.029)	0.281 (0.034)	0.296 (0.034)	0.262 (0.028)

A1 random effects panel model, A2 ordinary least squares (OLS) model, A3 OLS with no constant term, A4 median regression, *MO* mobility, *SC* self care, *UA* usual activities, *PD* pain/discomfort, *AD* anxiety/depression

Secondly, the difference between panel and OLS regression is negligible from a practical point of view, a convenient fact. Thirdly, the median regression yields lower parameters (and a negative constant term). An unsurprising finding, accounting for the negative skewness of health states' utilities (e.g. due to frequent non-trading for several states): the median regression establishes how the middle person's utility changes when dimensions are worsened; as the substantial fraction of the respondents still assigns high utility, the median decreases slower than the mean.

5.2 Fuzzy–crisp models

In this second class of models, we account for the TTO answers being usually imprecise (i.e. epas/spas ranges having non-zero length). Hence, the elicited (dis)utility is also imprecise. To formally model it, we use fuzzy sets (only basic introduction is offered; for details, see, e.g. Kilir and Yuan 1995).

A fuzzy set $\tilde{X}$ over real numbers (the tilde denotes fuzziness) is represented by a membership function: $\mu_{\tilde{X}}(\cdot) : \mathbb{R} \rightarrow [0, 1]$ (we drop the subscript in notation, if clear from the context); $\mu_{\tilde{X}}(x)$ is interpreted as the measure of belongingness of x to $\tilde{X}$. If $\mu_{\tilde{X}}(x) = 1$ for a single x and $\mu_{\tilde{X}}(\cdot) = 0$ otherwise, then $\tilde{X}$ is a regular, crisp number (equal to x). With fuzzy numbers, more than one value can be perceived by a respondent as a possible (dis)utility of a given state, and some values are less credible but not entirely ruled out.⁷

Equivalently, $\tilde{X}$ is defined by α -cuts: $A_{\tilde{X},\alpha} = \{x \in \mathbb{R} : \mu_{\tilde{X}}(x) \geq \alpha\}$, for $\alpha \in (0, 1]$, or strong α -cuts: $A_{\tilde{X},\alpha+} = \{x \in \mathbb{R} : \mu_{\tilde{X}}(x) > \alpha\}$, for $\alpha \in [0, 1)$. Then, A_1

⁷ In the paper, we do not differentiate between ontic or epistemic interpretation (Couso and Dubois 2014): respectively, whether the respondent believes there are multiple values of $u(Q)$ (e.g. because of ambiguity what being in Q truly means), or whether the respondent believes there is a single value $u(Q)$, only not being perfectly accessible through introspection into own preferences.

Fig. 1 An example of a fuzzy number representing health state disutility

is a (regular) set of values that belong to $\tilde{X}$ with full conviction, and A_{0+} is a set of values that belong thereto with non-zero conviction. In our study, both cuts can be directly inferred from epas and spas.

Lastly, we assume the membership function is continuous and linear where possible. In fuzzy set parlance, the disutility is a normal ($\max \mu = 1$) and trapezoidal (shape of μ) fuzzy number (α -cuts are convex).

With our assumptions, fuzzy disutility can be denoted by a four-tuple (l, L, H, h) , where $[L, H]$ are the bounds of A_1 and $[l, h]$ are the bounds of A_{0+} . An example of a fuzzy disutility, $(0, 0.2, 0.4, 0.4)$, is shown in Fig. 1.

In all models below, we ignore the panel structure; the observations for individual respondents are pooled and indexed by k . For each k , we have the four-tuple (l, L, H, h) denoting the observed, empirical disutility and, hence, the membership function, μ (we omit index k for clarity).

In fuzzy–crisp models, we assume model parameters, α_0 and $\alpha_{i,j}$, are crisp. Then, the theoretical disutility, $\widehat{1-u}$, is also crisp and given by:

$$\widehat{1-u} = \alpha_0 + \sum_{i,j} \alpha_{i,j} \times d_{i,j}.$$

Fitting models is done by setting such parameter values to make $\widehat{1-u}$ as close to (l, L, H, h) as possible, in a given sense. In B1, we minimize the sum (over k) of squared distance between $\widehat{1-u}$ and $(L+H)/2$. Hence, the imprecision is accounted for by basing the model on epas range (an alternative model for spas could be built), yet the range is reduced to considering its middle point.

In B2, the whole four-tuple is used, and we maximize the overall compatibility of the theoretical values given the imprecise preferences, i.e. we maximize the expression $\sum_k \mu(\widehat{1-u})$, (see also Celmiņš 1987). This model can be written as a linear optimization problem; still, as the resulting problem is overwhelmingly large, we numerically maximized the non-linear function with the `DEoptim` package in **R**. Therefore, the global optimum might not have been found.

Table 5 Coefficients of the fuzzy–crisp models (and standard errors, where available)

Variable	B1	B2	B3	B4
Const.	0.058 (0.016)	0.035	0.060	0.072
MO2	0.057 (0.022)	0.015	0.052	0.040
MO3	0.356 (0.041)	0.226	0.358	0.343
SC2	0.091 (0.024)	0.024	0.091	0.093
SC3	0.288 (0.051)	0.200	0.282	0.280
UA2	0.121 (0.023)	0.015	0.137	0.143
UA3	0.352 (0.029)	0.166	0.369	0.376
PD2	0.063 (0.022)	0.019	0.063	0.064
PD3	0.448 (0.043)	0.294	0.464	0.470
AD2	0.078 (0.022)	0.005	0.081	0.079
AD3	0.274 (0.034)	0.182	0.279	0.284

MO mobility, *SC* self care, *UA* usual activities, *PD* pain/discomfort, *AD* anxiety/depression

In *B3*, we extend the standard least squares by only calculating the error term if the theoretical value is beyond the $[L, H]$ range; formally, we use the directed Hausdorff distance of $\widehat{1 - u}$ from $[L, H]$, $d(\widehat{1 - u}, [L, H])$:

$$d(a, [b, c]) = \inf_{y \in [b, c]} |a - y|, \quad (1)$$

and minimize the sum of squared such distances. Finally, *B4* uses $[l, h]$ instead of $[L, H]$. From the technical side, the *B3* and *B4* approaches can be presented as quadratic programming tasks.

The results of the above specifications are collected in Table 5. The results of *B1*, *B3*, and *B4* are very similar between themselves and to the results of crisp–crisp models. Apparently, the imprecision around the TTO outcome has little impact when the modelling process is set to yield crisp values.

However, the results of *B2* are very different: much lower disutilities were found for all dimensions/levels (and the constant term), and we offer the following explanation. *B2* approach strives to set parameters to make the theoretical value fall into the range of as many respondents as possible; in this sense, *B2* can be viewed as mode-oriented regression. The disutilities of health states are positively skewed, and for variables with positive skew we often observe mode < median < mean. Analogously in our case, *B2* yields lower disutilities than *A4* (median regression), and *A4* lower than *A2*.

We illustrate the above phenomenon for state 32313 in Fig. 2: point answers and epas for all respondents are presented (spas left out for clarity). The average point utility amounts to -0.122 , and the median is greater (between 0.0 and 0.05, due to the even number of observations). To cross the maximal number of ranges with a horizontal line, we need to draw it at the utility of 0.5 (we have 13 crosses then), more than a median. Additionally, this cross-maximizing level can radically change with just a small change of the data. Also, the censoring of the answers at -1 (it is not possible to report lower utility) may lead to the grouping of the answers and impact the

Fig. 2 Results for state 32313 (EPAS drawn as black boxes, SPAS left out for clarity). The mean utility denoted with the horizontal dashed line (equal to -0.122); the median for the (sorted by point utility) respondents denoted by the vertical dotted line (between 0.0 and 0.05); the horizontal line for the utility of 0.5 crosses most of the EPAS ranges (15)

results. Thus, the results of the compatibility approach seem challenging to interpret and unstable.

5.3 Fuzzy-fuzzy models

In this third class of models, we attribute imprecision to individual dimensions; the results are collected in Table 6. In C1, we use a simplified, naïve approach: we build two separate OLS models for L and H , i.e. estimate two sets of crisp parameters $\alpha_{i,j}$ and α_0 .

In the remaining approaches, we explicitly redefine model parameters to be fuzzy numbers: $\tilde{\alpha}_0$ and $\tilde{\alpha}_{i,j}$. To conform with the definition of the dependent variable, we assume $\tilde{\alpha}$ s are trapezoidal and normal; hence, they can be represented by four-tuples of numbers. Calculating theoretical disutility requires multiplying parameters by dummies and adding the results. Standard fuzzy arithmetic yields: $1 \times (a, b, c, d) = (a, b, c, d)$, $0 \times (a, b, c, d) = (0, 0, 0, 0)$, and $(a, b, c, d) + (e, f, g, h) = (a + e, b + f, c + g, d + h)$. Then, the theoretical disutility is also fuzzy, and is denoted by $(\hat{l}, \hat{L}, \hat{H}, \hat{h})$. The models are estimated by making $(\hat{l}, \hat{L}, \hat{H}, \hat{h})$ as close to the observed (l, L, H, h) (for all k) as possible.

The fitting of the model requires calculating distances between trapezoidal fuzzy numbers, to minimize the sum over k of squared distances, and we use two methods. In the first one, we use the supremum of Hausdorff distance between their respective α -cuts of two fuzzy numbers (see Voxman 1998): $d(\tilde{X}, \tilde{Y}) =$

Table 6 Results of the fuzzy–fuzzy models (notation explained in the text)

Variable	C1	C2	C3
Intercept	(0.011; 0.103)	(0.002, 0.002, 0.102, 0.128)	–
MO2	(0.061; 0.054)	(0.065, 0.065, 0.065, 0.065)	(0.075, 0.075, 0.100, 0.111)
MO3	(0.353; 0.366)	(0.351, 0.351, 0.351, 0.351)	(0.347, 0.347, 0.347, 0.347)
SC2	(0.085; 0.098)	(0.069, 0.085, 0.085, 0.085)	(0.083, 0.088, 0.113, 0.118)
SC3	(0.282; 0.287)	(0.276, 0.276, 0.276, 0.299)	(0.276, 0.276, 0.276, 0.276)
UA2	(0.112; 0.131)	(0.115, 0.115, 0.115, 0.115)	(0.104, 0.104, 0.154, 0.169)
UA3	(0.345; 0.364)	(0.347, 0.347, 0.347, 0.349)	(0.343, 0.343, 0.443, 0.468)
PD2	(0.064; 0.056)	(0.051, 0.059, 0.059, 0.061)	(0.053, 0.073, 0.098, 0.107)
PD3	(0.442; 0.450)	(0.428, 0.428, 0.428, 0.429)	(0.423, 0.423, 0.498, 0.521)
AD2	(0.084; 0.083)	(0.079, 0.079, 0.079, 0.079)	(0.086, 0.086, 0.086, 0.086)
AD3	(0.262; 0.279)	(0.266, 0.266, 0.266, 0.288)	(0.277, 0.277, 0.302, 0.323)
Variable	C4	C5	
Intercept	(0.000, 0.015, 0.106, 0.150)	–	
MO2	(0.056, 0.056, 0.056, 0.056)	(0.064, 0.064, 0.089, 0.097)	
MO3	(0.352, 0.352, 0.352, 0.352)	(0.349, 0.349, 0.349, 0.349)	
SC2	(0.081, 0.090, 0.092, 0.092)	(0.080, 0.092, 0.123, 0.130)	
SC3	(0.283, 0.287, 0.287, 0.287)	(0.271, 0.271, 0.271, 0.271)	
UA2	(0.097, 0.116, 0.130, 0.130)	(0.099, 0.124, 0.175, 0.192)	
UA3	(0.322, 0.342, 0.362, 0.369)	(0.325, 0.359, 0.459, 0.506)	
PD2	(0.059, 0.062, 0.062, 0.062)	(0.054, 0.069, 0.095, 0.116)	
PD3	(0.420, 0.442, 0.452, 0.456)	(0.420, 0.455, 0.526, 0.561)	
AD2	(0.075, 0.075, 0.075, 0.075)	(0.075, 0.076, 0.078, 0.078)	
AD3	(0.254, 0.266, 0.282, 0.285)	(0.256, 0.270, 0.312, 0.328)	

In C1 the results of the L and H modelling is given in parentheses; in other models, the quadruples presented are (l, L, H, h)

MO mobility, *SC* self-care, *UA* usual activities, *PD* pain/discomfort, *AD* anxiety/depression

$\sup_{\alpha \in (0,1]} d_H(A_{\alpha, \tilde{X}}, A_{\alpha, \tilde{Y}})$, where $d_H(\cdot, \cdot)$ is a standard Hausdorff distance. In our case of trapezoidal fuzzy numbers, the distance can be shown to simplify to $\max(|l - \hat{l}|, |L - \hat{L}|, |H - \hat{H}|, |h - \hat{h}|)$. We use this distance in C2 and C3, with and without a constant term $\tilde{\alpha}_0$, respectively.

A possibly counter-intuitive and unwanted property of d_H is its insensitivity to some changes in the considered intervals:

$$d_H([0.1, 0.2], [0.2, 0.5]) = 0.5 - 0.2 = 0.3 = d_H([0, 0.2], [0.2, 0.5]),$$

even though we may think $[0, 0.2]$ differs more from $[0.2, 0.5]$ than $[0.1, 0.2]$ does. For this reason, we use another distance between four-tuples, directly comparing the corresponding elements of two four-tuples (see Chachi and Taheri 2016, for a similar distance for intervals):

$$d_I \left((l, L, H, h), (\hat{l}, \hat{L}, \hat{H}, \hat{h}) \right) = \sqrt{(l - \hat{l})^2 + (L - \hat{L})^2 + (H - \hat{H})^2 + (h - \hat{h})^2}.$$

We use $d_I(\cdot, \cdot)$ in C4 and C5, with and without a constant term $\tilde{\alpha}_0$, respectively. Models C2–C5 can be written as quadratic programming problems.

A counter-intuitive property of C1 is that some coefficients when modelling the lower bound of the disutility range (L) are greater than for the upper bound (H); hence, it is impossible to directly interpret the results as the imprecise disutilities of individual dimensions/levels.

In the remaining approaches, this difficulty does not arise. However, some parameters reduce to crisp numbers (e.g. MO3 in C2), especially when intercept captures most of the imprecision. Models with no constant term (C3 and C5) allow better to assign imprecision to the individual dimensions. Conforming with the results presented in Table 3, UA followed by PD is responsible for the largest part of the spread in answers. Looking at absolute values, the ranges of disutilities conform with the results of the standard crisp–crisp modelling in Table 4: PD3 causes the largest disutility, followed by UA3 and MO3.

5.4 Value sets

Supporting health technology appraisal requires not only understanding the impact of dimensions on disutility, but we also need to assign utility values to all the health states in the descriptive system, i.e. to construct a value set. We constructed and present in Fig. 3 the value sets constructed for selected types of models: standard panel model (A1, health state ordered decreasingly with this utility), compatibility (B2), fuzzy–crisp least squares (B3), naïve approach (C1, two lines), fuzzy–fuzzy Hausdorff distance model with no intercept (C3, gray area based on L – H range only). We also presented the result of a combined modelling of epas middle point (B1) and its length (Table 3, type 3 model). The value sets constructed based on fuzzy–fuzzy

Fig. 3 Value sets for selected models

models are also fuzzy: the utility of each health state is given as a fuzzy number (for clarity only $L-H$ range is presented in the figure).

For most models, the resulting value sets offer similar utilities. Notably, the compatibility approach used in $B2$ results in much higher utility values (due to the explanation provided above). Regular crisp–crisp ($A1$) and fuzzy–crisp ($B3$) approaches yielded nearly identical values. In this sense, imprecision of preferences used as input does not matter as long as we require a crisp output.

However, if we transport the imprecision to the value set, we obtain a band (not a line) of values that all can equally well be interpreted as utilities of consecutive health states. For fuzzy least squares ($C3$), for some health states there seemingly is no imprecision, and the band narrows to a single point (if only dimensions where disutilities degenerated to crisp numbers are worsened). The naïve approach ($C1$) and modelling the middle and length of the interval separately yield similar results; hence, the unwanted parameter reversal in $C1$ does not pose a problem for value set construction.

6 Concluding remarks

In the present paper, we studied the amount of imprecision in the preferences for health states. We showed that the imprecision of the utility values can be elicited by a simple extension of the standard TTO protocol. Preferences are indeed imprecise: there is typically a range of answers the respondent considers equally plausible. Importantly, this imprecision is not stochastic in nature: it will not diminish with increasing sample size (and even for a moderate sample used in the present study, the imprecision is larger than the standard error due to sample randomness). Hence, focusing solely on sample size or number of states used may be insufficient to increase the credibility of the results; it could be worthwhile to add elements that increase respondents' awareness of their own preferences (e.g. a qualitative part in the beginning of a study to discuss what aspects of health one values in life).

Some studies (Ramos-Goñi et al. 2017; Jakubczyk and Golicki 2018) notice the problem of imprecision in preferences and approach it by studying the paths of responses in TTO. As we show, the actual imprecision is often larger than what was rejected in TTO would suggest (as presented in Sect. 3, in $> 50\%$ a rejected value was included in the range of plausible answers). Perhaps, respondents need some time or other framing of the question to realize that the previously offered T values were acceptable; or perhaps respondents tend to only accept T being closer to the middle of the epas range. That comparing alternatives and valuing are two different tasks and sometimes their results diverge were observed in the literature (see, e.g. Butler et al. 2014; Grether and Plott 1979).

In particular, when respondents in TTO accept the first offered value of $T = 10$ years, the implied utility equals 1 (non-trading, present in $> 30\%$ cases for some mild states), and there is no path on which to base the inference (which might explain wide intervals in Jakubczyk and Golicki 2018). Asking for epas/spas directly helps to understand the non-trading behaviour better. As we found, in 34.7% of the non-trading cases, the respondents actually felt trading off some years as an equally good

representation of their preferences, and in 46.2% they did not rule out the trading off completely. This finding shows that observed non-trading may be largely due not to preferences being lexicographic (longevity first), but to the TTO protocol (starting with 10 years). As non-trading poses some technical challenges in modelling utilities (e.g. whether the utilities should be censored at 1, as done in Devlin et al. 2018), our finding suggests modifications to the protocol may be useful (e.g. using a finer set of answers near 10 years).

That the imprecision can explain some inconsistencies in answers (a dominated state being assigned greater utility by a given respondent) suggests that the standard quality checks may be overly conservative. This is not to say that respondents with inconsistent answers ought not to be removed from the sample (inconsistent answers as input may lead to counter-intuitive results), but perhaps inconsistencies in TTO should not be immediately treated as a sign of poor job by an interviewer or lead to also discarding the DCE data, if both are collected.

We found the preferences are typically not so imprecise to make respondents unsure if a given state is better or worse than dead (in only a few cases the epas/spas ranges contain 0 utility). This is a rather optimistic finding; for example, a discrete choice experiment between health states with no duration (another elicitation procedure, often paired with TTO) allows to estimate the utility values only on a latent scale that needs to be subsequently anchored at 0 value (dead) (see Norman et al. 2016, for information on anchoring), and apparently directly comparing health states versus dead can be relatively precisely done. Also other, novel approaches to estimating health state utilities (e.g. personal utility function presented in Devlin et al. 2019) require locating being dead in a set of ranked states; our finding supports this task can be done rather credibly.

In the paper, we demonstrated what state characteristics drive the imprecision mostly. Respondents are particularly unclear about the impact of usual activities, followed by pain/discomfort. This may result from usual activities being a wide concept (very different activities might be included—e.g. work, free time, family obligations) and most subjective (respondent must make a decision what activities they include in considerations). If the health state is internally diversified (some dimensions at level 1, and some at level 3), then it also increases the imprecision in how the states are valued.

The analysis of imprecision may help in developing descriptive systems; obviously, the EQ-5D-3L/5L descriptive system is well established and ought not to be changed due to the present study, but the analysis of imprecision may help when developing other systems or bolt-ons, i.e. additional dimensions added, e.g. for specific patient groups (see, e.g. Yang et al. 2015). In a larger and more diversified sample, one might verify how confident respondents feel with reporting problems in individual dimensions (i.e. re-estimate the model presented in Supplementary Material).

Also, the results presented here suggest what states are best used in valuation studies. As internal consistency increases the precision in assessment, it makes sense to focus on less internally diversified states. Also, the imprecision seems to be subadditive (the overall imprecision is smaller than the sum of its parts when worsening in two dimensions are compounded). This observation motivates using severe states in valuation studies.

In the paper, we tested several econometric approaches. Our findings suggest that for different purposes, different classes work best. To attribute the imprecise disutilities to individual dimensions, probably the fuzzy–fuzzy least squares with no intercept are best suited (C3 and C5), because in this approach, we directly attribute the imprecision of utility to parameters for individual dimensions/levels. However, if we want to construct a fuzzy value set, then separately modelling the middle and the length of the α -cuts seems most convenient, as it avoids the problem of possible sub-additivity of imprecision when problems in different dimensions are combined (see Supplementary Material). The naïve models proved similar in producing value sets, but modelling separately the location and the length has an additional advantage that it is possible to select the best class of model (e.g. semi-log) for both elements separately.

The compatibility approach (B2) seems least suited, as it yields counter-intuitive (and difficult to interpret) results with very low disutilities. On the one hand, that is a disappointment as the membership function of the fuzzy disutility is not used in other approaches (only the bounds of the two α -cuts). On the other hand, the assumptions regarding the actual shape of the membership function (is it trapezoidal) are less important and the findings are more robust in this sense.

The results of some modelling approaches are probably influenced by the selection of states used in the survey. For example, level 3 in SC was only used in three states and always in combination with some other dimension at level 3; this fact may explain the degenerate result of C3 and C5. On the one hand, if worsening in two dimensions usually coincides, it might be difficult to separate the impact of each on the imprecision. On the other hand, worsening only one of two dimensions to level 3 may make the state hardly imaginable, internally inconsistent, and difficult to assign utility. In future studies, probably a wider pool of states is needed. It might also be easier to use a richer EQ-5D-5L descriptive system in future studies, as the levels differ more subtly and most variation in dimensions can be introduced without making the state internally implausible (3L version was used here, as we wanted to cover a reasonable fraction of all the states with the design used).

In the present study, we neglected the problem of the estimation error (except for models where standard econometric techniques were used). Still, how the results of various approaches coincide is reassuring for the point estimates obtained. In future research, bootstrap methods could be used, for example, to inspect the impact of sample randomness on the estimated parameters.

The impact of imprecision on parameters is negligible, when crisp parameters are required as output (e.g. results of A1 and B1 or B3 are rather similar). This finding is rather fortunate for the whole area of research: if we want to estimate standard (crisp) value sets to support decisions on health technology provision, we may as well neglect preference imprecision. On one hand, that finding seems a bit unfortunate for the present paper (but was not known before the study). On the other hand, the imprecision of preferences should probably make us think more about the estimand (if not about the estimation process). If preferences are imprecise, what is it that we would like to maximize?

It is a bit paradoxical that we ask the respondents to report their imprecision in the form of (precise) numbers. The imprecision about how to quantify one's imprecision

can be modelled formally (see Zadeh 1975, for type-2 fuzzy sets). We believe, however, there is a trade-off between trying to learn more about the respondent and having to ask questions that may be unintelligible (asking for a range of values describing the lower bound of the epas range). Additionally, we believe that if the respondent is unsure about, say, lower bound of epas, they can report it more conservatively (i.e. report a lower bound of the range in a way accounting for the whole range of possible lower bounds of epas).

We have to acknowledge, however, that there are probably other forms of imprecision that the respondent cannot access by introspection. How questions are framed or what protocol is used to elicit responses may change the answers, as observed in other experimental contexts Beattie et al. (1998). This is especially true if we assume the preferences for health states are not available a priori but are only created during the experiment (just like regular markets shape the preferences for goods, see Isoni et al. 2016). Our present study is limited to studying only one type of imprecision. Still, perhaps the amount of imprecision (as defined in the present study) can serve as a proxy for measuring stability of preferences in repeated tasks. Then, future studies could inspect if weighting observations based on imprecision improves the quality of produced models (defined, e.g. by predictive validity, see Jakubczyk et al. 2018a).

Acknowledgements This research was financed by a grant from The EuroQol Research Foundation; however, the views expressed here are of the authors and not necessarily of the Foundation or of the EuroQol Group. We would like to thank the interviewers—E. Czeremużyńska, E. Grabowska, K. Jama, M. Karpiuk, J. Krawczyńska, M. Margielewicz, M. Musialik, M. Sternik, and M. Tuszyńska—and the respondents (valuing health states can be depressing as much as it is imprecise). We thank those who gave their comments at conferences and seminars where this paper was presented.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Beattie, J., Covey, J., Dolan, P., Hopkins, L., Jones-Lee, M., Loomes, G., et al. (1998). On the contingent valuation of safety and the safety of contingent valuation: Part 1-caveat investigator. *Journal of Risk and Uncertainty*, 17, 5–25.
- Brooks, R., & De Charro, F. (1996). EuroQol: The current state of play. *Health Policy*, 37, 53–72.
- Butler, D., & Loomes, G. (2007). Imprecision as an account of the preference reversal phenomenon. *American Economic Review*, 97(1), 277–297.
- Butler, D., & Loomes, G. (2011). Imprecision as an account of violations of independence and betweenness. *Journal of Economic Behavior and Organization*, 80, 511–522.
- Butler, D., Isoni, A., Loomes, G., & Navarro-Martinez, D. (2014). On the measurement of strength of preference in units of money. *Economic Record*, 90, 1–15.
- Celmiņš, A. (1987). Least squares model fitting to fuzzy vector data. *Fuzzy Sets and Systems*, 22, 245–269.
- Chachi, J., & Taheri, S. (2016). Multiple fuzzy regression model for fuzzy input-output data. *Iranian Journal of Fuzzy Systems*, 13, 63–78.
- Couso, I., & Dubois, D. (2014). Statistical reasoning with set-value information: Ontic vs. epistemic views. *International Journal of Approximate Reasoning*, 55, 1502–1518.
- Cubitt, R., Navarro-Martinez, D., & Starmer, C. (2015). On preference imprecision. *Journal of Risk and Uncertainty*, 50, 1–34.

- Décieux, J., Mergener, A., Neufang, K., & Sichka, P. (2015). Implementation of the forced answering option within online surveys: Do higher item response rates come at the expense of participation and answer quality? *Psihologija*, 48, 311–326.
- Devlin, N., Tsuchiya, A., Buckingham, K., & Tilling, C. (2011). A uniform time trade off method for states better and worse than dead: Feasibility study of the ‘lead time’ approach. *Health Economics*, 20, 348–361.
- Devlin, N., Shah, K., Feng, Y., Mulhern, B., & van Hout, B. (2018). Valuing health-related quality of life: An EQ-5-5L value set for England. *Health Economics*, 27, 7–22.
- Devlin, N., Shah, K., Mulhern, B., Pantiri, K., & van Hout, B. (2019). A new method for valuing health: directly eliciting personal utility functions. *The European Journal of Health Economics*, 20, 257–270.
- Golicki, D., Jakubczyk, M., Niewada, M., Wrona, W., & Busschbach, J. (2010). Valuation of EQ-5D health states in Poland: First TTO-based social value set in Central and Eastern Europe. *Value in Health*, 13, 289–297.
- Grether, D., & Plott, C. (1979). Economic theory of choice and the preference reversal phenomenon. *American Economic Review*, 69, 623–638.
- Horowitz, J., & McConnell, K. (2002). A review of WTP/WTa studies. *Journal of Environmental Economics and Management*, 44, 426–447.
- Isoni, A., Brooks, P., Loomes, G., & Sugden, R. (2016). Do markets reveal preferences or shape them? *Journal of Economic Behavior and Organization*, 122, 1–16.
- Jakubczyk, M., & Kamiński, B. (2017). Fuzzy approach to decision analysis with multiple criteria and uncertainty in health technology assessment. *Annals of Operations Research*, 251, 301–324.
- Jakubczyk, M., & Golicki, D. (2018). Estimating the fuzzy trade-offs between health dimensions with standard time trade-off data. In J. Kacprzyk, E. Szmidt, S. Zadrozny, K. Atanassov, & M. Krawczak (Eds.), *Advances in fuzzy logic and technology 2017* (pp. 266–277). Berlin: Springer.
- Jakubczyk, M., Craig, B., Barra, M., Groothuis-Oudshoorn, C., Hartman, J., Huynh, E., et al. (2018a). Choice defines value: A predictive modeling competition in health preference research. *Value in Health*, 21(2), 229–238.
- Jakubczyk, M., Kamiński, B., & Lewandowski, M. (2018b). Eliciting fuzzy preferences towards health states with discrete choice experiments. In C. Berger-Vachon, A. Gil Lafuente, J. Kacprzyk, Y. Kondratenko, J. Merigó, & C. Morabito (Eds.), *Complex systems: solutions and challenges in economics, management and engineering: Dedicated to Professor Jaime Gil Aluja* (pp. 131–147). Berlin: Springer. https://doi.org/10.1007/978-3-319-69989-9_9.
- Kilir, G., & Yuan, B. (1995). *Fuzzy sets and fuzzy logic: Theory and applications*. Upper Saddle River: Prentice Hall.
- Koenker, R. (2005). *Quantile regression*. Cambridge: Cambridge University Press.
- Lamers, L., McDonnell, J., Stalmeier, P., Krabbe, P., & Busschbach, J. (2006). The Dutch tariff: Results and arguments for an effective design for national EQ-5D valuation studies. *Health Economics*, 15, 1121–1132.
- Macran, S., & Kind, P. (2000). Valuing EQ-5D health states using a modified MVH protocol: Preliminary results. In: Badia X, Herdman M, Roset M (eds) *16th Plenary Meeting of the EuroQol Group, Sitges, 6–9 November 1999*. Discussion Papers (pp. 205–240).
- Norman, R., Mulhern, B., & Viney, R. (2016). The impact of different DCE-based approaches when anchoring utility scores. *Pharmacoeconomics*, 34, 805–814.
- Ramos-Goñi, J., Craig, B., Oppe, M., Ramallo-Fariña, Y., Pinto-Prades, J., Luo, N., et al. (2017). Handling data quality issues to estimate the Spanish EQ-5D-5L value set using a hybrid interval regression approach. *Value in Health*, <https://doi.org/10.1016/j.jval.2017.10.023>.
- Shiroiwa, T., Ikeda, S., Noto, S., Igarashi, A., Fukuda, T., Saito, S., et al. (2016). Comparison of value set based on DCE and/or TTO data: scoring for EQ-5D-5L health states in Japan. *Value in Health*, 19, 648–654.
- Simon, H. (1956). Rational choice and the structure of the environment. *Psychological Review*, 63, 129–138.
- Stieger, S., Reips, U. D., & Voracek, M. (2007). Forced-response in online surveys: Bias from reactance and an increase in sex-specific dropout. *Journal of the American Society for Information Science and Technology*, 58, 1653–1660.
- Tunçel, T., & Hammitt, J. (2014). A new meta-analysis on the WTP/WTa disparity. *Journal of Environmental Economics and Management*, 68, 175–187.
- Voxman, W. (1998). Some remarks on distances between fuzzy numbers. *Fuzzy Sets and Systems*, 100, 353–365.

- Xie, F., Pullenayegum, E., Gaebel, K., Bansback, N., Bryan, S., Ohinmaa, A., et al. Canadian EQ-5D-5L Valuation Study Group. (2016). A time trade-off-derived value set of the EQ-5D-5L for Canada. *Medical care*, 54, 98–105.
- Yang, Y., Rowen, D., Brazier, J., Tsuchiya, A., Young, T., & Longworth, L. (2015). An exploratory study to test the impact on three “bolt-on” items to the EQ-5D. *Value in Health*, 18, 52–60.
- Zadeh, L. (1965). Fuzzy sets. *Information and Control*, 8, 338–353.
- Zadeh, L. (1975). The concept of a linguistic variable and its application to approximate reasoning'1. *Information Sciences*, 8, 199–249.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.