Sample restrictions and the elicitation of a constant willingness to pay per quality adjusted life year

Nielsen, Jytte Seested; Gyrd-Hansen, Dorte; Kjær, Trine

Published in:
Health Economics

DOI:
10.1002/hec.4236

Publication date:
2021

Document version:
Final published version

Received: 20 March 2020 | Revised: 4 January 2021 | Accepted: 14 January 2021

DOI: 10.1002/hec.4236

RESEARCH ARTICLE

Health Economics WILEY

Sample restrictions and the elicitation of a constant willingness to pay per quality adjusted life year

Jytte Seested Nielsen[1] | **Dorte Gyrd-Hansen**[2] | **Trine Kjær**[2]

[1]Newcastle University Business School, Newcastle upon Tyne, UK

[2]DaCHE, Department of Public Health, University of Southern Denmark, Odense, Denmark

Correspondence
Jytte Seested Nielsen, Newcastle University Business School, 5 Barrack Road, NE1 4SE Newcastle upon Tyne, UK.
Email: jytte.nielsen@ncl.ac.uk

Funding information
European Community's Sixth Framework Program, Grant/Award Number: n° 044172

Abstract

It is well established that the underlying theoretical assumptions needed to obtain a constant proportional trade-off between a quality adjusted life year (QALY) and willingness to pay (WTP) are restrictive and often empirically violated. In this paper, we set out to investigate whether the proportionality conditions (in terms of scope insensitivity and severity independence) can be satisfied when data is restricted to include only respondents who pass certain consistency criteria. We hypothesize that the more we restrict the data, the better the compliance with the requirement of constant proportional trade-off between WTP and QALY. We revisit the Danish data from the European Value of a QALY survey eliciting individual WTP for a QALY (WTP-Q). Using a "chained approach" respondents were first asked to value a specified health state using the standard gamble (SG) or the time-trade-off (TTO) approach and subsequently asked their WTP for QALY gains of 0.05 and 0.1 (tailored according to the respondent's SG/TTO valuation). Analyzing the impact of the different exclusion criteria on the two proportionality conditions, we find strong evidence against a constant WTP-Q. Restricting our data to include only respondents who pass the most stringent consistency criteria does not impact on the performance of the proportionality conditions for WTP-Q.

KEYWORDS
chained approach, contingent valuation (CV), health state utility assessment, stated preference, willingness to pay (WTP), WTP per QALY

1 | INTRODUCTION

There is an ongoing debate about what value to place on a quality adjusted life year (QALY) and appropriate ways of estimating such a threshold. One approach is to establish a consumption value of a QALY based on public preferences in which individuals are asked to value hypothetical health states both in terms of QALYs and willingness to pay (WTP). It is well established that the underlying theoretical assumptions needed to obtain a constant proportional trade-off between QALY and WTP, and thus a constant value of a QALY (WTP-Q) estimate, are

restrictive (see Gyrd-Hansen, 2005 for an overview). Two conditions that need to be satisfied are *scope insensitivity* and *severity independence*. Whereas the former can be classified as a traditional test of the sensitivity of WTP to the size of the good, the latter tests the extent to which severity of the initial health states impacts on valuation. Therefore, for scope insensitivity, the WTP-Q must be independent of the size of the health gain individuals are asked to value[1], whereas severity independence implies that for equally sized QALY gains, the elicited WTP-Q must be independent of health state. Although there is evidence that WTP-Q varies across severity and scope, empirical studies, that focus systematically on the validation of the WTP-Q are scarce (exceptions include Bobinac, van Exel, Rutten, & Brouwer, 2012; Bobinac, van Exel, Rutten, & Brouwer, 2014; Pinto-Prades, Loomes, & Brey, 2009; Robinson et al., 2013; Sund & Svensson, 2018). As an example, Bobinac et al. (2014) investigated the impact of probability weighting on the theoretical validity of WTP-Q and found nonlinear probability weighting improved scope sensitivity, albeit WTP-Q still diminished with the size of the QALY gain. While the paucity in studies in this area, at least partly, can be explained by the data requirements needed to conduct such analyses, it remains unclear whether the failure to fulfill the proportionality conditions is a result of nonrandom variation in response patterns which could be mitigated.

In view of that, one area that remains greatly under-researched is the impact of applying different exclusion criteria on the theoretical validity of WTP-Q. Sund and Svensson (2018) examined the scope insensitivity condition and attempted, as a robustness check, to analyze the effect of excluding respondents who ranked health states inconsistently. Although the impact on the scope insensitivity was insignificant, their results remain inconclusive due to lack of statistical power. In the WTP-Q literature more broadly, the choice of, and reasoning behind, exclusion criteria is not well described and there is little consensus on what criteria to apply. This is also embodied in the literature review by Ryen and Svensson (2015) where the impact of exclusion criteria is not examined as a potential source of variation in WTP-Q estimates. From the included studies in Ryen and Svensson (2015), there seems to be a tendency to exclude WTP protest bids and WTP outliers whereas exclusion of responses/respondents based on consistency criteria of either the QALY or the WTP metric are rarely applied[2,3].

To the authors' best knowledge, no study has previously attempted to systematically investigate how the use of different prespecified exclusion criteria impact on the theoretical validity of WTP-Q. In this paper, we aim to investigate whether the two proportionality conditions for WTP-Q outlined above can be satisfied when data is restricted to include only respondents who pass a set of strict consistency criteria. We identify subsamples of respondents whose answers show consistency on an individual level according to (1) Positivity; WTP for a health gain must be strictly positive, (2) Internal sensitivity to scope; WTP must increase in size of the QALY gain, and (3) Internal sensitivity to health state; a strictly better health state must, all else equal, be given a higher QALY weight. Basing our sample restrictions on these consistency criteria allows us to verify whether exclusion of respondents that do not satisfy the conditions, can generate a constant WTP-Q. Further, we recognize that choice of exclusion criteria is likely to affect the magnitude of the WTP-Q estimates as well as the characteristics of the remaining sample; two elements that are relevant for the practical meaningfulness of the WTP-Q approach. We therefore also report the consequences of applying different sample restrictions on the sample representativeness and the WTP-Q estimates.

2 | METHODS AND ANALYTICAL APPROACH

For the purpose of this study, we apply a rich individual level data set comprising the Danish responses to the European Value of a QALY (EuroVaQ) study. The study design is described in much detail in Robinson et al. (2013) and further details can be found in the Final Report[4].

2.1 | Study design

Following the introductory questions, each respondent answered a total of six questions (Q1–Q6) in two "chains" as illustrated in (Figure 1).

In each chain, respondents were first asked to complete either a standard gamble (SG) or a time-trade-off (TTO) exercise, in order to ascertain the QALY value of a given health state (Q1 or Q4 depending on the random ordering of the two chains). Subsequently, respondents were randomized to one of two WTP question frames. Either they were to

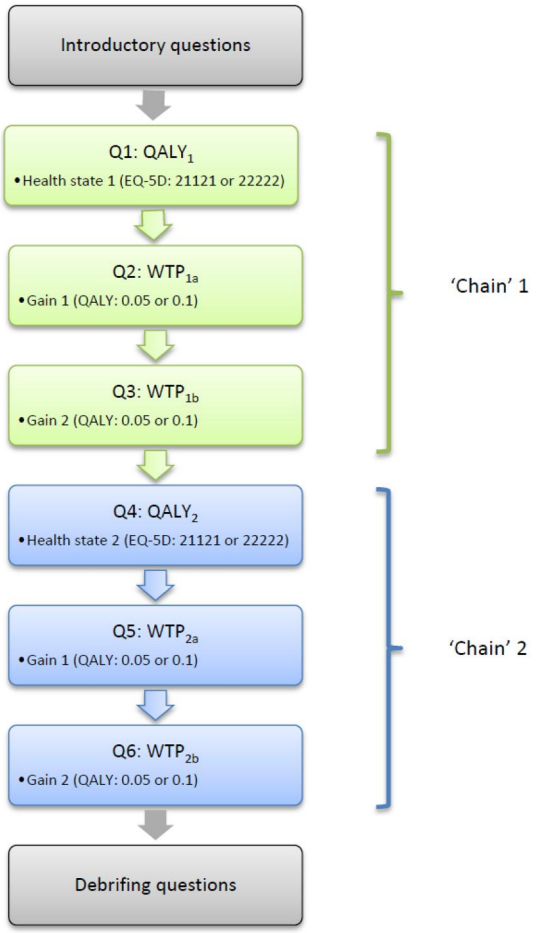

FIGURE 1 The "chained" approach

imagine that they were in the aforementioned health state and asked to state their WTP for reducing the time spent in the health state, or they were to imagine that they were at risk of falling into the health state and asked to state their WTP for reducing the risk. The magnitude of the time and risk reduction was tailored according to the responses to the previous SG/TTO questions such that all respondents were asked to state their WTP for the same QALY gains. For each respondent, WTP for a gain of both 0.05 and 0.1 was elicited in random order (Q2–Q3 and Q5–Q6). This was done using the random card sorting procedure. The set of payment cards presented to respondents aimed to keep the "range" constant in terms of implied WTP-Q[5] across the scenarios where 0.05 and 0.1 QALY was on offer.

The design thus constitutes four survey arms with four different combinations of elicitation method (SG-risk, SG-time, TTO-risk, TTO-time) which each contained two chains. The two chains varied in whether the health state description 22222 or 21121 (according to EQ-5D-3L) was used for the elicitation. These two health states were specifically chosen to ensure that one health state dominated the other. Each respondent assessed both health states, but the order in which this was done was randomized.

2.2 | Test of proportionality conditions

The novelty of the EuroVaQ study design was the chaining of the individual's assessment of the health state to the WTP such that the QALY gain valued was held constant across individuals while at the same time accounting for individual differences in the assessment of the quality-of-life of the presented health states. This allows us to test the following two proportionality conditions separately. Let $\Delta QALY_j^k$ be the health gain k from impaired health state j to perfect health and WTP_{ij}^k the WTP of individual i for that health gain.

2.2.1 | Proportionality condition 1: Scope insensitivity

According to this condition, WTP must increase in proportion to the size of the QALY gain *ceteris paribus* for WTP-Q to be constant. For two given health gains $k = 1,2$ from impaired health state j to perfect health that only differ with a factor α in size, this implies that:

$$H_0^1 : \Delta QALY_j^1 = \alpha\, \Delta QALY_j^2 \Rightarrow WTP_{ij}^1 = \alpha WTP_{ij}^2 \tag{1}$$

In our study, we operate with two QALY gains of 0.1 and 0.05 leading to a multiplicative factor of $\alpha = 2$ in Equation (1). The multiplicative factor was established by adjusting the risk or the time frame. The hypothesis is tested by multiplying WTP with 2 for the WTP responses relating to the gain of 0.05 and including a dummy variable for "scope" (0.1 = 1; 0.05 = 0). Therefore, by construct, the variable "scope" must be insignificant for H_0^1 not to be rejected.

2.2.2 | Proportionality condition 2: Severity independence

This condition states that for identical sized QALY gains, WTP must be constant and thus independent of the severity of the health states. For identical gains k obtained in two given health states $j = 1,2$ this implies that:

$$H_0^2 : \Delta QALY\,{}_1^k = \Delta QALY\,{}_2^k \Rightarrow WTP_{i1}^k = WTP_{i2}^k \tag{2}$$

We test this hypothesis by including a dummy variable labeled "severity" in our analysis indicating which of the two health states was used (22222 or 21121). By construct, "severity" must be insignificant for H_0^2 not to be rejected.

2.3 | Analytical approach

In the baseline model (Model 1), we follow standard practice in the WTP-Q literature and in line with Robinson et al., (2013) remove respondents with at least one observation classified as outlier (top-trimmed 1%) and respondents who in at least one of their responses provide a protest bid (defined as a zero WTP response, which is subsequently justified by "the government should provide the service"), and QALY nontraders[6]. Finally, to assure that all respondents can meet our consistency criteria R2 and R3 below, we also remove respondents at baseline who either a) assigned a value of above 0.9 to at least one health state ($w_{ij} > 0.9$), since for these extreme responses it was, by design, not possible to tailor the subsequent WTP question to a QALY gain of 0.1, or b) valued at least one of the health states as bad as death ($w_{ij} = 0$). In total, this implies that we remove 891 respondents in our baseline model corresponding to 3564 observations distributed as follows; outliers (91 observations), protest bids (976 observations), health state valued above 0.9 (1137 observation), health state valued as bad as death (301 observations). These exclusions entail that we lose another 1059 observations, as we only include respondents who meet all inclusion criteria, thereby ensuring that each respondent contribute four WTP-Q observations.

Table 1 provides an overview of the models and the consistency criteria we apply (R1–R3). All analyses are carried out on the data set applied in our baseline Model 1. We adopt a multistage modeling approach to examine if and how the outlined restriction criteria impact on the results of our two proportionality tests. We do this by examining one criterion at a time. Model 2 excludes all respondents who state a zero WTP in at least one of the four WTP questions (restriction R1). For Model 3, only respondents are included who state a strictly higher WTP for the larger QALY gain, and thus pass the weak internal sensitivity to scope test (restriction R2 i.e., $WTP_{0.1} > WTP_{0.05}$). In Model 4 we apply restriction R3 and only include respondents who provide consistent answers on the QALY metric by assigning a higher QALY weight to HS_{21121} compared to the inferior HS_{22222}. In Model 5 we limit our sample to the extreme and only include respondents who pass all three restriction criteria (R1–R3).

Data is analyzed using a standard log-linear model specification with clustered standard errors at the individual level[7]. We estimate the following model

TABLE 1 Overview of restrictions criteria and models

Sample restrictions		Models				
		Model 1[a]	Model 2	Model 3	Model 4	Model 5
R1:	Positivity (willingness to pay [WTP]>0)		X			X
R2:	Internal scope sensitivity ($WTP_{0.1}$>$WTP_{0.05}$)			X		X
R3:	Internal sensitivity to health state (HS_{21121}> HS_{22222}))				X	X
Number of respondents		1092	765	507	530	182
Number of observations		4368	3060	2028	2120	728

[a]The data set in Model 1 constitutes the baseline for the application of the sample restrictions (R1–R3).

TABLE 2 Description of variables

Variable name	Description	Premise[a]
Dependent variables		
ln (willingness to pay [WTP])	Elicited WTP values; four WTP per respondent; WTP is multiplied by 2 for the WTP responses relating to quality adjusted life year (QALY) gain of 0.05; WTP in Danish Krone (DKK)	N/A
Independent variables (test of proportionality)		
Severity	Severity independence severe (22222) = 1; less severe (21121) = 0	Insignificant
Scope	Scope independenceQALY gain 0.1 = 1; QALY gain 0.05 = 0	Insignificant

[a]According to assumption about constant proportional trade-off between QALY and money (i.e., constant WTP-Q).

$$\ln WTP_i = \beta_0 + \beta_1 severe + \beta_2 scope + \varepsilon_i \quad (3)$$

where$\beta_1$and β_2 are our coefficients of interests and β_0 the constant term capturing mean lnWTP of the sample. Regression variables are defined in Table 2 along with the premise as outlined in Equation (1), (2). To analyze the impact of the application of different restriction criteria on the sample's representativeness, we test whether respondent characteristics in terms of education, gender, age, household income, health state and duration of interview differ significantly from the baseline sample. Finally, to assess the impact of applying different consistency criteria on the absolute size of the WTP-Q threshold, we calculate the mean and median WTP-Q estimates for our three restricted models and test these up against the baseline results.

3 | RESULTS

Data were collected in an online survey between December 2009 and February 2010. In total, 1983 respondents answered the Danish version of the questionnaire. According to the set of exclusion criteria at baseline, we remove a total of 891 respondents leading to a final baseline sample of 1092 respondents (equivalent to 4368 WTP-Q estimates and 2184 QALY weights). Table 3 lists characteristics of those respondents included in each of the five models. *We find that the respondents in the baseline model (Model 1) are not significantly different when compared to the full sample (except that they spend slightly longer time on the questionnaire).*[8,9] *However, we see that the respondents in Model 4 and Model 5 (the most restrictive models) when compared to the baseline sample are significantly younger, higher educated, healthier and spend longer time on the questionnaire*. The same general pattern is found for the other models, however in most cases the differences are not statistically significant.

In addition, we find that only 182 (17%) of respondents meet all three consistency criteria whereas 405 respondents (37%) comply with two criteria and 446 (41%) respondents comply with only one criterion. This also implies that 59 respondents (5%) in our baseline sample fail to meet just one criterion.

TABLE 3 Summary statistics (standard error in parentheses)

	Full sample	Model 1 (Baseline)	Model 2 (R1)	Model 3 (R2)	Model 4 (R3)	Model 5 (R1-R3)
Age in years	47.9 (0.36)	47.5 (0.50)	48.2 (0.60)	45 (0.73)***	46.0 (0.73)*	43.4 (1.29)***
Male	0.49 (0.01)	0.47 (0.02)	0.46 (0.02)	0.50 (0.02)	0.45 (0.02)	0.48 (0.04)
Education						
Middle level education	0.17 (0.01)	0.17 (0.01)	0.16 (0.01)	0.16 (0.02)	0.15 (0.02)	0.13 (0.03)
Secondary level education	0.38 (0.01)	0.39 (0.01)	0.38 (0.02)	0.37 (0.02)	0.36 (0.02)	0.34 (0.04)
Tertiary level education	0.45 (0.01)	0.44 (0.02)	0.47 (0.02)	0.47 (0.02)	0.5 (0.02)**	0.53 (0.04)**
Household income in 1000 Euro	62.4 (1.8)	62.3 (2.4)	63.5 (2.9)	62.1 (3.6)	65.5 (3.3)	70.3 (6.4)
Health status[a]	0.86 (0.04)	0.87 (0.05)	0.88 (0.06)	0.87 (0.008)	0.89 (0.007)*	0.90 (0.01)***
Survey completion time in minutes	25.9 (0.23)	26.7 (0.33)**	28.2 (0.39)***	25.9 (0.46)	28.1 (0.47)**	29.0 (0.73)***
Restriction criteria (R1–R3) satisfied (no of respondents)						
Three criteria		182 (17%)				
Two criteria		405 (37%)				
One criterion		446 (41%)				
No criteria		59 (5%)				
No. of respondents	1983	1092	765	507	530	182

Note: Tested using *t*-test. M1 is compared to the full sample. M2–M5 are compared with M1

[a]EQ-5D Danish tariffs.

*,**,***Significant at 0.1, 0.05 and 0.01 levels, respectively.

TABLE 4 Regression results (using a log-linear specification with clustered standard errors at the individual level)

	Model 1 (Baseline) Coef.(std. error)	Model 2 (R1) Coef.(std. error)	Model 3 (R2) Coef.(std. error).	Model 4 (R3) Coef.(std. error).	Model 5 (R1-R3) Coef.(std. error).
Scope	0.14 (0.06)***	−0.36 (0.02)***	1.07 (0.1)***	0.32 (0.08)***	0.15 (0.04)***
Severity	0.28 (0.08)***	0.15 (0.03)***	0.3 (0.13)**	0.24 (0.12)*	0.12 (0.07)*
Constant	7.21 (0.12)***	9.31 (0.06)***	5.36 (0.2)***	7.37 (0.16)***	8.92 (0.11)***
Respondents	1092	765	507	530	182
Observations	4368	3060	2028	2120	728

*,**,***Significant at 0.1, 0.05 and 0.01 levels, respectively.

Regression results for all our models including test results of the proportionality conditions are shown in Table 4.

For all models (M1–M5), and thus irrespective of restriction criteria, we reject the two hypotheses H_0^1 and H_0^2 at the 0.1 significance level (with all p-values below 0.05 except in two cases where $p = 0.05$ and $p = 0.07$). The violations of the proportionality conditions were found even in the most restrictive model (Model 5) for which all the consistency criteria were applied simultaneously. The "scope" coefficient was found to be positive (except in Model 2) and significant implying that WTP for a QALY (WTP-Q) gain of 0.1 was more than twice the WTP for the half-sized (0.05) QALY gain[10]. Likewise, the "severity" coefficient was found to be positive and significant implying that for two equally sized health gains, a QALY gain is valued higher in the more inferior health state (22222) than in the less severe health state (21121).

As a secondary analysis we examine the consequences of applying the restriction criteria on mean and median WTP-Q. Results are presented in Table 5 with differences in WTP-Q tested up against our baseline estimates[11]. We see that

TABLE 5 Willingness to pay for a quality adjusted life year (WTP-Q) estimates

WTP-Q	Model 1 (Baseline)	Model 2 (R1)	Model 3 (R2)	Model 4 (R3)	Model 5 (R1-R3)
Mean (EUR[1])	35,598	45,013***	24,621***	34,094	32,842
% Change from baseline		26%	−31%	−4 %	−8%
Median (EUR[1])	8919	13,514***	3514****	8919	10,811***
% Change from baseline		52%	−61%	0	21%

Notes: 1 = Tested using *t*-test and Mann–Whitney test. All values in 2010 Euro

*** denotes significantly different from Model 1 at $p < 0.01$.

the use of consistency criteria impacts WTP-Q estimates significantly. The mean WTP-Q vary from EUR 24,621 (Model 3) to as high as EUR 45,013 (Model 2) both of which are significantly different from the WTP-Q of EUR 35,598 in our baseline sample. The biggest relative impact on the WTP-Q comes from excluding respondents who are insensitive to scope (R2), which reduces the mean (median) WTP-Q by 31% (61%) compared to our baseline estimate. Irrespective of sample restrictions applied, we see that the mean estimates are at least three times as high as the median estimates but the variability in the median estimates follow the same patterns as the mean estimates.

4 | DISCUSSION

We find that restricting our data to include only respondents who pass three consistency criteria limits the baseline sample drastically by 83%, but does not impact on the performance of the proportionality conditions for WTP-Q. Across the models, we find that our stated preference data do not meet the proportionality conditions required to obtain a constant WTP-Q. Importantly, these results are not driven by potential lack of statistical power as we obtain significant results for our two theoretical predictions that by construct should have been insignificant for our data to comply with the proportionality conditions. Specifically, we find evidence that WTP-Q increases in QALY gain (H_0^1) and severity (H_0^2). These results hold across all models, and even for the model applying the three consistency criteria simultaneously (Model 5). This indicates that the factors influencing WTP-Q are pertinent across respondents, also among the (few) "rational" respondents who pass the most stringent consistency criteria. Our findings suggest that the violations of the proportionality conditions that we (and others) observe, can be considered a widespread behavioral trait and not driven by inconsistencies in response patterns.

Whereas restricting our analyses to respondents who display internal scope sensitivity (R2) and internal sensitivity to heath state (R3) is well-accepted in the literature, it is more controversial to exclude all respondents with zero WTP bids (R1). In this specific study, where respondents were asked to state their WTP-Q gain involving a "*simple*, *safe and*, *painless treatment*," we would argue that zero bids could be considered invalid responses based on the premise that a rational individual should, all else equal, derive positive value for a health gain with zero opportunity costs. Moreover, this positivity criterion (WTP for a health gain must be strictly positive) is in line with restriction criteria recently applied by Hammitt, Geng, Guo, and Nielsen (2019) and accords with the decision rule in economic evaluations.

It is worth noting that whereas we from start exclude a large group of respondents in our baseline model (Model 1), our sample does not change in terms of observed characteristics. Hence, the respondents included at baseline seem to represent the total sample well. In contrast, respondents included in the most restricted subsample are significantly different from those of our baseline sample, indicating that the application of consistency criteria change the composition of the sample so that younger, higher educated, healthier and people spending longer time on the questionnaire are over-represented. Not surprising, this has important policy implications in terms of eliciting WTP-Q that accurately represents the preferences of the general public. We cannot rule out that there are other explanations for the rejection of the proportionality conditions that we observe across models. It is likely that respondents anchor on the first health state and/or size of the QALY gain they see in the survey and adjust their valuation insufficiently in subsequent questions for the proportionality conditions to be fulfilled (Gyrd-Hansen, Kjær, and Nielsen, 2012). Moreover, it could be argued that respondents do not pay sufficient attention to the small differences in size of the health gains and/or find it too difficult to distinguish between small QALY gains. Similar observations have been done in the literature on valuation of risk reductions where there is ample evidence that participants are insensitive to very small changes in risk

(Robinson et al., 2013). In order to elicit WTP-Q values which are not likely to be impacted by budget constraint it is imperative that valuations are elicited from marginal changes in QALY gains. We therefore elicit WTP for changes in health states that, based on either duration or risk adjustments, constitute smaller QALY gains (0.05 or 0.1 QALY). We thereby operate with small health gains but avoid the infinitesimally small health gains that may be prevalent in other WTP studies. Our lack of support for proportionality condition 1 (scope insensitivity) is therefore not driven by lack of sensitivity to such minutely small changes in health. On a related note, the set of payment cards was adjusted to keep the "range" of implied WTP-Q constant for the two QALY gains on offer[12]. Consequently, our rejection of insensitivity to scope (proportionality condition 1) is not likely to be driven by the elicitation approach.

Although we find that WTP-Q differ significantly across samples, we still obtain WTP-Q estimates within the bounds previously found in Ryen and Svensson (2015)[13]. This said, to allow for an easier comparison of WTP-Q estimates across studies, we would advocate for more transparency in the reporting of WTP-Q estimates including reflections on the choices of exclusion and inclusion criteria and how these are likely to affect the results.

5 | CONCLUSION

A constant WTP-Q requires a constant proportional trade-off between the two metrics: WTP and QALY. Some empirical studies have tested and found evidence of violation of the proportionality conditions. Using a rich data set, this paper investigates whether these violations can be attributed to respondents who fail to pass consistency criteria. We find strong evidence against a constant WTP-Q irrespective of the restriction criteria applied. This indicates that the factors influencing WTP-Q are pertinent across respondents and also present among the (few) respondents who pass all three consistency criteria. Our results suggest that eliciting a constant WTP per QALY estimate may be unfeasible, as we cannot improve the validity of our estimate through exclusion of inconsistent respondents.

ACKNOWLEDGEMENT

We thank the other members of the EuroVaQ team who contributed to the work presented here: Cam Donaldson, Rachel Baker, Helen Mason, Mark Pennington, Sue Bell, Michael Jones-Lee, John Wildman, Emily Lancsar, Angela Robinson, Philomena Bacon, Jan Abel Olsen, Ulf Persson, Annika Bergman, Christel Protière, Jean Paul Moatti, Stephane Luchini, Jose Luis Pinto Prades, Awad Mataria, Rana Khatib, Yara Jaralla, Werner Brouwer, Job van Exel, Roman Topór-Madry, Adam Kozierkiewicz, Darek Poznanski, Ewa Kocot László Gulácsi, Márta Péntek, Samer Kharroubi, Andrea Manca, and Phil Shackley. The research leading to these results received funding from the European Community's Sixth Framework Program under grant agreement n° 044172—the EuroVaQ project.

CONFLICT OF INTEREST

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

ETHICS STATEMENT

No ethical approval was required .

DATA AVAILABILITY STATEMENT

Data are available from the corresponding author with the permission of the EuroVaQ Management Board.

ORCID

Jytte Seested Nielsen https://orcid.org/0000-0003-0129-0225
Dorte Gyrd-Hansen https://orcid.org/0000-0003-1137-2304
Trine Kjær https://orcid.org/0000-0002-9554-374X

ENDNOTES

[1] This implies that WTP must exhibit strong sensitivity to the size of the health gain.

[2] Whereas all WTP-Q studies will have to collect individual level WTP data, QALY weights can either be elicited directly using for example, time-trade-off (TTO) or indirectly using official tariffs such as EQ-5D. Moreover, the number of valuation task per respondent differ across

studies. We acknowledge that this has implications for the applicability of exclusion criteria including the feasibility to restrict the sample according to certain consistency criteria (requiring with-in subject data).

[3] See Table S1 in the Appendix for a brief overview of exclusion criteria applied in the contingent valuation literature estimating WTP-Q.

[4] See the EuroVaQ website at http://research.ncl.ac.uk/eurovaq/

[5] For the QALY gain of 0.05, the payment amounts were ranging from DKK 65-200,000 whereas for the QALY gain of 0.1, the payment amounts were ranging from DKK 150-400,000. The Danish krone is pegged to the Euro, 1 Euro = DKK 7.46.

[6] As we operate with a balanced sample, we delete observations at the respondent level and not the response level as done by Robinson et al. (2013). This also implies that more observations are deleted.

[7] A constant value of 1 has been added to 0 WTP-values before the log transformation.

[8] That the completion time in Model 2 is significantly higher compared to Model 1 is likely explained by the removal of zeros as it takes much less time to state a WTP = 0 compared to completing the card sorting tool.

[9] The full sample (and by implication Model 1) was representative of the Danish population on average age and gender. Compared to the Danish population, the full sample was better educated and had a lower household income.

[10] This result is surprising given issues with scope insensitivity previously found in the literature and should be interpreted with caution. In Model 2 (where zero values are excluded), the "scope" coefficient was found to be negative which suggests that the positive coefficient to a large extent is driven by zero values being more prevalent for the 0.05 gain. In Table S2 in the Appendix, we report results using an interval regression approach. Reassuringly, this does not change our findings that all the coefficients are significant, implying that our results are robust to a change in model specification. However, the scope coefficient becomes negative in the interval regressions which indicate that a change in functional form is crucial to the sign of the coefficient. This is likely a result of right skewed data which is commonly found in WTP studies.

[11] For comparison, the mean(median) WTP-Q in the full sample is EUR 38,890 (9054) which is significantly different from the baseline model ($p < 0.05$).

[12] This implied that the bids presented in the card sort for the 0.05 QALY gain were halved relative to the bids in the scenario with 0.1 QALY gain.

[13] Ryen and Svensson (2015) find in their review trimmed mean and median estimates amount to 74,159 and 24,226 Euros (2010 price level), respectively. Whereas the estimates found in this study 24,621–45,013 and 3514- 13,514 Euros (2010 price level), respectively are below their finding, they are still within the range of previous studies. However, Ryen and Svensson also note that 80% of the mean estimates in their review are below 75,000 Euros

REFERENCES

Bobinac, A., van Exel, J., Rutten, F., & Brouwer, W. (2012). Get more, pay more? An elaborate test of construct validity of willingness to pay per QALY estimates obtained through contingent valuation. *Journal of Health Economics*, *31*, 158–168.

Bobinac, A., van Exel, J., Rutten, F., & Brouwer, W. (2014). The value of a QALY: Individual willingness to pay for health gains under risk. *PharmacoEconomics*, *32*, 75–86.

Gyrd-Hansen, D. (2005). Willingness to pay for a QALY. *PharmacoEconomics*, *23*(5), 423–432.

Gyrd-Hansen, D., Kjær, T., & Nielsen, J. S. (2012). Scope insensitivity in contingent valuation studies of health care services: Should we ask twice? *Health Economics*, *21*(2), 101–112.

Hammitt, J., Geng, F., Guo, X., & Nielsen, C. P. (2019). Valuing mortality risk in China: Comparing stated- preference estimates from 2005 and 2016. *Journal of Risk and Uncertainty*, *58*, 167–186.

Pinto-Prades, J. L., Loomes, G., & Brey, R. (2009). Trying to estimate a monetary value for the QALY. *Journal of Health Economics*, *28*, 553–562.

Robinson, A., Gyrd-Hansen, D., Bacon, P., Baker, R., Pennington, M., & Donaldson, C., & EuroVaQ-team (2013). Estimating a WTP-based value of a QALY: The `chained` approach. *Social Science & Medicine*, *92*, 92–104.

Ryen, L., & Svensson, M. (2015). The willingness to pay for a quality adjusted life year: A review of the empirical literature. *Health Economics*, *24*, 1289–1301.

Sund, B., & Svensson, M. (2018). Estimating a constant WTP for a QALY-a mission impossible?. *The European Journal of Health Economics*, *19*(6), 871–880.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Nielsen JS, Gyrd-Hansen D, Kjær T. Sample restrictions and the elicitation of a constant willingness to pay per quality adjusted life year. *Health Economics*. 2021;1–9. https://doi.org/10.1002/hec.4236