Improving the Design of EQ-5D Value Set

Studies for China and Beyond

Zhihao Yang

ISBN 978-94-6375-074-5

Printing and layout: Ridderprint BV, www.ridderprint.nl

Improving the Design of EQ-5D Value Set Studies for China and Beyond

Verbeteringen van het ontwerp van EQ-5D
waarderingsonderzoek voor China en daarbuiten

Thesis

to obtain the degree of Doctor from the
Erasmus University Rotterdam
by command of the
rector magnificus
Prof. dr. R.C.M.E. Engels

and in accordance with the decision of the Doctorate Board.
The public defence shall be held on
Wednesday 26 September 2018 at 9.30 hours

by

Zhihao Yang
born in Guiyang, China

Erasmus University Rotterdam

Promotiecommissie

Promotoren	Prof. dr. J.J. van Busschbach
Overige leden	Prof. dr. J. Passchier Dr. E. W. de Bekker Dr. M. M. Versteegh
Copromotoren	Dr. E.A. Stolk Dr. N. Luo

Table of Contents

CHAPTER 1

General introduction

This PhD thesis reports on recent experiences concerning the use of EQ-5D in China. The EQ-5D instrument is the most widely-used quality of life questionnaire in health economic evaluation world-wide and has recently been introduced in China. EQ-5D invites respondents to report their level of functioning on five basic dimensions of health. The responses define that person's EQ-5D health state. All EQ-5D health states can have a value attached to them which indicates how good or bad the quality of life is of people living in that state. In this way, EQ-5D describes and values health. The values attached to all EQ-5D states represent a key feature of EQ-5D, as they enable comparisons of health across population subgroups (e.g. stratified by region, disease area, or treatment received), and such indicators of health can inform health care investment decisions.

A basic requirement that follows from the context in which EQ-5D values are used is that values must reflect the health preferences of the target population. Hence, it is a recommended approach to establish values in the local context of the EQ-5D user. Health valuation is the field of science involved with constructing such value sets. The increased use of EQ-5D in China has spurred the development of the field of health valuation in that country. Initial research aimed to develop a local value set for EQ-5D in China, but this research has expanded and also addressed methodological questions around optimal ways to establish values in a valid and cost-effective way. These developments form the background for the studies reported in this thesis.

1.1 HTA AND ECONOMIC EVALUATION IN CHINA

The introduction of EQ-5D in China reflects an increased interest in economic evaluation and Health Technology Assessment (HTA). EQ-5D is a preferred outcome measure in economic evaluation, which can be seen as the most important component of HTA. HTA is a multidisciplinary field of policy analysis, studying medical, economic, social and ethical effects of the development, diffusion, and use of health technologies (1). HTA research supports decisions about the inclusion of health technologies in the collectively financed health benefit package. China started its HTA programme in the 1980s, encouraged by the World Bank. As in many other countries, when HTA was introduced in China, it was characterised as a body of academic activities. Since then it has become increasingly accepted and there is a growing use of HTA in listing, pricing, and reimbursement of pharmaceuticals (1).

In the last two decades, health expenditure grew at a rate of 11.6% per year, which outpaced the economic growth rate of 9.9% in China (2). Economic evaluation provides a tool for

containing increasing health care costs by promoting more efficient allocation of health resources (3). Since resources are limited, the funding decisions need to be made between different treatments/drugs/interventions etc. By comparing different alternatives' costs and effects (health outcomes), decisions can then be made against either some threshold values or by considering other possible concerns, e.g. disease burdens.

Cost-utility analysis (CUA) is one type of economic evaluation. This method uses Quality Adjusted Life Years (QALY) to measure health effects. The QALY measures health outcomes by combining length of life with health-related quality of life (HRQoL) (3). HRQoL is a broad concept capturing quality of life, utility or wellbeing that is related to health. By combining mortality and morbidity information, the QALY is a preferred measure in cost-effectiveness studies around the world as it allows comparison between different diseases and treatments. EQ-5D is the most widely-used questionnaire in the world to determine the 'adjustment factor', i.e. the quality part of the QALY.

In addition to its use in health economic evaluation, as explained above, EQ-5D can also be used to measure and describe health. For instance, when used to measure the health status of a given patient group, the disease burden of this group can be estimated by comparing its health with that of the healthy population. The health of the healthy population measured by EQ-5D is called a population norm, which provides normative values for the general public and has served as a benchmark in quantifying disease burdens.

A well-established EQ-5D value set is a necessary asset if China aims to expand its HTA activities. The intention of this thesis is to contribute in this endeavour by improving the validity and cost-effectiveness of the methods used for EQ-5D valuation research. An additional aim is to provide an EQ-5D population norm for the urban Chinese population.

1.2 EQ-5D AND ITS USE IN MEASURING HEALTH

The EQ-5D system includes two essential parts: page 2 displays the EQ-5D descriptive system and page 3 contains the EQ visual analogue scale (EQ-VAS). Figure 1 shows an example of EQ-5D-5L descriptive system in English. The EQ-5D descriptive system comprises five dimensions: mobility, self-care, usual activities, pain/discomfort and anxiety/depression (4). It has two versions, the three-level EQ-5D (EQ-5D-3L) and the five-level EQ-5D (EQ-5D-5L). In total, EQ-5D-3L defines a total of 243 unique health states, while EQ-5D-5L defines 3,125 health states. The higher number of health states in the 5L version is aimed at ensuring improved sensitivity.

Figure 1: EQ-5D-5L descriptive system

Under each heading, please tick the ONE box that best describes your health TODAY.

MOBILITY

I have no problems in walking about ❑
I have slight problems in walking about ❑
I have moderate problems in walking about ❑
I have severe problems in walking about ❑
I am unable to walk about ❑

SELF-CARE

I have no problems washing or dressing myself ❑
I have slight problems washing or dressing myself ❑
I have moderate problems washing or dressing myself ❑
I have severe problems washing or dressing myself ❑
I am unable to wash or dress myself ❑

USUAL ACTIVITIES *(e.g. work, study, housework, family or leisure activities)*

I have no problems doing my usual activities ❑
I have slight problems doing my usual activities ❑
I have moderate problems doing my usual activities ❑
I have severe problems doing my usual activities ❑
I am unable to do my usual activities ❑

PAIN / DISCOMFORT

I have no pain or discomfort ❑
I have slight pain or discomfort ❑
I have moderate pain or discomfort ❑
I have severe pain or discomfort ❑
I have extreme pain or discomfort ❑

ANXIETY / DEPRESSION

I am not anxious or depressed ❑
I am slightly anxious or depressed ❑
I am moderately anxious or depressed ❑
I am severely anxious or depressed ❑
I am extremely anxious or depressed ❑

By reporting one's health through ticking the corresponding response level from each of these five dimensions, EQ-5D health states can be simply described using five-digit codes, for example, 13245 represents a health state with no problems in walking about, moderate problems in self-care, slight problems in usual activities, having severe pain/discomfort and being extremely anxious/depressed. It is a credit to its descriptive richness and simple-to-use nature that EQ-5D has been used to measure population health in many countries, and population norms have been established by age, gender and socio-economic status (4). A set of norm scores provides an important reference point for clinical and health economic research outcomes, as the effects of medical conditions and/or treatments can be quantified by comparing patients and/or intervention groups with the general population (5). Currently, there are no EQ-5D-5L norms for the Chinese population, which hampers expansion in the use of EQ-5D-5L in China. In this thesis, we aim to provide such norms and to evaluate how health varies between demographic groups.

The EQ-5D descriptive system provides a way to classify and measure health, but direct comparison between two health states is difficult as a health state which is good in certain dimensions may not be good in others, for example, 13245 versus 51153. The former state is good in mobility, 'severe' in pain/discomfort, whilst the latter has extreme problems in mobility and pain/discomfort but no problems in self-care and usual activities. This comparison between states can be facilitated using the attached unidimensional value for each state, which reflects the desirability of that state. For instance, if the state 13245 has a value of 0.53 and state 51153 has a value of 0.31, then it could be concluded that state 13245 is better than state 51153 as 0.53>0.31. The next section describes how to obtain such values.

1.3 VALUATION OF EQ-5D

As mentioned above, the use of EQ-5D in economic evaluation requires its corresponding value set, which provides the index values (health utilities) of all defined health states. Such value sets are usually derived using a two-step approach: first, a subset of health states of the EQ-5D instrument is directly valued by members of the general public, and second, the observed values are modelled to predict values for all health states. This two-step approach is preferred over the direct valuation of all EQ-5D states, because the latter strategy requires a huge sample size and thus becomes extremely time-consuming and costly, which is deemed not feasible. The reason is that many health states need to be valued (243 for the 3L version and 3,125 for the 5L version of EQ-5D) by many respondents (N = 30 to 100) in order to obtain reliable mean values, but the maximum

number of health states respondents can value is usually around 20 to 30.

A challenge in conducting valuation research using the above-mentioned two-step approach is that a crucial aspect of designing such studies is not fully understood. In this thesis, 'design' typically refers to a specific question: the subset chosen for direct valuation. In the literature on health state valuation there has been much debate concerning how to select the subset of health states for which empirical values are collected. Different desirable properties for such subset have been identified (e.g. plausibility of the states, severity balance across the design), but these desirable properties cannot all be satisfied at the same time because of a resources constraint. In the absence of straightforward statistical rules to trade off the desirable properties, the selection of health states for the sample has thus far been consensus-based at the research team level. This approach caused studies to differ, without justification. In other words: when estimating a value set, researchers take a leap of faith because the trade-offs between available designs are unclear. This thesis attempts to shed light on these trade-offs, mostly in the context of EQ-5D work undertaken in China.

The dissemination of EQ-5D in China has been facilitated by the ground-breaking research of Dr. Nan Luo from National University of Singapore and Prof. Gordon Liu from Peking University. They pioneered the establishment of an EQ-5D value set for a large country such as China, using limited resources. In their research they were confronted with two difficulties, given these limited resources. First, they did not have the opportunity to investigate design properties beforehand, which resulted in the concerns mentioned above. Second, they did not have the resources to engage respondents from rural areas. It is likely that people living in such areas have different preferences compared to people in urban areas, since health preferences are known to be affected by demographic and cultural factors (6, 7). Since HTA decisions affect all residents equally, democratic principles suggest all people should have a chance to express their preferences. This is also true for China, where the distinction between rural and urban populations reflects a variety of social and economic inequalities. It is desirable to avoid value hegemony of the advantaged groups and deepening the divide. Thus, it is relevant to know that the values respondents give to EQ-5D health states relate to their experiences with ill health, personal interests and circumstances, and the environment etc.(8).

Both difficulties are linked in the sense that an efficient design can free up resources to engage the more difficult-to-reach respondents from rural areas. The way to establish a more efficient design, and hence better opportunities to arrive at such representative samples, is the theme of the thesis.

1.4 AIMS AND OUTLINE OF THIS THESIS

Six studies were conducted aimed at improving the use, and especially the valuation, of EQ-5D in China. First in Part 1 (Chapter 2 and Chapter 3), there is an exploration of the 2012 Chinese valuation data to see how demographic factors affect individual's self-reported health states and understanding of the TTO task. Then part 2 (Chapter 4 to Chapter 7) reports on how different design choices affect the predictions of health state values for both EQ-5D-3L and EQ-5D-5L.

Part 1:
The first part of the thesis is focussed on how demographic factors affect individuals' self-reported HRQoL and understanding of the TTO valuation task. EQ-5D-5L data from China's 2012 valuation study was utilized, the same data that was used to establish the EQ-5D-5L value set for urban China (9). With this data, the thesis aims to answer the following research questions:

Research question 1 – What are the EQ-5D-5L norm scores for the urban Chinese population and are there disparities in self-reported HRQoL in urban China?
A set of norm scores provides an important reference point for clinical and health economic research outcomes, as the effects of medical conditions and/or treatments can be quantified by comparing patients and/or intervention groups with the general population (5). Moreover, previous research has shown HRQoL inequalities between different socio-demographic groups and regions in China (10-13). Reporting the norm scores by demographic groups helps us to understand this issue further: this is accomplished in **Chapter 2** for the urban Chinese population which also shows how demographic factors affect individuals' self-reported HRQoL.

Research question 2 – Is the TTO valuation method equally valid across respondents/ interviewers in China?
We know from previous studies that the TTO task is difficult for some respondents (14). This is more problematic if certain groups of respondents (e.g. those with a low level of education) are excluded due to data quality reasons, as the representativeness of the sample would be compromised. Similarly, an interviewer could prove problematic if his/ her respondents consistently showed higher levels of inconsistency. Hence, in **Chapter 3**, the validity of the composite time trade-off method in the Chinese population is assessed by looking at individual-level inconsistencies.

Part 2:

In this part, I focus is upon an important design issue for valuation studies: how to select health states for direct valuation? As discussed above, a modelling approach is used to obtain the values of all defined health states in EQ-5D. In this approach, first a subsample of health states is selected for direct valuation, then the values of other health states are predicted from these empirical values. For EQ-5D-3L valuation studies, different designs were used in different countries (15). For EQ-5D-5L, a standardized EQ-VT protocol was established and, using the same design, value sets were established for different countries (9, 16-22). An open question is how the different design choices for EQ-5D-3L valuation studies and the standardized EQ-VT design of EQ-5D-5L valuation studies performed in predicting the values of all health states.

Research question 3 – How to select health states for EQ-5D-3L valuation studies?

Published EQ-5D-3L valuation studies have utilized from 17 to 43 states for direct valuation and the performance of these designs is unknown. Additional to the published designs, an examination of two oft-used, but competing criteria, in selecting health states is proposed. The first criterion is commonness of the states. In relying on the general public to value health states, these health states should be imaginable to the respondents, otherwise reliable values may not be obtained. The second criterion is that the selected states, taken together, should possess balanced statistical properties, allowing unbiased decomposition of health effects. In **Chapter 4**, the validity of the published designs versus newly proposed designs in selecting health states for direct valuation in EQ-5D-3L is assessed, using an external saturated VAS dataset as validation.

Research question 4 –What are implausible EQ-5D health states and how do members of the general public value implausible EQ-5D-5L health states?

As many members of the general public do not have much ill-health experience, some EQ-5D health states are inevitably hypothetical for them (23). One issue in valuing hypothetical health states is that some states may be considered implausible or unrealistic by respondents. Perceived implausibility may prevent respondents from accurately imagining the concerned health states, which is pivotal to the thought process for valuation. In **Chapter 5**, the characteristics of implausible health states are identified and there is an examination concerning how values differed over plausible and implausible observations.

Research question 5 – can we use a smaller design to estimate EQ-5D-5L value sets?

Arguably, the fewer health states used for direct valuation, the more feasible a valuation study would be. Nonetheless, the selected health states for direct valuation should enable

adequate predictions for the non-valued health states. Previous EQ-5D-3L research has shown that, by optimising the statistical efficiency of a design, less states can be used for direct valuation without compromising prediction accuracy. Hence, it would be helpful to know how many health states are needed for an acceptable level of prediction accuracy and how to select these health states in EQ-5D-5L. In **Chapter 6**, applying a similar method used in examining design performance in EQ-5D-3L, the current EQ-VT design and a possibly more efficient design in terms of prediction accuracy is evaluated. To achieve this, an EQ-5D-5L VAS saturated dataset is collected. In **Chapter 7**, as TTO is the main method of collecting valuation data for EQ-5D, a 25-state orthogonal design is applied to TTO data to assess whether findings from Chapter 6 using VAS data can be generalized to TTO data.

Finally, in **Chapter 8**, the findings of the thesis are discussed alongside some other relevant matters.

1.5 TERMINOLOGY

As health economics is a multi-disciplinary field, the terms used are not always standardized. Terms used throughout the thesis are employed as consistently as possible. EQ-5D 'index value' is sometimes referred to as 'health state utility'. The 'Misery Index' is sometimes referred to as 'the sum of five digits'. Some terms look similar but are fundamentally different, notably, 'implausible health states' are different from 'uncommon health states'. The commonness of a health state is defined as the prevalence of that state in reality, whereas the plausibility of a health state is defined as whether a respondent considers it as likely to exist and therefore imaginable. Another example is the word 'design'. In this thesis, 'design' typically refers to a specific question: the subset chosen for direct valuation.

CHAPTER 2

EQ-5D-5L norms for the urban Chinese population in China

Zhihao Yang, Jan Busschbach, Gordon G Liu, Nan Luo

Submitted for Publication

2.1 INTRODUCTION

EQ-5D is a health-related quality of life (HRQoL) questionnaire widely used in economic, clinical, and population health studies. The EQ-5D descriptive system comprises five dimensions: mobility, self-care, usual activities, pain/discomfort and anxiety/depression (4). It has two versions, a three-level EQ-5D (EQ-5D-3L) and a five-level EQ-5D (EQ-5D-5L). Although EQ-5D-3L has been widely used, it is reported to suffer from ceiling effects and measurement insensitivity (24). By increasing the number of levels in the descriptive system, EQ-5D-5L has demonstrated improved measurement properties in ceiling effects, and in discriminatory power in comparison to EQ-5D-3L (9, 25-27). In addition to classifying health states in terms of the 5 dimensions of health, EQ-5D permits the valuation of these health states. This is accomplished from both the respondent's own perspective by using a Visual Analogue Scale (EQ-VAS) and from the perspective of the general public's by attaching the appropriate EQ-5D index score to the described health state of the respondent.

EQ-5D has been used to measure population health in many countries, and population norms have been established by age, gender and socio-economic status (4). A normative data set provides an important reference point for clinical and health economic research outcomes, as the effects of medical conditions and/or treatments can be quantified by comparing patients and/or intervention groups with the general population (5). At this juncture, there are no EQ-5D-5L norms for the Chinese population, which hampers the increasing use of EQ-5D-5L in China.

The objective of this paper is to provide normative data, including the prevalence of EQ-5D-5L health problems, and EQ-VAS and EQ index scores by age and gender, in the Chinese urban population. In addition, we also examine the relationships between socio-economic factors and (i) the components of the EQ-5D-5L descriptive system, (ii) EQ-VAS scores and (iii) EQ index scores.

2.2 METHODS

2.2.1 Sampling and recruitment

The study drew data from a large EQ-5D-5L valuation study in China (9). The aim of this study was to estimate a country specific scoring algorithm to calculate EQ-5D-5L index scores. The scoring algorithm has been reported elsewhere (9). The sample size was decided by the EQ-5D-5L valuation protocol, which was aiming at constructing country

specific EQ-5D-5L value set (28). Members of the general population were randomly recruited from five urban areas of five cities (Beijing, Shenyang, Nanjing, Chengdu, and Guiyang). From each city, respondents were recruited from at least five difference administrative districts and at different time of day. Specific recruitment sites included library, hospital, university, local community, park and shopping areas etc (6). These five cities were selected as representative urban areas in terms of size of population, geographical region and economic development status in China (9). Within each city, quotas were set to recruit equal numbers of participants from each city and to ensure the study sample resembled the general Chinese urban adult population with respect to age, gender, and education level according to the Sixth National Population Census (6). In each city, members of the general public who were at least 16 years old, and were literate and able to understand survey questions, were recruited through personal invitation (9). Response rate was calculated.

Each respondent was interviewed face-to-face by a trained interviewer using the EuroQol valuation technology (EQ-VT) (9, 29). EQ-VT is a standardized software design by the EuroQol Group in order to facilitate the data collection for valuation study (28). The interview had four sections. The first section was for respondents to report their own health using the EQ-5D-5L questionnaire: the five-dimensional descriptive system and the EQ-VAS. In the second section respondents were asked to value 10 different EQ-5D-5L health states using a composite time trade-off (cTTO) method (8). The third section contained 7 pairs of EQ-5D-5L discrete choice tasks. The fourth section assessed respondents' socio-economic and other background characteristics. This paper used data collected in the first and fourth sections only.

2.2.2 The EQ-5D questionnaire

The EQ-5D-5L descriptive system consists of five dimensions (mobility, self-care, usual activities, pain/ discomfort and anxiety/depression) with five ordinal severity levels each (no problems, slight problems, moderate problems, severe problems, and extreme problems/unable to), thus defining 3,125 (5^5) distinct health states (24). The respondent is asked to indicate his/her health state against the most appropriate statement in each of the 5 dimensions and this leads to a 1-digit number expressing the level selected for each dimension (4), i.e. 12211 means the respondent had no problems in mobility, pain/discomfort, and anxiety/depression, but had slight problems in self-care and usual activities. A VAS was used in the interview, with anchor points 0 ('worst imaginable health state') and 100 ('best imaginable health state'). Respondents first report their own health state using the EQ-5D-5L descriptive system and then their overall health on the EQ-VAS based on their health on the day of survey.

In 2012, the Chinese version of EQ-5D-5L was translated using a response scaling method (24), and its descriptors were proven to have similar interpretations to those of the English, Spanish and French versions (30). This version demonstrated validity and increased sensitivity in diabetes and hepatitis B patients (31, 32).

2.2.3 Data Analysis

For each respondent, the EQ-5D-5L health state and the EQ-VAS were directly observed from respondent's' self-report questionnaire while the EQ index score was derived from the Chinese EQ-5D-5L value set (9). In the EQ-5D-5L value set, the EQ index score of all 3,125 health states were estimated (9). For each respondent, we derived their corresponding EQ index score from their self-reported health states.

First, descriptive statistics of EQ-5D-5L health state, EQ-VAS and EQ index score were calculated for the whole sample and by different demographic variables and cities (age, gender, employment status etc.). For each demographic variable, the percentage of reported problem in EQ-5D dimension, the means (and 95% confidence interval) of EQ-VAS and EQ index scores were calculated for each subgroup and the difference were tested statistically. Second, we used multivariable analysis to examine the associations between demographic characteristics with reported problems in EQ-5D-5L, EQ-VAS and EQ index scores respectively. For the reported problems in each dimension, we used logistic regression ('no problems' coded as 0; 'slight problems', 'moderate problems', 'severe problems', or 'extreme problems/unable' coded as 1)(4). For EQ-VAS and EQ index scores, we used linear regression. All demographic variables including age and education level were entered into the models as categorical variables. Multivariable analysis was used to identify significant demographic characteristics using a backward selection procedure to remove covariates with $p>0.05$. Odds ratio was reported for logistic regression and coefficient was reported for linear regression respectively, the corresponding 95% CI was calculated using robust standard error.

For this study, ethical approval was not needed in China at the time of data collection. A waiver of the informed consent was approved as this study did not provide any intervention to participants. Participants can withdraw at any time without any consequences.

2.3 RESULTS

A total of 1332 individuals (response rate: 68.6%) who met the inclusion criteria were recruited. Among these, 1296 (97.3%) who successfully completed the questionnaire

were included in the analysis. The mean age of the sample was 42 years (SD: 16 years), the age ranged between 16 years to 85 years old. Females comprised 49.9% of the sample. Other demographic information is shown in Table 1.

Table 1: Demographic characteristics of all respondents

Variables	Our sample	
Age group, years	N	%
<19	109	8.4
20-29	229	17.7
30-39	244	18.8
40-49	272	21.0
50-59	220	17.0
60-69	155	12.0
>70	67	5.2
Gender		
Female	646	49.9
Male	650	50.2
Education		
Primary or lower	138	10.7
Junior & Senior high school	867	66.9
College or higher	291	22.5
Employment status		
Full time employees	382	29.5
Temporary worker & freelancer	451	34.8
Retired	240	18.5
Student	132	10.2
Other	91	7.0
Residence of origin		
City	757	58.4
County	86	6.6
Township or village	453	35.0
Health insurance		
Urban employee	551	42.5
Urban residence	304	23.5
New rural	296	22.8
Other	88	6.8
No	57	4.4

In total, 54% of the sample reported their health as '11111', followed by '11121', '11112', '11122', and '21121'. The percentages of 'no problems' were: 94.37% for mobility, 98.92% for self-care, 95.45% for usual activity, 70.14% for pain/discomfort, and 73.15% for anxiety/depression. The mean EQ-VAS and EQ index scores were 86.0 (SD: 11.4) and 0.957 (SD: 0.069), respectively.

Table 2: Percentage of a general population sample reporting levels 1 to 5 by dimension, EQ-VAS & EQ index score by age group for **males**

EQ-5D dimension		Age Groups							Total N=650
		<19 N=56	20-29 N=116	30-39 N=123	40-49 N=135	50-59 N=110	60-69 N=84	>70 N=26	
Mobility	No problems	100%	98.3%	98.4%	91.9%	96.4%	85.7%	69.2%	94.0%
	Slight problems	0%	1.7%	1.6%	8.2%	3.6%	13.1%	26.9%	5.7%
	Moderate problems	0%	0%	0%	0%	0%	1.2%	3.9%	0.3%
	Severe problems	0%	0%	0%	0%	0%	0%	0%	0%
	Unable to	0%	0%	0%	0%	0%	0%	0%	0%
	Z (P value)							5.69 (0.000)	
Self-care	No problems	100%	100%	100%	98.5%	100%	96.4%	96.2%	99.1%
	Slight problems	0%	0%	0%	1.5%	0%	3.6%	3.9%	0.9%
	Moderate problems	0%	0%	0%	0%	0%	0%	0%	0%
	Severe problems	0%	0%	0%	0%	0%	0%	0%	0%
	Unable to	0%	0%	0%	0%	0%	0%	0%	0%
	Z (P value)							2.65 (0.008)	
Usual Activity	No problems	96.4%	94.8%	95.9%	93.3%	99.1%	90.5%	92.3%	94.9%
	Slight problems	3.6%	5.2%	4.1%	5.9%	0.9%	7.1%	7.7%	4.6%
	Moderate problems	0%	0%	0%	0.7%	0%	2.4%	0%	0.5%
	Severe problems	0%	0%	0%	0%	0%	0%	0%	0%
	Unable to	0%	0%	0%	0%	0%	0%	0%	0%
	Z (P value)							0.95 (0.342)	
Pain/ Discomfort	No problems	78.6%	75.9%	78.1%	71.1%	64.6%	64.3%	57.7%	71.4%
	Slight problems	19.6%	23.3%	20.3%	26.7%	29.1%	31.0%	30.8%	25.4%
	Moderate problems	1.8%	0%	0.8%	1.5%	6.4%	4.8%	11.5%	2.8%
	Severe problems	0%	0.9%	0.8%	0.7%	0%	0%	0%	0.5%
	Extreme problems	0%	0%	0%	0%	0%	0%	0%	0%
	Z (P value)							3.44 (0.001)	
Anxiety/ Depression	No problems	67.9%	65.5%	66.7%	78.5%	75.5%	77.4%	88.5%	72.8%
	Slight problems	30.4%	32.8%	29.3%	20.7%	21.8%	20.2%	11.5%	25.1%
	Moderate problems	1.8%	1.7%	2.4%	0%	1.8%	0%	0%	1.2%
	Severe problems	0%	0%	0.8%	0.7%	0.9%	2.4%	0%	0.6%
	Extreme problems	0%	0%	0.8%	0%	0%	0%	0%	0.3%
	Z (P value)							-2.94 (0.003)	
EQ-VAS	Mean	87.4	86.9	85.5	85.5	84.8	82.9	83.9	85.4
	95%CI	84.4 90.4	85.2 88.5	83.8 87.2	83.3 87.8	82.6 87.1	79.9 85.9	76.9 90.9	84.5 86.3
	Z (P value)							-1.68 (0.093)	
EQ index score	Mean	0.968	0.963	0.961	0.959	0.956	0.943	0.932	0.957
	95%CI	0.957 0.978	0.953 0.973	0.950 0.972	0.948 0.971	0.946 0.967	0.921 0.964	0.897 0.966	0.952 0.962
	Z (P value)							-2.21 (0.027)	

Table 3: Percentage of a general population sample reporting levels 1 to 5 by dimension, EQ-VAS & EQ index score by age group for **females**

EQ-5D dimension		Age Groups							Total N=646
		<=19 N=53	20-29 N=113	30-39 N=121	40-49 N=137	50-59 N=110	60-69 N=71	>=70 N=41	
Mobility	No problems	96.2%	96.5%	99.2%	97.1%	95.5%	90.1%	73.2%	94.7%
	Slight problems	3.8%	3.5%	0.8%	2.9%	3.6%	8.5%	19.5%	4.5%
	Moderate problems	0%	0%	0%	0%	0%	1.4%	7.3%	0.6%
	Severe problems	0%	0%	0%	0%	0.9%	0%	0%	0.2%
	Unable to	0%	0%	0%	0%	0%	0%	0%	0%
	Z (P value)							4.68 (0.000)	
Self-care	No problems	98.1%	99.1%	99.2%	100%	99.1%	97.2%	95.1%	98.8%
	Slight problems	1.9%	0.9%	0.8%	0%	0%	1.4%	4.9%	0.9%
	Moderate problems	0%	0%	0%	0%	0%	1.4%	0%	0.2%
	Severe problems	0%	0%	0%	0%	0.9%	0%	0%	0.2%
	Unable to	0%	0%	0%	0%	0%	0%	0%	0%
	Z (P value)							1.42 (0.156)	
Usual Activity	No problems	96.2%	99.1%	98.4%	97.8%	96.4%	93.0%	78.1%	96.0%
	Slight problems	3.8%	0.9%	1.7%	2.2%	1.8%	7.0%	22.0%	3.7%
	Moderate problems	0%	0%	0%	0%	0.9%	0%	0%	0.2%
	Severe problems	0%	0%	0%	0%	0.9%	0%	0%	0.2%
	Unable to	0%	0%	0%	0%	0%	0%	0%	0%
	Z (P value)							4.36 (0.000)	
Pain/Discomfort	No problems	66.0%	74.3%	76.0%	69.3%	65.5%	64.8%	51.2%	68.9%
	Slight problems	30.2%	24.8%	23.1%	28.5%	32.7%	32.4%	39.0%	28.8%
	Moderate problems	1.9%	0.9%	0.8%	1.5%	0.9%	2.8%	7.3%	1.7%
	Severe problems	1.9%	0%	0%	0.7%	0.9%	0%	2.4%	0.5%
	Extreme problems	0%	0%	0%	0%	0%	0%	0%	0.2%
	Z (P value)							2.56 (0.010)	
Anxiety/ Depression	No problems	56.6%	62.8%	76.9%	75.9%	76.4%	85.9%	78.1%	73.5%
	Slight problems	37.7%	31.9%	20.7%	21.9%	21.8%	14.1%	19.5%	23.7%
	Moderate problems	5.7%	4.4%	2.5%	1.5%	1.8%	0%	2.4%	2.5%
	Severe problems	0%	0.9%	0%	0%	0%	0%	0%	0.2%
	Extreme problems	0%	0%	0%	0.7%	0%	0%	0%	0.2%
	Z (P value)							-4.02 (0.000)	
EQ-VAS	Mean	88.3	85.8	87.8	87.5	86.2	84.5	85.3	86.6
	95%CI	85.4 91.2	83.6 88.0	86.0 89.6	85.6 89.3	84.0 88.3	81.8 87.2	82.0 88.6	85.8 87.5
	Z (P value)							-1.75 (0.081)	
EQ index score	Mean	0.945	0.959	0.971	0.962	0.954	0.957	0.912	0.957
	95%CI	0.926 0.963	0.949 0.968	0.962 0.979	0.952 0.972	0.933 0.975	0.943 0.971	0.881 0.943	0.951 0.962
	Z (P value)							-1.04 (0.300)	

Tables 2 and 3 show the percentage of reported problems for each severity level and EQ-5D dimension, and the mean (SD) of EQ-VAS and EQ index scores for males and females by age groups, respectively. In both male and female groups, the number of problems increased with age in the dimensions of mobility, self-care, and pain/discomfort ($p<0.05$, trend test for ordered groups). In contrast, anxiety/depression was more prevalent in younger age groups ($p<0.01$, trend test for ordered groups). As could be expected, the means of both EQ-VAS and EQ index scores decreased with age, but only the EQ index score for male was statistically significant ($p<0.05$, trend test for ordered groups). Females reported higher EQ-VAS values than males ($p<0.05$, two-sample t-test). The highest mean EQ index score was observed for females of 30-39 years (0.971), the lowest mean score for females of > 70 years (0.912). The mean VAS score ranged between 88.3 for females of <19 years and 82.9 for males of 60-69 years.

Beside age and gender, Table 4 shows the percentage of any reported problem for each EQ-5D dimension, and the mean (SD) of EQ-VAS and EQ index scores by other socio-demographic characteristics. Lower education indicated more problems in mobility, usual activities and more pain ($p<0.05$, chi2 test). Lower education also had lower EQ index score ($p<0.05$, one-way analysis of variance). Percentage of any reported problem all differed by employment status ($p<0.01$, chi2 test), full time employees reported least problems with self-care and usual activities; students reported the least problems with mobility and less pain/discomfort; retired reported least anxiety/depression. Students reported the highest score in EQ-VAS and EQ index score. Insurance status seem did not affect the percentage of reported problems in any dimension, but the EQ-VAS of the insured was higher than those without insurance ($p<0.05$, two-sample t-test). In terms of original place of residence, residents from the city reported less anxiety ($p<0.01$, chi2 test). Difference were also found between cities in pain/discomfort, anxiety/depression, EQ-VAS and EQ index score.

Socio-demographic characteristics which significantly predicted any problems in EQ-5D dimensions, and EQ-VAS and EQ index scores, are reported in Table 5, where the reported problem in each dimension was reported as an odds ratio, and the EQ-VAS, EQ index scores were reported as regression coefficients. Notably, reported problems with anxiety/depression declined along age groups (odds ratio: 0.58 for 30-59 years; 0.40 for >=60 years respectively). Males had 1.45 lower EQ-VAS value than females. All outcomes varied with employment status. For example, compared to the group with full time job, unemployed group reported 4.04 lower EQ-VAS value and 0.03 lower EQ-index score, retired group reported 3.93 lower EQ-VAS value and 0.02 lower EQ-index score. Respondents from the county were found more reported problem in usual activities (odds ratio: 2.58).

Table 4: Percentage of a general population sample reporting any problem by dimension, EQ-VAS & EQ index score by other demographic variables

	Mobility	Self-care	Usual activities	Pain/ discomfort	Anxiety/ depression	EQ-VAS (95%CI)	EQ-index (95%CI)
Highest education							
Primary school & lower(n=138)	10.9%	1.4%	9.4%	37.0%	25.4%	84.8 (82.9, 86.8)	0.943 (0.924, 0.961)
High school(n=867)	5.4%	0.9%	4.1%	30.3%	24.6%	86.2 (85.5, 87.0)	0.959 (0.954, 0.963)
College & above(n=291)	3.8%	1.0%	3.4%	25.1%	34.4%	85.9 (84.7, 87.0)	0.959 (0.952, 0.965)
P value	0.01	0.91	0.01	0.04	0.00	0.40	0.04
Employment status							
Full time employee(n=382)	2.6%	0%	1.8%	28.3%	29.1%	87.5 (86.6, 88.5)	0.963 (0.957, 0.968)
Part time & freelancer(n=451)	4.2%	0.7%	4.7%	29.3%	26.6%	85.6 (84.5, 86.7)	0.960 (0.955, 0.966)
Retired(n=240)	13.3%	2.9%	7.1%	37.5%	16.7%	83.8 (82.2, 85.5)	0.948 (0.937, 0.958)
Student(n=132)	1.5%	0.8%	3.0%	17.4%	40.1%	88.7 (87.3, 90.0)	0.964 (0.957, 0.972)
Others(n=91)	11.0%	3.3%	11.0%	37.4%	26.4%	83.7 (80.7, 86.8)	0.930 (0.902, 0.957)
P value	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Insurance status							
With insurance(n=1,239)	5.7%	1.1%	4.4%	29.9%	26.8%	86.1 (85.5, 86.8)	0.957 (0.953, 0.961)
Without insurance(n=57)	3.5%	0%	7.0%	29.8%	28.1%	82.8 (79.4, 86.1)	0.953 (0.933, 0.974)
P value	0.48	0.42	0.36	1.00	0.83	0.03	0.71
Residence of origin							
City(n=757)	6.7%	1.1%	4.1%	31.2%	23.7%	85.6 (84.8, 86.5)	0.957 (0.952, 0.962)
County(n=86)	5.8%	1.2%	8.1%	32.6%	34.9%	85.4 (83.1, 87.7)	0.952 (0.941, 0.964)
Township or village(n=453)	3.8%	1.1%	4.6%	27.2%	30.7%	86.8 (85.8, 87.8)	0.957 (0.950, 0.965)
P value	0.09	0.99	0.23	0.29	0.00	0.19	0.82
Cities							
Beijing	3.0%	0%	2.3%	28.2%	17.9%	88.5(87.4, 89.7)	0.968(0.962, 0.974)
Chengdu	6.6%	1.2%	6.3%	34.8%	31.6%	84.9(83.4, 86.5)	0.949(0.941, 0.957)
Guiyang	7.7%	1.2%	5.8%	21.8%	28.0%	86.0(84.7, 87.2)	0.959(0.949, 0.969)
Nanjing	5.6%	1.5%	4.5%	36.7%	34.1%	85.2(83.8, 86.6)	0.948(0.939, 0.956)
Shenyang	5.2%	1.6%	4.0%	27.6%	22.4%	85.4(83.8, 87.0)	0.961(0.952, 0.969)
P value	0.21	0.41	0.21	0.00	0.00	0.00	0.00

Table 5: The association between HRQoL data and demographic factors (N=1,296)

Variables	Mobility	Self-care	Usual activity	Pain /discomfort	Anxiety /depression	EQ-VAS	EQ-index score
	Odds Ratio 95%CI					Coefficients 95%CI	
Age group (Ref: <=29 years group)	1.00	1.00	1.00	1.00	1.00		
30-59 years groups	1.14		0.87		**0.58 (0.44,0.77)**		
>=60 years groups	**4.89 (1.94,12.32)**		**3.67 (1.40,9.60)**		**0.40 (0.26,0.59)**		
Gender (Ref: female)	1.00	1.00	1.00	1.00	1.00		
Male						**-1.45 (-2.69,-0.22)**	
Health Insurance (Ref: no insurance)	1.00	1.00	1.00	1.00	1.00		
With Insurance						**3.36 (0.08,6.63)**	
Employment status (Ref: full time job)	1.00	1.00	1.00	1.00	1.00		
Temporary worker& freelancer	1.48	**0.20 (0.04,0.99)**	2.31	1.05		**-1.84 (-3.30,-0.39)**	-0.00
Retired	1.83	0.88	1.55	**1.52 (1.08,2.15)**		**-3.93 (-7.22,-0.85)**	**-0.02 (-0.03,-0.00)**
Student	0.65	0.22	1.52	**0.54 (0.32,0.88)**		0.98	0.00
Unemployed & others	**3.05 (1.22,7.62)**	1.00	**4.54 (1.74,11.87)**	1.51		**-4.04 (-6.63,-1.45)**	**-0.03 (-0.06,-0.00)**
Residence of origin (Ref: city)	1.00	1.00	1.00	1.00	1.00		
County			**2.58 (1.03,6.48)**				
Village			**1.19**				

Note: CI: Confidence interval

2.4 DISCUSSION

This is the first EQ-5D-5L norms study from China. These general population-based norms provide insights into HRQoL in China and how HRQoL varies between different socio-economic groups. More importantly, it facilitates interpretation of the cost effectiveness studies which use QALY as a health outcome. As HRQoL instruments measure postulated constructs, the set of normative values provides a reference point to interpret an HRQoL study's results by comparing HRQoL between the general population and patients with specific conditions from similar age and gender groups (33, 34).

Compared to the Chinese EQ-5D-3L norms reported in 2008 (11), our study showed a significant increase in problems reported in the last two dimensions. This could be either because there were more problems in these two dimensions compared to the past, or that the five-level EQ-5D was more sensitive in identifying the mild problems in these dimensions. While it is not possible to detangle such change in our study, in several studies comparing normative data between EQ-5D-3L and EQ-5D-5L, the researchers reported the 5L questionnaire suffered less ceiling effect , had less standard deviation in the index value, and had wider spread of health states, which all suggests the improved sensitivity for the 5L questionnaire (25, 32, 35). HRQoL inequalities were shown in China between different socio-demographic groups and regions, based on previous research (10-13). Such disparities were confirmed by our multivariable analysis, with lower socio-economic status related to lower HRQoL.

Some results from our study were in line with other countries' EQ-5D-5L norms (25, 27, 30, 32, 35-38): the first three dimensions of EQ-5D had less reported problems compared to the last two dimensions, with pain/discomfort being the most prevalent dimension; women reported lower EQ index score than men; EQ-VAS & EQ index score declined with age. Two differences were noted, first, in previous EQ-5D norms studies conducted in China and other countries, the percentage of reported problems in anxiety/depression increased with age (4, 25, 27, 30, 32, 39), our results suggest the opposite: the anxiety/depression problem was more prevalent in the younger population. One possible explanation is that the younger generation living in urbans areas perceived more psychological pressures than the older generation due to the fast-paced life in urban China. Second, females reported slightly higher EQ-VAS values than males, which is inconsistent with EQ-5D-3L norm values in China (11): this discrepancy could be due to the difference in the two study samples' compositions. The EQ-VAS score is predicted by several demographic variables and in our study sample, females were in higher socio-economic groups.

One limitation of this study is that the sample was collected in five urban areas in China, which is not representative of the whole Chinese population. As socio-economic differences exist between different areas, also between urban and rural areas in China, the health status of residents may differ by type of area (40). Furthermore, most respondents were recruited in public locations, therefore the sample may have left out those who were not able to go outside. This may have led to a selection bias towards healthy respondents and underreported problems with mobility and usual activities. Nevertheless, we did not correct for this bias in our result as we did not know the exact proportion of respondents missed out in the sample. Third, this is a cross-sectional study, which provided insights into relationship between HRQoL data and socio-demographic variables. In terms of understanding the causal relationship between variables and controlling for unobserved heterogeneity, longitudinal data is needed (41-43).

2.5 CONCLUSIONS

This study has offered the first EQ-5D-5L urban population norms for China. Disparities exist in self-reported health status measured by EQ-5D-5L across socio-economic groups. Further research into rural HRQoL and into using a national representative sample is warranted.

CHAPTER 3

Logical inconsistencies in time trade-off valuation of EQ-5D-5L health states: whose fault is it?

Zhihao Yang, Jan van Busschbach, Reinier Timman, M.F. Janssen, Nan Luo

We thank Elly Stolk from the EuroQol Office for her constructive suggestions for the early version of this manuscript, but the conclusion does not necessarily reflect her views.

Publication: Yang Z, van Busschbach J, Timman R, Janssen MF, Luo N (2017) Logical inconsistencies in time trade-off valuation of EQ-5D-5L health states: Whose fault is it? PLoS ONE 12(9): e0184883. https://doi.org/10.1371/journal.pone.0184883

3.1 INTRODUCTION

EQ-5D-5L is a preference-based quality of life instrument which is mainly designed to generate health-state utility values that are required for calculation of quality-adjusted life years (QALYs) and cost-utility analysis (44). With a classification system consisting of five dimensions (mobility, self-care, usual activities, pain/discomfort and anxiety/depression) and five levels of severity for each dimension (1=no problems, 2=slight problems, 3=moderate problems, 4=severe problems and 5=extreme problems), the instrument defines (5^5) = 3,125 unique health states, each of which can be represented using a 5-digit number or vector between 11111 (no problems in any dimension) and 55555 (extreme problems in all five dimensions). An important component of the instrument is the social tariff or value set that contains the utility values for all the health states it defines. With the value set available investigators can easily obtain the utility values of the EQ-5D-5L health states of interest, or find the utility values for their study populations by describing their health using the EQ-5D-5L classification system. Establishing the value set for a preference-based health related quality of life instrument is not a trivial task. The general approach is to elicit the utility values for a subset of the health states defined by the instrument and develop a regression model to predict the values for all the health states, including those not directly valued. In the case of EQ-5D-5L, the currently recommended study protocol (29) requires 1,000 or more members of the general public each to value 10 different health states using the time trade-off (TTO) technique. After the TTO task, the current EuroQol Valuation Technology (EQ-VT) protocol also includes 7 pairs of discrete choice experiment (DCE) for each respondent. A number of countries have used the study protocol to establish their local EQ-5D-5L value sets (16, 18). In this paper, we focus mainly on the TTO task.

One issue that has occurred in the valuation of EQ-5D-5L health states is that some respondents give logically inconsistent values. That is, better health states are valued as more undesirable than worse health states (45). For example, the state 11121 is valued lower than the state 22321. Logical inconsistency could be due to random mistake, however, if it occurs among a large proportion of respondents, it could signify the failure in the way the valuation technique is implemented. Regardless of the reason, such data lowers the precision of the estimated values. Specifically, logical inconsistency may attenuate the differences in values between health states (46) and consequently lead to underestimated health improvements when the values are used in cost-utility analysis (47). In some valuation studies, inconsistent observations were excluded when constructing the value set, thereby potentially affecting representativeness if certain sub-groups of respondents score more inconsistencies than others (45, 47-49). Hence

the magnitude of this issue and the underlying reasons should be investigated and, if possible, interventions should be implemented to minimize the potential bias caused by inconsistency.

Previous EQ-5D-3L valuation studies found that older and less-educated respondents were more likely to make inconsistent valuations (46, 49). EQ-5D-3L is similar to EQ-5D-5L except that there are only three descriptive levels for each dimension (no problems, moderate problems, and extreme problems). This result is not surprising as logical inconsistency could be due to poor understanding or misinterpretation of the valuation task (50, 51) . However, it is not clear whether this is the case in the valuation of EQ-5D-5L health states and to what extent logical inconsistency is related to interviewers. In EQ-5D-5L valuation studies, interviewers play an important role in the conduct of the valuation tasks, and they are trained to follow a standardized protocol. Nevertheless, interviewer effects have been observed in previous studies (52).

The aim of the present study was to ascertain the factors underlying individual-level logical consistency in an EQ-5D-5L valuation study. We hypothesized that logical inconsistency was related to multiple factors with respect to interviewers, the interview process, and respondents' background characteristics.

3.2 METHODS

3.2.1 Data source

This study makes use of data collected in the EQ-5D-5L valuation study in China. The purpose of the valuation study was to establish the EQ-5D-5L value set in China from a societal perspective. The target population was urban residents in China (9). Detailed description of the valuation study have been published elsewhere (9). In the valuation study, the EQ-5D-5L was translated through a response scaling approach, which ensured the Chinese descriptors have similar interpretations with English counterpart (24). Briefly, the study recruited members of the general population from five cities, namely: Beijing, Nanjing, Shenyang, Chengdu, and Guiyang (9). In each city, members of the general population were recruited from a number of public places including community centers, parks, shopping centers, and university campuses. Sampling quotas were applied so that the sample resembled the target population in terms of age, sex, and education (6, 9). Inform consent was given to the respondent before conducting the interview (9), and ethics approval was not needed for this study in China as the valuation task is not seen as a medical intervention. Each respondent was interviewed face-to-

face by a trained interviewer using the EQ-VT platform (28). The interview had four sections. The first section was for respondents to report their own health using the EQ-5D-5L questionnaire, and their experience with serious illness. The second section asked respondents to complete 10 TTO tasks, each valuing a different EQ-5D-5L health state. The third section contained a set of discrete choice questions designed for valuation of selected EQ-5D-5L health states based on random utility theory. Data collected in this section was not used in the present study. The fourth section assessed respondents' socio-economic and other background characteristics.

The 'composite' TTO technique was used in the study. This employs conventional TTO and lead-time TTO (53) to value better-than-dead and worse-than-dead states, respectively. The two TTO variants are described in detail elsewhere (54). Briefly, conventional TTO elicits the raw value x ($0 \leq x \leq 10$) at which the respondent is indifferent between two alternatives: 1) living in full health for x years, and 2) living in an EQ-5D-5L health state for 10 years. The utility value is given by x/10. For health states considered to be worse than dead, the two alternatives in the valuation task are: 1) living in full health for x years, and 2) living in full health for 10 years and then in an EQ-5D-5L health state for another 10 years. The utility value is given by x/10 -1.

At the interviews, the interviewer demonstrated and explained how the composite TTO works to the respondent using the state of 'in a wheelchair' as an example, before proceeding to the formal TTO tasks for the valuation of 10 different EQ-5D-5L health states (29). The EQ-VT platform was designed to value a total of 86 EQ-5D-5L health states considered sufficient for the estimation of a value set. These 86 health states were divided into 10 blocks in such a way that each block consisted of the worst state (55555), one of the five mildest states (21111, 12111, 11211, 11121, 11112), and eight other unique health states. Each respondent was randomized to value one block of health states which were presented to the respondent in a random order.

A total of 20 interviewers, 4 for each city, conducted the interviews (9). The interviewers were students and researchers from local universities. They were trained at a full-day workshop by their respective site project leaders who were trained in the same way by the principal investigator. The training focused on the use of a standardized protocol to conduct the interview, the principles of the TTO technique, and the objectives of the valuation study. As the TTO task was difficult to conduct, interviewers were instructed to perform multiple 'practice' interviews during and after the workshop with their peers and friends or family members.

3.2.2 Measures of inconsistency

At the respondent level, the magnitude of logical inconsistency was assessed using three indicators: inconsistency rate, distance, and ΔTTO. Inconsistency rate was the number of inconsistently valued pairs of health states divided by all possible logical pairs. Inconsistency distance was calculated as the sum of the squared difference in levels for corresponding dimensions of the two health states involved. For example, the level differences between health states 12344 and 44444 were respectively 3, 2, 1 in the first three dimensions and 0 in the latter two, and thus the distance was $3^2 + 2^2 + 1 = 14$. ΔTTO was the difference in utility values of two inconsistently valued health states. For example, if one respondent gave 21222 a utility 0.8 and 11112 a utility 0.5, the ΔTTO of this inconsistency would be 0.3.

Owing to the highly skewed distribution of inconsistency in all 3 indicators across respondents, as in other studies (46, 50), respondents were categorized into 3 levels: none, slight, and severe. 'None' was defined as no observed inconsistency; 'severe' was defined as inconsistency rate higher than 10%, average inconsistency ΔTTO larger than 0.2, and average inconsistency distance larger than 9; and 'slight' was applied for respondents whose inconsistency profiles were neither 'none' nor 'severe'(48, 55). So, a respondent is classified as severe inconsistent if he/she made more inconsistencies and those inconsistencies were more severe.

3.2.3 Data analysis

Inconsistency factors studied included respondents' demographic characteristics, interviewer identity, and interview process indicators. Respondents' characteristics were age (16-24 years, 25-34 years, 35-44 years, 45-54 years, 55-64 years, 65-74 years, ≥75 years), gender, and education (primary or lower, junior high school, senior high school, college or university, Masters or PhD). Interview process indicators were: time spent on the wheelchair example, number of iterations in the wheelchair example, and time spent on the 10 TTO tasks. The number of iterations indicated how many steps a respondent had moved before the indifferent point was reached in a TTO task. The number of iterations and the time spent on the wheelchair example, and the formal TTO tasks may reflect to what extent respondents and interviewers were engaged in the valuation tasks.

An additional process characteristic examined was the sequence of the interviews, that is, the rank order of the interviews conducted by the same interviewer in terms of the interview date and time. It was hypothesized that there was a learning curve for the interviewers in the study such that the quality of the interviews increased with the number of interviews that an interviewer completed. As a result, more interview

experience would lead to a lower level of logical inconsistency.

A two-level multi-nominal logistic model (Equation 1) with the interviewer as the upper level and the respondent as the lower level was used to explore logical inconsistency factors. This model estimated the average effects of the lower-level factors among the interviewers. The requirement to discern levels was determined using likelihood ratio tests (56). Age, gender, education level (edu), interview sequence, TTO time, TTO iteration (ttoit), wheelchair time and wheelchair iteration were entered as covariates. The covariates sequence, times, and iterations were standardized (by dividing the raw data with its Standard Error) in order to enhance interpretation of the relative risk ratios (RRR) for category i compared to the reference category no inconsistencies. A RRR > 1 suggests an increased risk of that outcome compared to the reference group. A RRR between 0 and 1 suggests a reduced risk compared to the reference group.

$$RRR = e^{\beta 00 + u_{0j} + \beta_{1i} age + \beta_{2i} edu + \ldots + \beta_{8i} ttoit} \quad (1)$$

Where β_{00} is the overall mean intercept and u_{0j} is the random intercept to identify clusters, here: interviewers.

Additional analysis determined whether there were differences in inconsistencies between the interviewers. As 'interviewer' was included as a between-subject factor in this analysis, a single-level multi-nominal regression model (Equation 2) which included both interviewer and the above-mentioned covariates was used. Relative risk ratios, their 95% confidence intervals, and p-values of the independent variables were estimated using STATA version 13.1. Covariates were deleted in a backward procedure, with p>0.05 as the criterion for deletion. Interaction terms between statistically significant covariates were created and examined based on the results of the two models.

$$RRR = e^{\beta_{0i} + \beta_{1i} age + \beta_{2i} edu + \ldots + \beta_{8i} ttoit + \beta_{9i} inter2 + \ldots + \beta_{23i} inter20} \quad (2)$$

3.3 RESULTS

3.3.1 Data description

Of 1,302 participants in the valuation study, 1,296 finished the interview. Each of the 20 interviewers conducted at least 50 interviews. Table 1 summarizes the demographic information of the interviewees and the summarized information of the interview process.

Table 1: Demographic information of interviewees and the summarized information of interview process

Variables	Total sample
Age group (years)	(N, %)
16-24	235, 18%
25-34	231, 18%
35-44	237, 18%
45-54	258, 20%
55-64	222, 17%
65-74	79, 6%
≥75	34, 3%
Gender	(N, %)
Male	650, 50%
Female	646, 50%
Education	(N, %)
Primary or Lower	138, 11%
Junior high school	405, 31%
Senior high school	462, 36%
College or University	225, 17%
Masters or PhD	66, 5%
Interview Sequence (Rank orders)	(Mean, SD)
	33.4,19.6
Time spent on TTO task (Minutes)	(Mean, SD)
	14.2,5.3
Time spent on Wheelchair example task (Minutes)	(Mean, SD)
	6.3,3.2
Iterations spent on TTO task (steps)	(Mean, SD)
	7.9,2.5
Iterations spent on Wheelchair example task (steps)	(Mean, SD)
	22.1,11.9

Out of 1,296 respondents, 723 (56%) did not display any inconsistency; the remaining 44% gave at least one inconsistent response. The numbers of respondents who were 'slightly' and 'severely' inconsistent amounted to 499 and 74 respectively. The rate, distance, and ΔTTO of logical inconsistency are summarized in Table 2.

3.3.2 Factors associated with inconsistency

Significant variables associated with logical inconsistency and their effects in the two-level model are displayed in Table 3. The likelihood ratio test showed that both levels (interviewers and respondents) were statistically significant ($P < 0.01$). Three variables were significantly associated with slight inconsistency and another two variables were associated with severe inconsistency (Table 3). Specifically, more time spent on the

Table 2: Inconsistency severity measured by three criteria

Measurement criteria	Severity degree	Numbers identified	Total inconsistency rate	Average inconsistency distances	Average inconsistency ΔTTO
Inconsistency rate	Slight	447	0.045	14.287	0.235
	Severe	126	0.169	22.713	0.333
Inconsistency distance	Slight	160	0.040	4.966	0.254
	Severe	413	0.085	20.469	0.257
Inconsistency ΔTTO	Slight	325	0.059	15.317	0.096
	Severe	248	0.090	17.219	0.467
Inconsistency fulfilled all criteria	Slight	499	0.056	14.946	0.223
	Severe	74	0.189	24.194	0.482

wheelchair example, less time spent on the TTO task, and interviews completed at a later sequence, were associated with less likelihood of slight inconsistency; female respondents, and interviews completed at a later sequence were associated with less likelihood of severe inconsistency. The RRR is interpreted as, for example, compared to reference group, the risk of being slightly inconsistent is 1.246 times higher for every one unit of more time spent on TTO task.

Table 3: Inconsistency: multi-level multinomial logistic model in full dataset, N=1,269

Variables	RRR (unadjusted)	95%CI	RRR (adjusted)	95% CI
0 (Reference level: no inconsistency)	Base outcome		Base outcome	
1 (Slight inconsistency)				
Sequences (Rank orders)	0.810**	0.720, 0.912	0.806**	0.707, 0.918
Standardized time spent on TTO task	1.081	0.957, 1.220	1.246**	1.076, 1.441
Standardized time spent on wheelchair example	0.855*	0.755, 0.967	0.815*	0.699, 0.952
2 (Severe inconsistency)				
Sex	1.997**	1.230, 3.243	2.347**	1.429, 3.855
Sequences (Rank orders)	0.540**	0.417, 0.699	0.511**	0.385, 0.678

"Sex" is coded "0" for female respondent, and "1" for male respondent.
** Significant at 0.01 level.
* Significant at 0.05 level.

Two interviewers were found to be associated with a higher likelihood of slight and/or severe logical inconsistency in the single-level model (Table 4). One of the interviewers was particularly unusual as the relative risk ratio were found to be much higher compared to those conducted by an averagely performed interviewer, after adjusting for covariates. Interaction terms (i.e. education level of respondent*interviewer) were explored and proved less interesting in terms of statistical significance.

Table 4: Interviewer effect on inconsistency: multinomial logistic model in full dataset, N=1,296

Variables	RRR (unadjusted)	95% CI	RRR (adjusted)	95% CI
0 (Reference level: no inconsistency)	Base outcome		Base outcome	
1 (Slight inconsistency)				
Interviewer 7	3.486**	1.506, 8.071	3.476**	1.475, 8.191
Interviewer 9	2.242*	1.073, 4.683	2.659*	1.241, 5.696
2 (Severe inconsistency)				
Interviewer 7	8.054**	2.205, 29.411	7.335**	1.908, 28.195

Dummy variables 'interviewer' represent different interviewers, the reference level is 'interviewer1' from Shenyang, whose inconsistency level is the median among all interviewers.
** Significant at 0.01 level.
* Significant at 0.05 level.

3.4 DISCUSSION

As hypothesized, the factors interviewer, interview process, and respondent were all related to individual level logical inconsistency in the valuation of EQ-5D-5L health states. In terms of respondents' characteristics, male gender was associated with severe logical inconsistency. One explanation could be that male respondents might have had poorer engagement than females in the present study. In the previous EQ-5D-3L valuation study conducted in China, young and well-educated respondents were more likely to give inconsistent TTO answers (40). Unlike previous studies (46, 49, 50), older age was not associated with logical inconsistency in the present valuation study. This could be due to the efficiency of the survey tool: a computerized software program was used to demonstrate the valuation tasks in the EQ-5D-5L valuation study while a time board was used in previous studies. It should be noted that respondents' characteristics such as gender are not modifiable factors in valuation studies aiming at establishing a societal value set. For such studies, samples should be representative of the general population in terms of demographics. Hence, respondents who are more susceptible to logical inconsistency, cannot be removed from EQ-5D-5L valuation studies; the only intervention is to have interviewers pay more attention to these respondents.

More importantly, we found that interviewer and interview process indicators were independently associated with logical inconsistency. Specifically, interviews conducted by certain interviewers, those conducted earlier on by interviewers (sequence effect), and those in which less time was spent on the wheelchair example, suffered more from this issue. The variations across interviewers suggest that some interviewers did not perform to the expected standards. This could be due to poor understanding of the valuation tasks or poor compliance to the interview protocol. The sequence effect suggests that interviewers might still have been on a learning curve, that is, they had not

been versed enough in conducting the interviews at the time they started. Wheelchair time might be an indicator of training adequacy: when this was inadequate, logical inconsistency would increase. It is notable that the more time spent on TTO tasks, the more inconsistency occurred. One explanation could be that if the respondents did not understand or engage in the task, it took them longer to finish the TTO tasks while this did not warrant consistent responses.

Therefore, our study supports the extension of EQ-5D-5L valuation protocol with a quality control (QC) tool (57). It also should be noted that this data collection was done in the first version of EQ-VT protocol. The new protocol with the several modification to the original protocol, including the QC process lower the inconsistency rate from 11% to 3% (57). By using the new valuation protocol with QC tool, individual interviewers are monitored during the entire data collection period for their performance including time spent on explaining the wheelchair example (57). This monitoring is possible because the information is collected by the survey program and uploaded by interviewers on a daily basis. Nevertheless, our study suggests that future EQ-5D-5L valuation studies could benefit from more training for interviewers. In addition, our findings could be generalizable to other interviewer-administered health-state valuation study. The role of interviewers and the importance of interviewer training might be more crucial than hitherto considered, especially for the valuation study that is done without proper QC process during the data collection.

This study raised the question concerning how to handle logical inconsistency in establishing an EQ-5D-5L value set: should the inconsistent data be removed? Past studies showed that keeping inconsistent data will attenuate the differences in values between health states (58). On the other hand, if inconsistent responses are systematically higher in certain groups of respondents (e.g. male respondents), removing these data will affect the representativeness of population samples (45). Only a few EQ-5D-3L value sets were estimated by excluding some of the logically inconsistent data (40, 47, 55). Nevertheless, it can be postulated that values of extreme health states may be biased if logical inconsistency occurs with respect to these states. For example, good health states are unlikely to be overestimated because the logical inconsistency is one-sided: such health states are more likely to be valued lower rather than higher because the valuation tasks are designed in a way that no health states can be valued as > 1.0, the upper bound of utility value. Hence it is advisable to assess the effect of logical inconsistency on the estimated EQ-5D-5L value set.

One limitation of this study is that we limited our analysis of logical inconsistency to logistic analysis due to the skewed distributions of inconsistency at individual level. Moreover, the classification of inconsistency in the logistic model was arbitrary. There is no a well-accepted definition for 'slight' inconsistency or 'severe' inconsistency. However, in this study, in order to identify between "those who made careless mistakes" and "those who seem do not understand the task at all", the line was drawn.

In conclusion, logical inconsistency in the valuation of EQ-5D-5L health states is associated not only with respondents' characteristics but also with interviewers' performance and the interview process. Our study has highlighted the importance of interviewers for health-state valuation using the TTO elicitation procedure.

CHAPTER 4

Selecting health states for EQ-5D-3L valuation studies: statistical considerations matter

Zhihao Yang, Nan Luo, Gouke Bonsel, Jan Busschbach, Elly Stolk

Publication: Yang Z, Luo N, Bonsel G, Busschbach J, Stolk E. Selecting health states for EQ-5D-3L valuation studies: statistical considerations matter. Value in Health: 21 (2018) 456 – 461. https://doi.org/10.1016/j.jval.2017.09.001

4.1 INTRODUCTION

The EQ-5D-3L instrument is the most widely used preference-based health-related quality of life questionnaire (59). The EQ-5D descriptive system consists of 5 dimensions (mobility, self-care, usual activities, pain/ discomfort and anxiety/depression) with 3 ordinal severity levels each (no problems, some problems and extreme problems), thus defining 243 distinct health states (60). The key feature of a preference-based instrument is that health state 'values' (some prefer 'utilities' or 'index values') are derived for all of its health states, which indicate how good or bad each health state is. These numbers are assumed to have ratio properties and can be used to estimate quality-adjusted life years (QALYs).

The QALY is a preferred health outcome measure in cost-effectiveness studies around the world (61-64). The results of cost-effectiveness studies could be biased if values of health states cannot be well-estimated. Nevertheless, arriving at a set of 243 values for all separate states is a challenge: the valuation methods used can be demanding to the respondent, leading to data collection methods where every respondent usually values only a defined subset of all 243 states. From such a data subset, typically, parametric regression analysis enables prediction/extrapolation of the values for all the health states. Different design choices (selection of subsets) have been documented in other areas (65, 66), but in the literature on health state valuation, it is an open question how to select health states for inclusion in the subset for direct valuation. Different desirable properties have been identified which cannot all be satisfied at the same time. In the absence of straightforward statistical rules, selection has thus far been consensus-based. Hence, researchers take a leap of faith when estimating a value set while the advantages and disadvantages of their design choice are unknown. In this paper, we aim to find the best design (subset of health states) for EQ-5D valuation studies by comparing the performance of different desirable design properties in prediction accuracy.

A key EQ-5D valuation study in the context of design choices was the Measurement and Valuation of Health (MVH) study conducted in the UK. Its consensus-based design has been replicated frequently (67). In the MVH study, the following design criteria were selected (67):

- The set of states should spread widely over the valuation space so as to include as most combinations of levels across the five dimensions.
- Prima facie implausible states were excluded to sustain the credibility of the task, and to reduce errors in assigned values.

- All plausible combinations of dimension levels were to be included to allow identification of (first level) interactions.
- All respondents should value 2 out of the 5 mildest states which were most prevalent, and all respondents had to value the anchor state '33333', so that for every individual the utility range was known ('11111' was defined as 1.0).
- Some of the selected health states had been used in previous studies, thus maintaining a link to these studies.

Using the above criteria the MVH study selected 42 health states, where the health data of the respondents themselves were used to determine common health states. The valuation of these 42 states was used to predict the values of all 243 states, including the 201 not included in the data subset (68). In later research , the possibility that the MVH approach contained redundant values was investigated (60), but without returning to the desirability of the requirements shown above that were originally imposed on this approach.

While the face validity of the above sampling criteria appears apparent, a disadvantage is the lack of attention to the statistical properties of the resulting design. From a statistical point of view, desired properties of a design are level balance and level pair balance (i.e. orthogonality). These two properties allow for statistical decomposition of all separate dimension effects dependent on the level of other dimensions. Nevertheless, both properties were ignored in order to adhere to the desirable properties emphasized in the MVH design. Whether this design reflects the best compromise with given resources is hitherto unknown.

Any comparison of design strategies requires quantitative criteria on what constitutes the 'best' strategy. Bonsel et al (69) systematically compared designs on the MAE (mean absolute error) criterion, exploiting the possibilities of a 'saturated data set' that contained observed values of all 243 EQ-5D-3L health states. The emphasis was on questions around the type (flat, stressed, mild, and severe) and number of health states that may be selected for valuation exercises (affecting estimation bias), and on whether or not the introduction of interaction terms in the model increased or reduced the risk of mis-specifying the value of out-of-sample health states. In their study, random selection of health states worked better than selection of states of specific types (as above) and the predictions became acceptable when the sampling ratio reached 10% (in EQ-5D-3L, that is 25 health states.) In addition the 'main effects' model appeared to be a crude but 'safe' tool for estimation bias analysis (69).

The present study built upon the work outlined in the last paragraph, with a specific focus on strategies for selecting health states. There were 2 competing design principles at issue that both have their merits. 1) Historically, health states for inclusion in a design tended to be hand-picked, based on the properties of those health states (such as whether they were common), and on easily recognizable properties of the set as a whole (spanning the value range), but without an eye for the statistical properties of the set. 2) Optimization of statistical properties was an alternative route to follow, with an orthogonal design looking like a promising alternative to the designs that had been used historically.

This study aimed to assess the comparative performance of the designs created, while giving different weight to these principles. For this purpose, we compared designs on the basis of root mean squared errors (RMSEs), whilst also taking into account that from a user's perspective, misprediction of common health states could be considered as a mistake to be penalized more heavily than misprediction of rare health states (defined below). At first glance, over-representation of common health states (with assumed better face validity and data quality compared to rare health states) could lead to more accurate estimations for the common health states which shared the mild levels. On the other hand, this could also lead to reduced statistical efficiency compared to balanced designs. Thus it would be unknown whether the best prediction of a common state was achieved.

While we were aware that employing more design choices (the number of respondents, the number of health states per respondent, and the use of blocks (69)) may affect misprediction, this study focused on the above 2 design principles. Hence 'design' refers only to the deliberate selection of a subset of health states. Our research questions were addressed by testing a variety of designs, using a pre-existing saturated data set with observations on all health states for reference purposes, and by using different RMSE-based performance measures with and without focus on common health states.

4.2 METHODS

4.2.1 Research strategy

We used an existing data set with visual analogue scale (VAS) values from 126 students, each of whom valued all 243 EQ-5D-3L states. We generated a series of designs and subsequently modelled data subsets derived from each design.

Some of these designs were used previously, e.g. the MVH subset; others were newly generated, based on our proposed design strategies. The performance of the different

designs was evaluated in terms of the lowest RMSEs for all health states taken together, and for common and rare states separately. Common health states were defined in terms of the frequency of their occurrence in the 3 reference data sets (see below).

4.2.2 Existing data set with VAS values

In 1996 a students' panel (n=126) provided EQ-VAS values for all 243 EQ-5D-5L health states. Such a dataset is called 'saturated' and if the resulting data set was regarded as suitable, no regression was needed to generate a value for each health state.

The EQ-VAS was displayed as a standard vertical 20 cm scale to record an individual's rating for a health state. By using the EQ-VAS, each health state could be valued on a scale from 0 (the worst score) to 100 (the best score). The students received 41 sheets of paper, each containing 6 EuroQol health states, except for the last sheet which contained 1 health state plus the states 'unconscious' and 'dead'. 11111 and 33333 were valued twice at the outset before valuing all other states, and then with all the other health states. The first time the values of 11111 and 33333 were used as anchors while the second time these values were used for analysis. The EQ-VAS was shown on a separate paper. The standard abbreviations for the health states (e.g. '11132') were printed above the health states in order to provide a shortcut for the stimuli. Next to the standard abbreviations, the students were able to fill in the values of the health states. In order to eliminate framing effects, the health states were presented in 10 different random orders. The students were instructed to value all health states, even when they thought that a health state was unrealistic. The investigators piloted the procedure (70). Each student was awarded 35 Dutch guilders (equivalent to € 16.35 today).

4.2.3 Three reference data sets for the identification of common health states

The classification of each of the 243 possible health states as common or rare in our study was based on the frequency of their occurrence in 3 patient and population pooled data sets holding data on N=5,269 people in total. The data sets are described elsewhere (26, 71, 72). We defined the health states that never occurred in our sample as 'rare health states', and the ones we did observe were classified as 'common health states'.

4.2.4 Tested experimental designs

Table 1 provides a summary of the 10 different designs that were compared in this study.

Table 1: Tested designs & their characteristics

ID	Design names	# health states	# mild, moderate, severe health states	#common health states in design (percentage)
1.	MVH	42	13,12,17	32 (76.2)
2.	Japan	17	08,04,05	13 (76.5)
3.	Paris	25	09,05,11	19 (76.0)
4.	Bagust	47	16,10,21	45 (95.7)
5.	*Small Orthogonal*	*18*	*/*	*/*
6.	*Random (Common states only)*	*18*	*06,06,06*	*18 (100)*
7.	*Random (All states)*	*18*	*06,06,06*	*/*
8.	*Large Orthogonal*	*54*	*/*	*/*
9.	*Random (Common states only)*	*54*	*18,18,18*	*54 (100)*
10.	*Random (All states)*	*54*	*18,18,18*	*/*

* Italicized designs (5-10) repeated 100 times as such designs include randomization sampling.

Designs 1-3 have been used historically in EQ-5D valuation studies (59, 68, 73-76). The MVH' design is one of the earliest designs used in EQ-5D valuation studies (68), and both 'Paris' and 'Japan' designs were generated on the basis of the 'MVH' design, aiming at lowering the number of states that needed to be valued in valuation studies (59, 60, 75, 76).The designs of EQ-5D valuation studies have been critiqued by Bagust (67), and design 4 was proposed by him as an alternative. Designs 6, 7, 9 &10 comprised health states that were randomly selected from 3 groups with differing average severity. We began by categorizing all health states into 3 groups (mild, moderate and severe) based on their misery index (e.g., misery index of '32121'=3+2+1+2+1=9). In designs 6 and 9 the candidate set was restricted to common health states; in 7 and 10 all health states were included. Designs 5 and 8 were built starting from the experimental generating principle, hence these were orthogonal, and the arrangement of levels across all different health states is commonly called an 'orthogonal array'. Design 9 was obtained from an orthogonal main effects plan [http://neilsloane.com/oadir/index.html; design: oa.18.7.3.2](77), whereas design 10 [Hedayat et al 1997; design: oa.54.5.3.3.c](78) represented an orthogonal array that also enabled identification of two-way interactions. From the emerging set, a large number of equivalent yet different designs could be derived which maintained orthogonality.

As neither the principle of randomization nor the principle of orthogonality always produced the same design, we created multiple (100 times) variants of each (designs 5 to 10) to test by how much, if any, the misprediction performance was dependent on the variant.

4.2.5 Normalization and modelling of VAS values

To deal with the issue that each respondent may have utilized the EQ-VAS scale differently, we used Formula 1 to rescale the VAS value to anchor it on the extremes full health

(11111) and worst health state (33333). Note that we did not rescale on the value of death, because the value of death is controversial and would introduce relatively more variance (error) than the value of 33333.

V (health state) = (VAS (health state) -VAS (33333))/ (VAS (11111) -VAS (33333)) (1)

As the VAS scores 11111 and 33333 were required as anchored health states, we dropped the observations with missing values in 11111 and/or 33333. After this data cleaning, we estimated the 'main effects' ordinary least squared (OLS) model to predict values for all health states (68). In this model, each dimension was assumed to be independent of the others and no interactions between dimensions were used (17, 59, 69). The background characteristics of respondents such as age and gender were not entered into the model, as the purpose of the valuation study was to predict the values of health states rather than predicting how an individual evaluates a health state (68).

The equation was:

$$V(\text{health state}) = \alpha + \beta_1 MO_2 + \beta_2 MO_3 + \beta_3 SC_2 + \beta_4 SC_3 + \beta_5 UA_2 + \beta_6 UA_3 + \beta_7 PD_2 + \beta_8 PD_3 + \beta_9 AD_2 + \beta_{10} AD_3 + \varepsilon. \quad (2)$$

V (health state) is the rescaled VAS value given by formula 1 and it is explained by 10 variables and one intercept. Each dimension (MO for mobility, SC for self-care, UA for usual activity, PD for pain/discomfort, AD for anxiety/depression) has two dummy variables to represent the move from level 1 to level 2 or level 3, e.g. MO_2 takes 1 if the health state has a problem in the second level of mobility, takes 0 if otherwise (68). The coefficients β_i indicate the 'disvalue' of each variable associated with its move away from 11111. Given that the value of 11111 is 1 and the intercept represents any deviation from 11111, the intercept is then interpreted as a discontinuity between the value of full health and all other health states (68, 75).

4.2.6 Analysis

The performance of different designs in predicting health states values was quantified through computation of the RMSE as the primary measure of misprediction. We used two approaches to report the RMSE, by distinguishing between: i) health states included/ omitted in the design, and ii) commonness of the health states. For overall comparison of the designs, we also reported the general RMSE for all 243 health states. For the designs with 100 variants (random selection and orthogonal designs), we also presented the mean and variance of RMSEs, using Box plots. For reference, we listed the mean,

standard error, 95% confidence interval and predictions from three designs for 10 health states from different misery index groups. Additionally, we also estimated the number of health states with large prediction error (RMSE not applicable, absolute error -AE- used instead, where AE>0.05 & AE>0.10 were applied as criteria) to see whether a design predicted fairly for all health states.

4.3 RESULTS

4.3.1 EQ VAS data

All 126 students completed the main 243 health states valuation task. The average time to complete this task was 1 hour 16 minutes. The time for the fastest decile was 50 minutes, the time for the slowest decile was 1 hour 43 minutes. Data cleaning was minimal, with only 40 state values missing. Missing data was not imputed. Individual inspection of the outliers revealed that one respondent reversed the scale and one respondent gave highly inconsistent responses, valuing many clearly good states much lower than clearly bad states. These 2 respondents were excluded.

The mean V (health state) across all states and respondents was 0.370 (SD 0.214). The mean top and bottom values were V (11111) = 0.965 (SD 0.165), and V (33333) = 0.011 (SD 0.132).

4.3.2 Reference data sets for identification of common health states

In total, 55% (143) of health states out of 243 were observed in our data sets. Among the common health states, 43% (57) occurred at least 5 times. The most frequently observed health state was 11111.

4.3.3 Performance of various designs

Table 2 summarizes the RMSEs produced by different designs. The RMSEs were calculated at an aggregate level. Orthogonal designs (5 and 8) dominated performance compared to the published designs (1-4) and the random selection designs (6, 7, 9 and 10). Increasing the proportion of common health states in a design did not necessarily improve the predictions for the common health states, but resulted in larger RMSEs for rare health states. The regression estimates can be found in the Appendix.

To show some examples of differences in performance, Table 3 lists the observed means and predictions for ten health states in our sample.

Table 2: RMSEs of different designs

ID	Design name	i) RMSE by inclusion for modelling			ii) RMSE by commonness		RMSE for all health states, n=243
		#Number of included health states	Included states, n=#	Not included states, n=243- #	Common states, n=133	Rare states, n=110	
1	MVH	42	0.035	0.048	0.043	0.050	0.046
2	Japan	17	0.037	0.079	0.063	0.092	0.077
3	Paris	24	0.035	0.065	0.055	0.070	0.062
4	Bagust	47	0.031	0.090	0.064	0.099	0.082
5	*Small Orthogonal*	18	0.023	0.042	0.047	0.031	0.041
			0.021,0.025	0.040,0.042	0.045,0.048	0.029,0.032	0.040,0.041
6	*Random (Common states only)*	18	0.022	0.062	0.056	0.063	0.060
			0.020,0.024	0.057,0.065	0.053,0.058	0.057,0.068	0.056,0.063
7	*Random (All states)*	18	0.019	0.052	0.055	0.043	0.051
			0.018,0.020	0.049,0.055	0.052,0.057	0.039,0.046	0.048,0.053
8	*Large Orthogonal*	54	0.033	0.035	0.041	0.026	0.035
			0.031,0.034	0.034,0.036	0.040,0.042	0.025,0.026	0.035, 0.035
9	*Random (Common states only)*	54	0.035	0.043	0.040	0.043	0.042
			0.033,0.036	0.042,0.044	0.039,0.041	0.041,0.044	0.041,0.043
10	*Random (All states)*	54	0.030	0.038	0.042	0.028	0.037
			0.029,0.030	0.0.37,0.039	0.041,0.043	0.027,0.029	0.036,0.037

*Italicized designs (5-10) repeated 100 times; shown here are the average results for their 100 variants and the 95%CI below the mean.

Table 3: Observed means & predicted means for 10 random health states from each misery index group

Health states	Observed means	Standard error(SE)	95%CI		Predicted by Japan design	Predicted by MVH design	Predicted by Large Orthogonal design (mean, SE)*
11121	0.761	0.014	0.735	0.788	0.731	0.705	0.651,0.002
11311	0.592	0.017	0.558	0.626	0.571	0.605	0.591,0.001
11132	0.427	0.020	0.388	0.466	0.494	0.494	0.468,0.001
13212	0.468	0.015	0.440	0.497	0.567	0.515	0.502,0.001
12133	0.304	0.016	0.274	0.335	0.308	0.347	0.321,0.001
33113	0.284	0.014	0.257	0.311	0.449	0.352	0.300,0.000
13332	0.229	0.013	0.204	0.254	0.200	0.221	0.234,0.001
33313	0.169	0.012	0.145	0.192	0.213	0.177	0.162,0.001
23333	0.118	0.014	0.091	0.145	0.049	0.063	0.078,0.001
33333	0.011	0.012	-0.012	0.034	-0.043	-0. 023	-0.031,0.001

*For the large orthogonal designs with 100 variants. the averaged predicted means and the SEs were estimated.

For the designs repeated 100 times (5-10), Figure 1 shows the distributions of RMSEs based on all 243 health states. The variations in the RMSE values of the random selection designs (6, 7, 9 and 10) were greater than those of the corresponding orthogonal designs (5 and 8). The large orthogonal design offered a stable performance over the 100 variants, the RMSEs of which were all below 0.04.

Figure 1: RMSE distributions for random& orthogonal designs

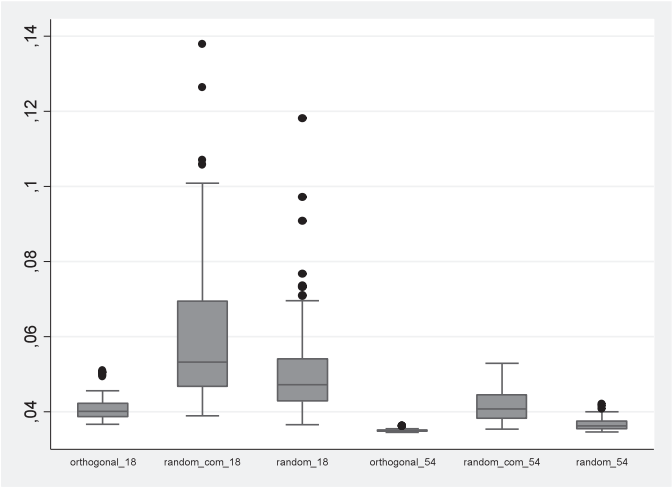
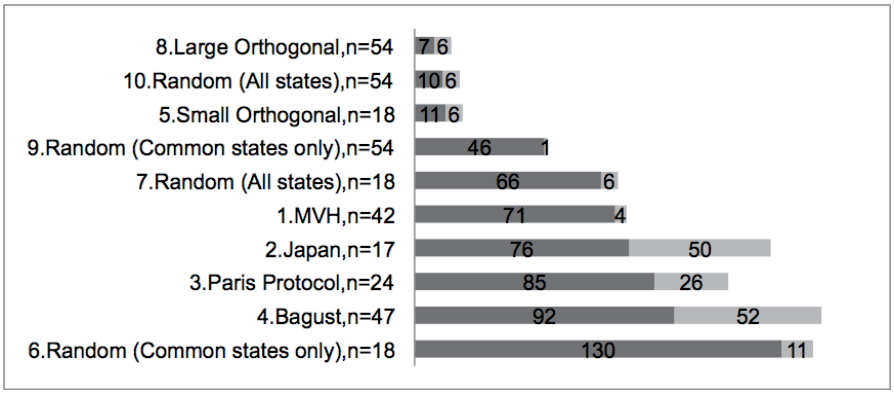


Figure 2 shows the number of health states for which the AEs were larger than 0.05 and 0.10 for each design. Again, orthogonal designs (9 and 10) performed better than the other designs.

Figure 2: Number of health states with AE larger than 0.05 and 0.10



*Dark grey: number of health states with AE >=0.05 & <0.10; Light grey: number of health states with AE >=0.10.

Figure 2 indicates that current designs produced large prediction errors for many health states. In contrast, the orthogonal designs provided considerably fewer large mispredictions.

4.4 DISCUSSION

The saturated data set allowed us to judge the consequences of different design strategies concerning health state selection in valuation studies. To address our research question, the first principle of ensuring design orthogonality clearly outweighed the second principle of the over-representation of common health states in the design. In other words, when weighing up design properties, increased statistical efficiency outweighed an increased error rate - if any - in rare health states. In addition, previously published and commonly used designs performed worst among all the designs examined. The MVH design performed best among the published designs, but was still worse than either orthogonal designs or random selection designs. The 243 health state values were best predicted by the large orthogonal design (with 54 health states), despite the fact that this design contained many rare health states. Large random selection design also performed well, but not when restricted to common health states only. Notably, random selection designs restricted to common health states did not improve the estimation of common health states, but led to misprediction for rare health states. Furthermore, keeping the number of health states in the design constant, the orthogonal designs outperformed the random designs in terms of overall performance and performance stability.

While our investigation used the MVH study as a point of reference, it should be noted that other published valuation studies, including instruments such as HUI (the Health Utility Index) and SF-6D (the Short Form Health Survey), were consensus-based regarding sampling, as in the MVH study, and often used similar criteria (29).

It is noteworthy that published designs provided less accurate predictions overall, and large estimation errors for many health states. The magnitude of the large prediction error exceeded the often-used minimum important difference (MID) in EQ-5D shown in previous studies (79-81). Hence the continued use of the current designs may not be appropriate. Generally, the predictions improve when the large design approach is used in a valuation study. By focusing on the statistical properties required in a study design, we have demonstrated that orthogonal designs may be viable alternatives for selecting health states in future valuation studies. The common health states design did not predict the rare states well. In contrast, orthogonal designs offered the best predictions

universally for common and rare states. These results suggest that orthogonal designs can safely be used, especially when they are also designed to capture level interactions (i.e. the large orthogonal design), but it is the user's decision whether to minimize sample size or to maximize prediction accuracy.

Since most EQ-5D valuation studies have been conducted using the time trade-off (TTO) approach, a limitation of this study is that the designs have been tested and compared using VAS data. Theoretically, VAS values do not have ratio properties and should not be used for estimating QALYs (82). For the purpose of this study, we rescaled the VAS values using two extreme health states and assumed the rescaled values had interval scale properties. It also should be noted that after rescaling the VAS values, the valuation space of our data was more compressed than raw VAS data and TTO data (83, 84). Thus caution is in order when applying our results in valuation studies that use a different valuation methods. We also restricted our analysis to a 'main effects' model only, as in previous studies the use of an interaction term increased the likelihood of mis-specification (69).

The good performance of the large orthogonal design warrants further consideration as this could point to a mis-specification problem in the sense that main effects models perhaps do not capture all effects on valuations. For instance, we could imagine that people's values are subject to interaction effects. Interactions across severity levels might reflect diminishing marginal disutility, or reflect that in some health profiles both causes (say, mobility issues) and consequences (problems in usual activities or self-care) appear. Interactions will affect the value that people express for a given health state and hence model parameters that are identified when values are decomposed to the underlying health characteristics are conditional on unobserved interactions. The large orthogonal designs are balanced in the presence of all possible two factor interactions, and hence their performance was robust even over different variants. Further research is required.

It is an open question what our results imply for EQ-5D-5L valuation. In EQ-5D-5L, a 5-level response scale is used as a replacement for the 3-level response scale, while the descriptive system is the same as for EQ-5D-3L (28). By increasing the response scale levels from 3 to 5, the sensitivity of the instrument was improved and the number of unique health states increased from 243 to 3,125. A big difference between 3L and 5L valuation studies is the ratio between observed and predicted health states, for example, in the MVH design, this ratio was 21% (42: 201), whereas in the current EQ-5D-5L design, it was 3% (86: 3,039). Hence it should be noted that the risk of misprediction would increase since more health state values would rest on extrapolation when using EQ-5D-

5L. Thus, future saturation data should confirm the use of the current EQ-5D-5L design against other possible design choices, especially for orthogonal designs.

In conclusion, there is scope for improvement in health state valuation design strategies. Published design strategies suffer to a large extent from misprediction that can be avoided by promoting statistical efficiency within the design, and by reducing the emphasis placed on over-representing common health states. The orthogonal designs may be an alternative design for future EQ-5D-3L valuation studies

4.5 APPENDIX

Appendix 1. Regression estimates from Japan, MVH, Large orthogonal designs

	Japan design. n=2.139		MVH design. n=5.283		Large orthogonal design. n=6.795	
mo2*	-0.032	-2.5	-0.065	-11.79	-0.054	-11,02
mo3	-0.124	-7.96	-0.151	-21.77	-0.155	-31,43
sc2	-0.068	-5.49	-0.054	-9.15	-0.042	-8,55
sc3	-0.059	-4.02	-0.099	-13.94	-0.085	-17,33
ua2	-0.126	-10.06	-0.082	-12.71	-0.056	-11,3
ua3	-0.236	-15.56	-0.175	-25.75	-0.131	-26,62
pd2	-0.076	-7.45	-0.076	-13.72	-0.071	-14,45
pd3	-0.257	-22.35	-0.201	-34.4	-0.186	-37,65
ad2	-0.056	-4.93	-0.085	-14.59	-0.053	-10,7
ad3	-0.175	-15.54	-0.178	-29.9	-0.146	-29,52
Intercept	0.807	105.96	0.78	128.72	0.695	104,06
R squared	0.663		0.574		0.392	
Root MSE	0.168		0.163		0.166	

*Dummy variable coding: mo2 means 'mobility' at second level etc

CHAPTER 5

How prevalent are implausible EQ-5D-5L health states and how do they affect valuation? A study combining quantitative and qualitative evidence

Zhihao Yang, Zeyun Feng, Jan Busschbach, Elly Stolk, Nan Luo

Submitted for Publication

5.1 INTRODUCTION

In developing an EQ-5D value set, it is common that a subset of health states is directly valued by the general public to provide a basis for predicting values of all health states using regression models. An important decision concerns the selection of health states for direct valuation (15, 69). However, there is no evidence-based guidance on health state selection criteria. One commonly used criterion is the plausibility of health states (67, 68, 85). Health states being 'implausible' refers to those health states that respondents may find unrealistic (86). The inclusion of such states in valuation studies has been assumed to compromise respondents' engagement in the valuation tasks and to increase variations in health states' values (68, 87, 88). For example, in the Measurement and Valuation of Health (MVH) study, the investigators excluded the *prima facie* implausible health states from direct valuation in order to sustain motivation and credibility, and to reduce error (68).

It has been reported that the EQ-5D valuation method was cognitively demanding for respondents as they needed to understand the concept of time trade-off and imagine various 'health states' based on textual descriptions in a short interview (8, 14, 29). Perceived implausibility may have further increased the difficulty in imagining the health states concerned, which is pivotal to the thought process for valuation. According to Karimi et al (6), lay respondents use several steps to value EQ-5D states. First, respondents use their imagination and experiences to give substance to the EQ-5D health states. Then the consequences of these health states are determined by combining with conversion factors (defined as personal and social factors that affect how participants value health states), and are then perceived and weighted to evaluate the health states. In their research, the investigators found respondents encountered difficulties when valuing implausible EQ-5D health states.

Dealing with the implausibility issue is not easy as implausibility is a subjective judgement. For example, health states that are easy to imagine for some individuals may be considered implausible by others. The question thus is to what degree a health state is implausible, rather than whether a health state is implausible or not. So far, concerns about implausible health states have not been formally studied. First, it is not known how plausible or implausible are the health states defined by standardized descriptive systems such as EQ-5D. Second, little is known about the effect of plausibility on health state valuation. While it is reported that implausible health states make valuation tasks more difficult in some qualitative studies, quantitative investigation of how implausibility affects health state values is lacking. What is clear, however, is that imposing a plausibility

constraint on the selection of health states potentially affects the severity spread and the level balance of the selected subset, which may impact on data modelling (15, 87). Whether it is still relevant to account for plausibility in health state selection is unknown.

In this study, we investigated how likely EQ-5D-5L health states were perceived to be implausible, were there common characteristics for implausible states, and how implausibility affected health state valuation. We conducted a large health state valuation study among students, who valued health states and rated the plausibility of each state. The dataset allowed us to observe what health states had the highest chance of being considered plausible or implausible, and to analyze whether values obtained from respondents who considered a state plausible agreed with their counterparts who considered that same state implausible. In addition, we used qualitative interviews to explore directly from the respondents the reasons for implausibility judgements and the effect of implausibility on values.

5.2 METHODS

5.2.1 Data collection

University undergraduate students (N=1,600) at Guizhou Medical University, China were recruited for the study through e-mail and/or personal invitations. The inclusion criteria were aged 18 years or above, full-time students, agreement to complete a questionnaire that may take up to 1 hour, and informed consent. Data collection was conducted in 3 stages. Consenting students were first invited to a classroom to complete a valuation questionnaire in a group (stage 1). Two weeks later, the students were invited back to rate the plausibility of the health states they had valued previously (stage 2). Lastly, a small group of students was invited to participate in a focus group/individual discussion of their experiences in completing the valuation questionnaires, with an emphasis on the valuation of implausible states (stage 3). The detailed methods employed at each stage are elaborated upon below.

Stage 1 – Valuation of EQ-5D-5L health states

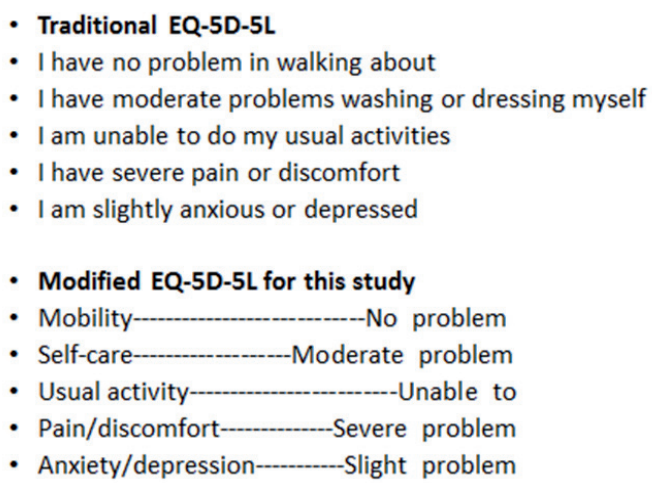
All 3,125 EQ-5D-5L health states were valued using the EuroQol visual analogue scale (EQ-VAS). EQ-VAS is a vertical, 20-cm-long, hash-marked numerical rating scale ranging from 0 (‘the worst health you can imagine’) to 100 (‘the best health you can imagine’). Previous research demonstrated the feasibility of university students valuing 243 EQ-5D-3L health states using EQ-VAS in a single survey session (89). In this research, each student was asked to value 196 or 197 health states in a self-administered, questionnaire-

based survey. We used a stratified random selection procedure to divide all EQ-5D-5L states (except for the best and the worst states, i.e. 11111 and 55555) into 16 blocks of 196 or 197 states. Details of the data collection protocol, which aimed at an equivalent response burden across respondents, are provided in the Appendix.

The valuation questionnaire was administered using paper and pencil. It had three sections. The first section was for respondents to classify their own health on the EQ-5D-5L questionnaire and rate their health using the EQ-VAS. The second section was for respondents first to value the states 11111 and 55555 on one page and then to value a randomly selected block of states presented in separate pages with 10 health states per page. Students were asked to value all states using an EQ-VAS that was presented on a separate piece of paper and were instructed to write down the VAS value for each health state beside the description of the health state.

In order to highlight the differences in the health states for valuation, we presented the EQ-5D-5L health state descriptions by separating the dimensions and severity levels. Figure 1 illustrates this presentation style using the state '13542' as an example (the original questionnaire was in Chinese). In total, we organized 16 group data collection sessions, with each session being attended by around 100 students. In each session, an investigator briefed the students using a standard script before they started to complete their questionnaires and was available to answer any of their questions.

Figure 1: An example of the health states presentation used in this study



Stage 2 – Rating of health states' implausibility

2 weeks after the valuation survey, about half of the total sample were invited back to rate the plausibility of health states they had valued. The survey settings were similar to the valuation survey in the first stage. Students received the questionnaires they had previously valued and were asked to rate each health state using a binary scale by responding 'Y' = 'yes, this state is implausible to me'; 'N' = 'no, I don't think this state is implausible'. In total, 9 sessions of group interviews were organized, each being attended by around 100 students. In the first and second stages, students were paid 100 Chinese Yuan (approximately equivalent to 14 euros) per session.

Stage 3 – Qualitative interviews

A small group of students who participated in both surveys was invited to share their thoughts about implausible health states during in-depth interviews or focus group discussions. The interviews/focus group discussions were conducted 2 weeks after the implausibility rating task. All interviews were conducted and recorded by the same investigator who conducted the stage 1 and 2 surveys. In the interviews/ focus group discussions, interviewees were asked to share their experiences/thoughts about implausible health states.

5.2.2 ANALYSIS

Quantitative data

First, to evaluate the validity of the VAS data, we examined the relationship between the VAS values of all health states and their misery index. The misery index was defined as the sum of the five digits of an EQ-5D-5L health state, e.g., the misery index of 45133=4 +5 +1 +3 +3 =16. It is used as an approximation of severity, with higher values indicating worse health. The misery index ranges from 5 (state 11111) to 25 (state 55555); the number of health states with the same misery index ranges from 1 (for misery index group 5 and 25) to 381 (for misery index group 15). In this analysis, boxplots were produced to compare the distribution of VAS values for health states with different misery indexes. We hypothesized that the mean VAS values monotonically decreased with an increasing misery index.

In order to measure implausibility, we calculated an implausibility score for each health state. The rating of implausibility was coded as '1' for implausible, and '0' for plausible, respectively, for each health state valued by each respondent. By averaging the rate for each health state, an implausibility score was calculated for each health state as

the percentage of respondents who rated that health state as implausible. For example, the implausibility score for state 45133 was 0.71 as 71% of respondents rated it as implausible. The higher the implausibility score, the more implausible a health state. We listed the top ten most implausible states and the least implausible states for reference.

We examined the relationship between implausibility and health severity. First, a total of 21 ., one for each group of health states with the same misery index, were generated and arranged in increasing order of the misery index to observe any trend. Second, we used a regression model to investigate which combinations of extreme levels between dimensions (e.g. 51xxx, 5x1xx) contributed most to the implausibility. Based on the qualitative results, health states with such combinations were more likely to be perceived as implausible in valuation studies. In the model, we created 20 dummies for all the 5-1 level combinations (see formula 1). The analysis was performed at the individual level using logistic regression. We ranked the most influential contrasted pairs from the top downwards.

$$\textit{Implausibility rate} = \alpha + \beta_1 MO_5SC1 + \beta_2 MO_5UA_1 + \beta_3 MO_5PD_1 + \beta_4 MO_5AD_1 + \beta_5 SC_5MO_1 + \beta_6 SC_5UA_1 + \beta_7 SC_5PD_1 + \beta_8 SC_5AD_1 + \beta_9 UA_5MO_1 + \beta_{10} UA_5SC_1 + \beta_{11} UA_5PD_1 + \beta_{12} UA_5AD_1 + \beta_{13} PD_5MO_1 + \beta_{14} PD_5SC_1 + \beta_{15} PD_5UA_1 + \beta_{16} PD_5AD_1 + \beta_{17} AD_5MO_1 + \beta_{18} AD_5SC_1 + \beta_{19} AD_5UA_1 + \beta_{20} AD_5PD_1 + \varepsilon \quad (1)$$

To evaluate the effect of implausibility on VAS values, we calculated the mean VAS values for all states using observations from respondents who valued the states and rated them as plausible (mean VAS value $_{plausible}$), and the mean VAS values for the same states using observations from respondents who valued and rated them as implausible (mean VAS value $_{implausible}$). The agreement of the two sets of mean VAS values was examined using a histogram of the mean VAS values' difference of each state (mean VAS value $_{plausible}$ - mean VAS value $_{implausible}$). We also used a paired t-test to examine the difference between two means.

Qualitative interviews

All interviews and focus groups were audio-recorded. One researcher (ZY) transcribed all interviews and imported all data into Nvivo. Two independent researchers analyzed the qualitative transcriptions using the thematic framework. During the coding, two coders discussed if disagreement occurred. After coding, codes were classified into themes and sub-themes and were further discussed by two coders to finalize the definition and names for each theme.

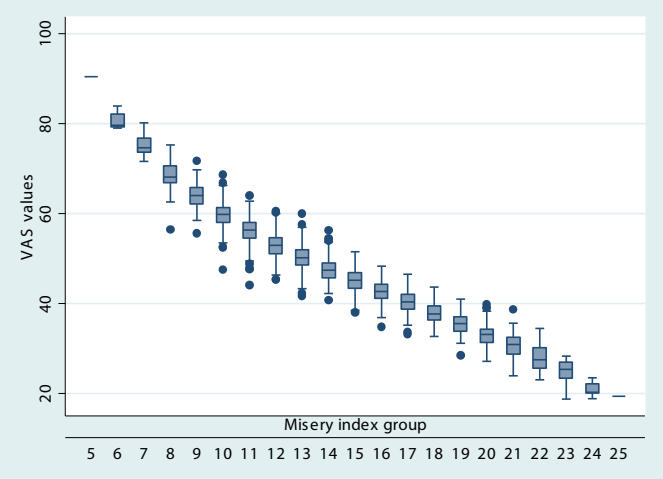
5.3 RESULTS

5.3.1 Participant characteristics

A total of 890 students completed both the valuation and implausibility rating tasks. The average time to complete the valuation task was 65 minutes (range: 23 to 180). On average, the students were around 21 years old and 62% female. All students had a health-related education background, such as pharmacy and public health. Twenty-one students were recruited for qualitative interviews. Nine were interviewed individually and the remaining twelve students attended two focus group discussion sessions with six students in each session.

5.3.2 Distributions of VAS and implausibility scores

Figure 2: Empirical mean VAS values of all health states by misery index



The box plot was sorted on the misery index group; it should be noted that one misery index value could result from more than one health state.

Figure 2 shows the mean VAS values of 3,125 health states plotted with their corresponding misery index group, which ranged between 5 and 25. An outlier was interpreted as the mean value of one health state. On average, the value of 11111 was around 90, and the value of 55555 was around 10, which together represented the range of the values. The health state values decreased along the misery index.

The number of observations for implausibility ratings ranged from 46 to 73 per health state except for "11111" and "55555", both of which had 890 observations. No health state was rated unanimously as "implausible" by all respondents. In contrast, four health states had an implausibility score of 0, indicating which were universally rated as plausible by all respondents. Among all 3,125 health states, 910 (29.1%) health states had an implausibility score over 0.5. The mean implausibility score was 0.386 (SD:0.211). Table 1 lists the top 10 most and least implausible states and their mean VAS values (SD) for reference. The full implausibility score list can be found in the Appendix.

Table 1: Top 10 most implausible and least implausible EQ-5D-5L states.

The least implausible states				The most implausible states			
Health state	Implausibility score	Mean VAS value	SD	Health state	Implausibility score	Mean VAS value	SD
33334	0.000	39.07	20.54	55111	0.932	52.90	17.17
32322	0.000	55.46	17.18	44151	0.930	44.95	17.19
33333	0.000	48.06	14.07	55121	0.927	48.62	14.43
32233	0.000	50.06	19.08	54151	0.926	44.51	14.89
22322	0.000	60.00	14.56	55151	0.924	41.15	16.63
23333	0.017	53.02	15.50	54114	0.907	42.00	14.84
21333	0.017	53.05	17.05	35114	0.907	48.49	15.31
12222	0.019	64.85	17.36	55141	0.889	43.26	15.34
54443	0.019	33.74	15.96	55152	0.889	35.50	13.06
55545	0.020	17.92	14.01	45121	0.885	53.98	14.85

Figure 3: The implausible score of health states over misery index

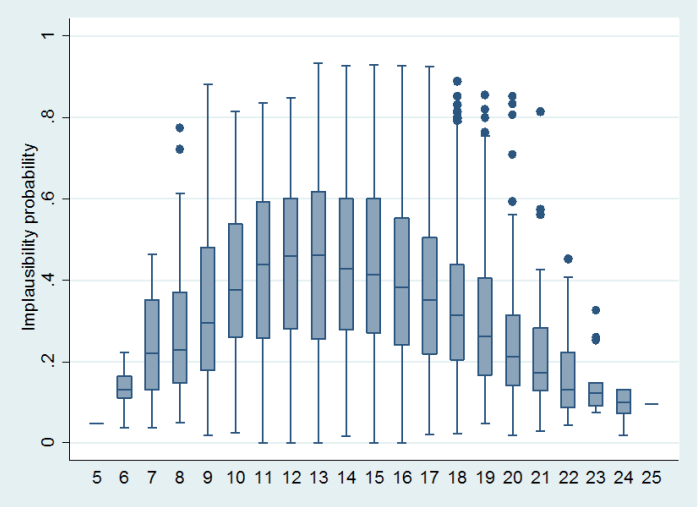


Figure 3 shows the distributions of the implausibility scores by misery index using box plots. It can be observed that moderately impaired health states defined by the misery index were more likely to be rated as implausible.

Table 2: Odds ratio of 5-1 dimension combinations on implausibility

Dimensions combinations	Odds ratio	Standard error	Z	P-value
moua	3,525	0,108	40,960	0,000
scua	3,358	0,102	40,010	0,000
pdua	2,631	0,077	33,030	0,000
adua	2,319	0,067	28,960	0,000
uamo	2,305	0,063	30,390	0,000
uasc	2,205	0,061	28,750	0,000
moad	1,798	0,051	20,840	0,000
uaad	1,763	0,049	20,450	0,000
scad	1,637	0,046	17,590	0,000
pdad	1,555	0,043	15,960	0,000
mopd	1,456	0,041	13,390	0,000
uapd	1,450	0,040	13,390	0,000
mosc	1,371	0,039	11,220	0,000
scpd	1,308	0,037	9,580	0,000
pdsc	1,184	0,033	6,100	0,000
pdmo	1,171	0,033	5,650	0,000
scmo	1,169	0,033	5,540	0,000
admo	1,050	0,030	1,750	0,080
adpd	1,047	0,030	1,610	0,107
adsc	1,047	0,029	1,620	0,105

*all 5-1 levels combinations, in each dummy, the former was on 5th level, the latter was 1st level, e.g. 'moua' is a health state with pattern 5X1XX.

Table 2 shows the ranking of 20 pairs of extreme/no problem combinations in causing implausibility. Some patterns can be spotted, for example, usual activities was the most prominent dimension, i.e. extreme problems in other dimensions would always cause problems in usual activities. In contrast, the last three 'not significant' combinations suggested that extreme anxiety/depression would not cause a problem in any other dimension except for usual activities, but not the other way around: extreme problems in other dimensions would cause anxiety/depression.

Figure 4: Histogram of the mean VAS difference (mean VAS value plausible - mean VAS value implausible)

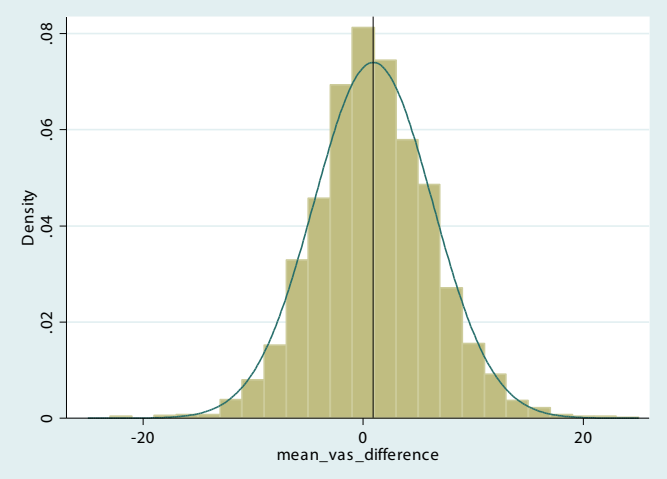


Figure 4 shows the distribution of the mean VAS difference (mean VAS value $_{plausible}$ - mean VAS value $_{implausible}$). Four states were not included in this analysis because of nil implausibility ratings. In general, the distribution was almost symmetrical around 0 with more states having higher mean VAS values based on plausible observations. The paired t-test suggested that the mean VAS value based on plausible observations was 0.911 higher than the mean VAS value based on implausible observations (P=0).

5.3.3 Thematic analysis results

Most of the respondents reported that they noticed some 'strange' states and the values they gave were vague. Four broad themes came up during the thematic analysis regarding implausible health states: 1) reasons for states being rated as implausible, 2) difficulties in valuing implausible health states, 3) strategies for valuing implausible health states, and 4) values of implausible health states.

Reasons for states being rated as implausible

Different respondents interpreted EQ-5D health states differently and had different judgements about the plausibility of health states. For example, some respondents reported that '11111' & '55555' were implausible as they believed that one's health cannot be perfectly good or extremely bad. More often, respondents reported implausible health state as having 'logical conflict between dimensions'. Within the latter idea, the interpretation of relationships between dimensions varied across respondents.

(1) Mutual inclusiveness: some dimensions were deemed to cover other dimensions. For example, many respondents stated that usual activities included mobility or self-care, so if experiencing mobility or self-care problems, one's usual activities could not be without any problem.
(2) Causality: a problem in one dimension would certainly cause a problem in other dimensions. Some respondents followed the sequence of the five dimensions and interpreted the sequence as a causality, e.g. if the first four dimensions had no problems, then the health state should not have any anxiety, as it was believed there was no reason to be anxious or depressed if one was healthy in the first four dimensions.
(3) Close relatedness: some dimensions were deemed to be closely related (not necessarily having the causal and inclusion relationships), so if someone had problems in one dimension, they would also experience some problems in a related dimension. For example, usual activities and self-care were perceived to be two closely related dimensions as they both required individuals to use their limbs.

Difficulties in valuing implausible health states

When asked what kinds of problem respondents encountered when valuing implausible health states, they replied that they were:

(1) Reluctant to put more effort in imagining a health state that they could not think of / perceive. Once a respondent deemed a state to be implausible, she or he became reluctant to value it 'properly'. Other respondents claimed that they simply could not imagine implausible health states.
(2) Unable to foresee the consequences of being in that health state. Respondents reported that after perusing the health states, they could not think of what the impact of such health states on life would actually be.

Strategies for valuing implausible health states

The strategies respondents employed to value implausible health states included:

(1) To focus on the severe dimension: respondents focused on valuing the dimension with the most severe problem. Respondents reported that they normally valued one dimension at a time, starting from the most important dimension, and then perceived all dimensions together as a health state. When two dimensions conflicted, they focused more on the more severe dimension and valued that dimension instead of the whole health state.
(2) Reference to a similar health state: when valuing an implausible health state, they referred to the value of a similar health state and made some adjustment to that value.

(3) Rationalize the health state: respondents re-wrote the severity levels of some dimensions to render an implausible health state plausible. For example, if respondents noticed that the first four dimensions did not have many problems, but the last had rather severe problems, they would assume that the first four dimensions could be affected by the last dimension and then value the health state.

Values of implausible health states

The valuation task required respondents to attach a value to a health state. Since implausible health states imposed more difficulty during the thought process, we asked respondents how they felt about the values they gave to implausible health states.

(1) 'Vague'/ 'less reliable'/ 'just about right': respondents suggested that the values they gave to implausible health states were less precise compared to the values they gave to plausible health states.
(2) Low: most respondents stated that they felt that the values they gave to implausible health states were lower than they should have given.

5.4 DISCUSSION

This is the first study to investigate the effect of implausible health states on respondents' values. Unlike previous studies searching for evidence from existing datasets (67, 88), this study took its point of view from that of its respondents. For convenience reasons, we used the term 'implausible health states' to refer to those states with high implausibility scores judged by our student panel. It should, however, be emphasized from our results that there was strong heterogeneity concerning the judgement of implausibility, and there was not one health state that could be deemed totally implausible.

In summary, around 30% of the 3,125 EQ-5D-5L health states were considered as implausible by at least half of the sample (i.e. an implausibility score >0.5). From both quantitative and qualitative analysis, we observed that the major reason for implausibility was due to perceived conflicts between two dimensions. Thus, the moderate health states on the misery index scale were more likely to be judged as implausible states, as health states in this range could have some extreme/no problems combinations between dimensions. By analyzing the extreme/no problems combinations, a common pattern could be found, such as usual activities being more likely to interact with other dimensions, for example. Nevertheless, from the thematic analysis, respondents had different interpretations concerning how EQ-5D dimensions interacted with each other.

We found the mean VAS values from plausible observations to be slightly higher than the counterparts from implausible observations, but the difference did not vary along the severity scale nor among different implausibility scores. By reviewing the strategies used for valuing implausible states, two strategies could be linked with the lower scores, i.e. focusing on the severe dimension and rewriting the health state. Using either strategy, respondents tended to pay more attention to the severe part of a health state.

Using a saturated VAS dataset, Yang et al confirmed that health state selection strategy focusing on using common states for valuation led to large mispredictions for other states (15). Instead, a statistically efficient design (i.e. the subset of health states used for direct valuation) could produce more accurate predictions for non-valued health states. As a statistically efficient design often includes implausible states, then the trade-off is between a statistically efficient design with some implausible states versus a plausible states-only design with limited statistical efficiency. If the purpose of a valuation study is to provide as accurate as possible predictions for all defined health states, then we would prioritize the concern of statistical efficiency over the concern of implausible states. As the output of design generators like N-gene could permit any transformation of the basic permutation scheme from, for example, an orthogonal design, researchers could opt for a variant with the most plausible states.

There are several limitations of this study. First, instead of the general population, our respondents were university students. Second, each respondent was asked to value around 200 health states using EQ-VAS. The number of health states valued was much higher than in the standard valuation task using composite TTO. Last, as all respondents were interviewed several weeks after the valuation task and the implausibility rating task, there could be a recall bias and interviewees may mix the thoughts they had during the valuation task and implausibility rating task.

5.5 CONCLUSION

To conclude, health states with a logical conflict between dimensions were more likely to be judged as implausible states. Health states considered as implausible were more difficult to value accurately and the values from respondents who deemed certain states implausible tended to be lower than those from respondents who thought they were plausible. Evidently, respondents interpreted health states differently, and showed large heterogeneity with respect to views about implausible states.

5.6 APPENDIX

The process of assigning EQ-5D-5L states to blocks was as follows: we computed the 'misery index' of all EQ-5D-5L states, as the sum of the five digits of a health state, e.g. the misery index for state '12345' is 15. The misery index defines 19 strata (sum score ranges from 5 to 25) and the number of health states in each stratum differs: e.g. in the stratum with misery index 6, there are only 5 health states, 21111, 12111, 11211, 11121, 11112, while in others there might be many more. From each stratum, we randomly selected health states proportionally to the total number of health states in that stratum. Hence each block contained similar numbers of health states from each stratum, ensuring general severity balance across blocks.

CHAPTER 6

The effect of health state sampling methods on model predictions of EQ-5D-5L values: small designs can suffice

Zhihao Yang, Nan Luo, Gouke Bonsel, Jan Busschbach, Elly Stolk

Accepted in Value in Health. In Press.

6.1 INTRODUCTION

The EQ-5D is a health-related quality of life (HRQoL) questionnaire widely used in health economic, clinical, and population health studies. EQ-5D has two validated versions, which both compromise five dimensions: mobility, self-care, usual activities, pain/discomfort and anxiety/depression, the three-level EQ-5D (EQ-5D-3L) describes each dimension at three levels (roughly corresponding to no, moderate, extreme problems) and the five-level EQ-5D (EQ-5D-5L) expands its descriptions to five levels (roughly corresponding to no, slight, moderate, severe, extreme problems) (9). Compared to EQ-5D-3L, the descriptive richness of the EQ-5D-5L is an advantage when the goal is to understand the health state of a respondent, but potentially complicates the development of value sets. Through a valuation study, all health state 'values' (some prefer 'utilities' or 'index values') can be derived from the corresponding value set. These 'values' indicate how desirable the health states are. Performing such a valuation study for EQ-5D-5L is a challenge in terms of the trade-off between feasibility and validity. EQ-5D-5L defines 3,125 states, which ideally should all be valued, but that is infeasible under standard conditions. Hence, in practice, only a subset of the health states is directly valued, and from this subset the values of all health states can be estimated through statistical modelling. Value sets for SF-6D and EQ-5D-3L have also been developed using this statistical modelling approach (29).

Selecting the subset of health states for direct valuation ('the empirical state set') is an important design matter for valuation studies and it is still evolving. For EQ-5D-3L, the Measurement and Valuation of Health (MVH) study protocol containing an empirical set of 42 EQ-5D-3L health states is most widely used (60). Without applying explicit statistical considerations, the MVH study oversampled mild and commonly seen health states (15). For EQ-5D-5L, the current valuation protocol was built on the results of several iterative pilot studies (28, 29, 53). It was decided that the number of states in the design should be somewhere between 80 and 100, as the EQ-5D-5L main effects model has 21 parameters (5 health dimensions x 4 dummy variables for severity levels + intercept). By ensuring that the total number of health states was four times larger than the number of parameters in the main effects model, multi-level modelling could be applied, i.e. a random coefficient model to account for the effects of individual background variables (53). Next, the number of health states to be valued by a single respondent was maximized at 10. To arrive at around 80 health states, a blocked design (10 blocks, each block with 10 states) was utilized, employing a balanced selection of states with respect to their utility values. Hence each block was planned to include the pits state, i.e. the most severe health state: 55555 and 1 of the 5 very mild states: 21111, 12111, 11211, 11121,

11112. This left 8 unique states per block, in total 80 health states, to be defined; these were randomly selected out of the remaining 3,118 health states. The selection of the 80 health states for the protocol was based on Monte Carlo simulations to predict the prior values obtained from the multi-national pilot study, instead of choosing predominantly mild states (28). The 'optimal' set of 80 states was selected on the mean squared error (MSE) between the prior parameters and estimated parameters from a 'main effects' model, and level balance, but without making orthogonality an explicit criterion (53).

Using an EQ-5D-3L VAS saturated dataset (a dataset where the values of all 243 health states are known), two studies investigated the effect of health-state selection on prediction adequacy (15, 69). Both studies found that by improving the statistical efficiency of the design, the number of health states in the empirical state set in a valuation study could be reduced without loss of precision or validity (69, 90). In particular, the orthogonal design appeared ideal as it possessed two statistical properties: level balance and orthogonality (i.e. level pair balance) (90). As the EQ-5D-5L empirical state set of 86 states (also known as the 'EQ-VT set') was selected without constraints concerning orthogonality, the design choice of EQ-5D-5L may have suffered from misprediction effects, as found in some design choices of EQ-5D-3L (90).

Furthermore, while larger designs may be favoured, given the advantages that they offer in the context of model exploration, we note that published EQ-5D-5L value sets have never used models with more than 22 parameters, leaving a surplus of 64 degrees of freedom (19). This indicates that there could be redundancy in the current design, but we must proceed with caution when we aim to investigate this. In EQ-5D-3L, we have seen that a reduced design with 17 states from the original MVH design (42 states) introduced large prediction errors in the final value set (90). Nevertheless, utilizing a small design could reduce the cost of a valuation study and increase the feasibility of such a study for countries with limited resources. Hence, for any given degree of prediction accuracy, the smallest design with the least number of health states to be directly valued is sought, so that the cost of a valuation study can be minimized.

In this paper, we revisit the EQ-VT design through two research questions.

(1) Is there a more efficient (thus less costly) empirical set of health states than the current 86 EQ-VT set to derive an equally valid EQ-5D-5L value set?
(2) Since 86 states in the EQ-VT design were divided into 10 blocks, and the pits state and 5 mild states were over-sampled given they were in all the blocks, what was the impact on prediction performance of oversampling these particular states in

the current EQ-5D-5L design?

To address these questions, we collected values for all 3,125 EQ-5D-5L health states in a dedicated direct EQ-VAS valuation study. This saturated VAS dataset enabled us to compare the prediction performance of the 86 health states subset with any alternative subset of health states. VAS was used in this research for its simplicity and VAS values served as proxies for TTO values. Even though TTO (a trade-off exercise involving duration) and VAS (a direct scaling exercise) are two different tasks (91), they are both used to elicit cardinal preference data on the same object. Moreover, from past experience, we know that a VAS data set can be close to its TTO counterpart (82, 83, 92, 93). Nevertheless, the results of this research should be seen in the light of the assumption that the selection artefacts are independent of the valuation methods employed.

6.2 METHODS

6.2.1 Protocol to collect the saturated VAS dataset

The current EQ-VT protocol requires each health state to have at least 100 observations so that the estimate of the (mean) value of each health state is sufficiently precise (53). Adopting this sample size requirement, we obtained a saturated dataset by inviting 1600 university students as respondents, each of whom provided VAS values for approximately 197 health states: (1600 students x 197 health states/respondent)/ 3125 states = 100 observations/health state. We divided 3,123 health states into 16 blocks using a stratified random selection process so that each block contained around 197 states (11111 and 55555 were presented in all blocks). For details of the data collection protocol, which aimed at an equivalent response burden across respondents, see the Appendix.

We organized 16 sessions of group interviews. Around 100 students were recruited to participate in each session, and each student received a randomly chosen block of health states. Each student received 100 RMB (equivalent of €15) as an incentive payment.

6.2.2 Tested Designs

After we obtained the empirical values for all 3,125 health states, we tested how well the EQ-VT set with 86 health states and other candidate health state sets predicted the values for all 3,125 EQ-5D-5L health states. In short: subsamples of the dataset were drawn to mimic the data obtained using a particular design, then a model was applied to estimate all 3,125 health states, and finally these predictions were compared with the empirical values.

Using the EQ-VT 86 states set as a reference selection, we investigated the performance of orthogonal, random or D-efficient designs of different sizes (number of health states in the subset). We started the size selection at 25 health states, as this was the smallest size for orthogonal design in a five-factor five-level classification system (main effects modelling only). For each design, size selections of 25, 50, 75, 100 and 200 health states were created. For each design of a different size, 100 variants were produced.

Both the orthogonal design and D-efficient design are standard design choices in conjoint analysis and both designs aim to optimize statistical efficiency (94). An orthogonal design defines an empirical state set, which satisfies the criterion that all severity levels and all severity level combinations (to a defined degree of level interaction: 2e or 3e etc.) are equally prevalent and therefore balanced (78). An orthogonal design is not always available as some combinations of dimension levels are not feasible (in the case of EQ-5D, the combination of 'unable to walk' with 'no problems in usual activities' appears to conflict). Alternatively, D-efficient design can be used. A D-efficient design aims at minimizing the geometric mean of the eigenvalues given |(X'X)-1|1-p (94) from the empirical state set, taking into account level balance. Hence, a D-efficient design is efficient as the matrix of the vector of parameter estimates in a least squares analysis is proportional to |(X'X)-1|1-p, which is minimized (94). In our study, orthogonal designs were provided by N-gene (36) and D-efficient designs generated through Stata 14.0 by selecting the 100 most D-efficient designs from 5,000 random candidates. The Stata code can be found in the supplementary materials. For comparison, we created a series of random designs, imposing the restriction that the design should be severity balanced. For this purpose we first computed the 'misery index', that is the sum score of the digits that represent the EQ-5D health states: 54321 = 5+4+3+2+1 = 15. We then classified all 3,125 states into five misery index groups (<=10, 11-13, 14-16, 17-19, >=20) and randomly selected health states from each group. Hence across empirical sets, balance was present in terms of the number of health states in each of the five 'misery strata'. It should be noted that there are also other designs, e.g. Bayesian, which take both prior information and statistical efficiency into consideration.

6.2.3 Analysis

First, to obtain some insight into the data, we described the saturated dataset by plotting the relation between the mean VAS values of all health states to their misery index scores and showed the distribution of all observations along the VAS scale.

The performance of the different principles in selecting health states was quantified through computation of the Root Mean Squared Error (RMSE) as the primary measure

of prediction performance (the higher, the worse the performance). For each design, an ordinary least squares (OLS) main effects model was used to fit the model for the empirical data of that particular design. In this paper, we fitted the model using individual-level data (100 raw VAS observations per state) (68-70). In the main effects model, the VAS value of a health state was explained by 20 dummy variables and one intercept. For each dimension (MO for mobility, SC for self-care, UA for usual activity, PD for pain/discomfort, AD for anxiety/depression), four dummy variables were used to represent the deviation from level 1 to the other 4 levels, e.g. MO_3 takes 1 if the health state has a problem in the third level of mobility, and takes 0 if otherwise (68).

$$\text{VAS value} = \alpha + \beta_1 MO_2 + \beta_2 MO_3 + \beta_3 MO_4 + \beta_4 MO_5 + \beta_5 SC_2 + \beta_6 SC_3 + \beta_7 SC_4 + \beta_8 SC_5 + \beta_9 UA_2 + \beta_{10} UA_3 + \beta_{11} UA_4 + \beta_{12} UA_5 + \beta_{13} PD_2 + \beta_{14} PD_3 + \beta_{15} PD_4 + \beta_{16} PD_5 + \beta_{17} AD_2 + \beta_{18} AD_3 + \beta_{19} AD_4 + \beta_{20} AD_5 + \varepsilon. \quad (1)$$

In the modelling, 100 observations per state were used across all design choices. This meant all data was used, except for 11111 and 55555, because these states were sampled in every block of the questionnaire. To avoid 'over-weighting' 11111 and 55555, the number of observations was limited to 100.

To answer our first research question, we summarized the RMSE of all designs (orthogonal, random, and D-efficient, all with different sizes and 100 variants), using a boxplot to combine the results of the simulations per specific design. The RMSE of the EQ-VT design was added in the boxplot as a reference. We defined as the most efficient design that which systematically achieved the lowest RMSE relative to sample size. The most efficient design was reported in detail, with further descriptive tables including comparisons with the EQ-VT design.

To test our second research question, we fitted the model using weighted OLS regression: 2 times for the 5 mildest states and 10 times for the pits state 55555, as undertaken in the EQ-VT protocol. The comparison was made with the EQ-VT design with an equal 100 observations for all 86 states. Similarly, we examined how adding the 5 mildest states and the pits state in the most efficient design identified from the above analysis would impact upon the misprediction. In the detailed comparison of the most efficient designs, we reported on the RMSE separately for the empirical state set only, on the validation state set only, and for all 3,125 states combined. We also considered whether prediction error depended on health state severity. For this purpose, we categorized the values into 10 groups along the VAS scale: <30, >=30 & <35, >=35 & <40, >=40 & <45, >=45 & <50, >=50 & <55, >=55 & <60, >=60 & <65, >=65 & <70, >=70. Finally, we estimated

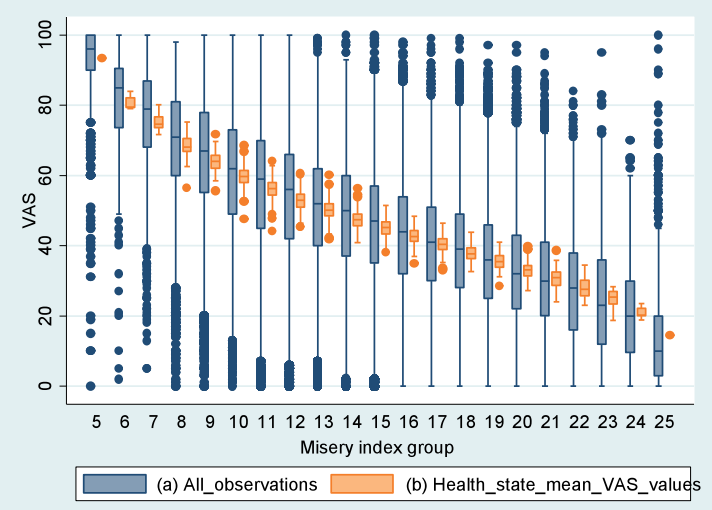
the number of health states with large prediction errors, defined by the Absolute Error (AE): AE>5 & AE>10. For reference, we listed the following for 10 random health states: observed mean VAS values, standard error, 95% confidence interval, and predicted VAS value.

6.3 RESULTS

6.3.1 Description of the saturated dataset

In total, 1,603 students participated in the study and finished the valuation task. This resulted in 100 observations for all states except 11111 and 55555, which each had 1600 observations.

Figure 1: Empirical VAS values and mean VAS values of all health states by misery index



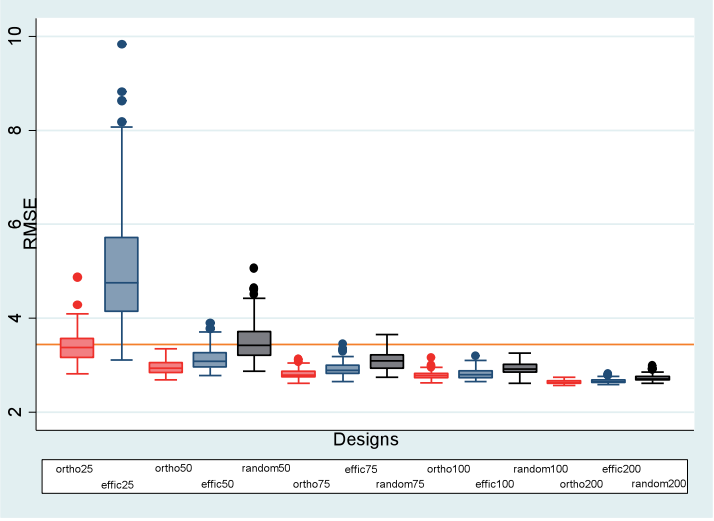
The box plot was sorted on the misery index group, it should be noted that one misery index value could result from more than one health state. The 'All_observations' is based on all VAS observations; the 'Health_state_mean_VAS_values' is based on the mean VAS values of health states.

The misery index for the EQ-5D-5L ranged from 5 (state 11111) to 25 (state 55555), but the number of different health states with the same misery index ranged from 1 to 381. In Figure 1, for each misery index score (5-25) the following were plotted separately: (a: All_observations) its relationship with all VAS value observations of a given misery index (blue boxplot), and (b: Health_state_mean_VAS_values) its relationship with the mean

VAS value per health state with that misery index (orange boxplot). For comparison, we put (a) and (b) side by side in Figure 1. In (a) an outlier was interpreted as one observation, in (b) an outlier was interpreted as the average value of one health state. On average, the value of 11111 was around 90, and the value of 55555 was around 10, which together represented the range of the values. The health state values decreased along the misery index, as expected. Detailed descriptions concerning the quality of the saturated dataset can be found in the Appendix.

6.3.2 Comparison of design performance

Figure 2: Boxplot showing the variations of different designs' Root Mean Squared Error (RMSE) of the predictions for 3,125 health states



The reference line is EQ-VT design with RMSE=3.44. The random design with 25 states was excluded due to a large RMSE.

We summarize design performance in Figure 2 using boxplots. The box plots show the median RMSE for designs of each type and the variance observed across the 100 variants. The reference line represents the EQ-VT design. This design has no variance, as it is fixed by protocol. The EQ-VT design performed well with the RMSE = 3.44, but we can also note that all the other designs of the same size performed even better. When the sample size was limited below 50, the orthogonal design performed better than the D-efficient and random designs. In designs of a size = 75, the orthogonal and D-efficient designs performed similarly, while the random selection design was slightly worse. Random

selection designs had many more variations. Noticeably, the small orthogonal design with 25 states on average performed about as well as the EQ-VT design. Other designs of size 25 performed poorly. The random designs with 25 states were not plotted in Figure 4 as their RMSE =7.65 were beyond the range of the Y-axis. The D-efficient design with 25 health states performed the worst among all plotted designs.

When inspecting the outlier variants in the orthogonal design, we noticed that the outliers were mainly due to the inclusion of state 11111. Given the favourable outcomes for the small orthogonal design, in the following analysis we compared this in detail with the standard EQ-VT design.

Table 1: Root Mean Squared Error (RMSE) by empirical/validation state set for EQ-VT design and 25 orthogonal design

	No. of states	Empirical state set	Validation state set	All 3,125 states
EQ-VT Protocol (weighted for pits & 5 mildest)	86	2.69	3.69	3.66
EQ-VT Protocol	86	2.65	3.45	3.44
EQ-VT Protocol (excluding pits & 5 mildest)	80	2.39	3.02	3.00
25 orthogonals	*25*	*1.03*	*3.41*	*3.40*
25 orthogonals (extending pits & 5 mildest)	*31*	*2.61*	*3.88*	*3.87*

*the italic design was repeated over 100 times.

In Table 1, we report the RMSEs for the empirical health state set, the validation health state set, and all health states taken together, for the small orthogonal design and the EQ-VT design, and the variants of both designs in adding/weighting/removing 5 mild states and the pits state. Excluding the 5 mildest states and the pits state in EQ-VT, or restricting the design to an orthogonal design only, improved the overall RMSE. Furthermore, over-representing the 5 mildest states and the pits state following the current EQ-VT protocol increased the overall RMSE.

Figure 3: Root Mean Squared Error (RMSE) over VAS values for EQ-VT design and 25 orthogonal design

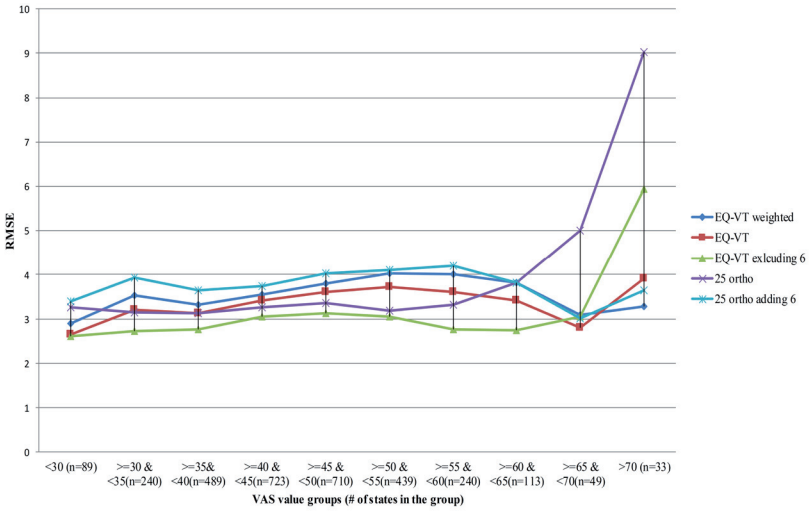
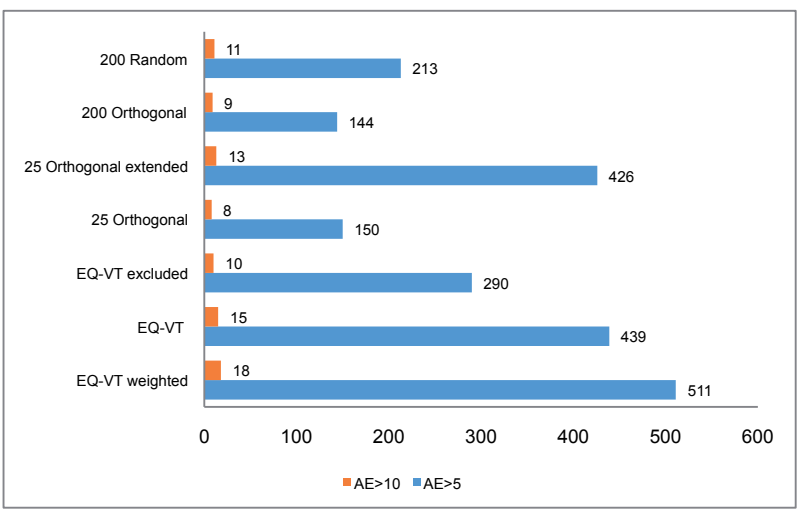


Figure 3 shows that the EQ-VT set predicted evenly along the scale when the five mildest states were included. In contrast, removal of the 5 mildest states from the EQ-VT set and/or restriction to only a small orthogonal design, improved the fit for severe states but increased mispredictions for mild states.

Figure 4: Count of large misprediction errors (Absolute Error > 5, AE > 10)



* random design and orthogonal design with 200 states were added for reference

Figure 4 shows that large mispredictions occurred least frequently in the orthogonal designs, regardless of size.

Table 2: Observed values & predicted values for 10 random health states

Health state	Observed value	SE	95% CI	Predicted by orthogonal* (mean, SE)	Predicted by EQ-VT
21112	73.7	1.5	70.7. 76.6	67.2. 3.1	71.7
13112	68.0	1.6	64.9. 71.1	65.3. 2.9	70.5
23113	61.3	1.5	58.2. 64.3	58.8. 2.6	60.4
25511	51.9	1.8	48.3. 55.4	50.2. 2.5	53.2
14334	43.0	1.8	39.6. 46.5	45.0. 2.1	45.0
44513	41.0	1.5	38.1. 43.9	40.9. 2.3	42.3
13455	35.5	1.7	32.2. 38.8	37.6. 2.0	37.1
24445	35.0	1.6	31.9. 38.0	33.0. 2.0	33.3
45354	30.5	1.5	27.6. 33.3	28.9. 2.1	26.1
55555	14.5	0.4	13.8. 15.2	20.2. 2.3	18.2

CI, confidence interval; SE, standard error.
*for the large orthogonal designs with 100 variants, the averaged predicted means and the SEs were estimated.

Table 2 lists the observed VAS values and predicted VAS values of a random set of 10 health states with different severity levels.

6.4 DISCUSSION

We obtained a saturated dataset that allowed for head-to-head comparison of different principles in the selection of health states in valuation studies. We found that the EQ-VT design performed well in terms of misprediction effects measured by the overall RMSE. In addition, we observed that designs with fewer states can perform as well as the EQ-VT design if they are constructed with attention to their statistical properties. The orthogonal design with 25 states performed closely to the standard EQ-VT with 86 states in terms of overall RMSE. Importantly, values generated on the basis of a small orthogonal design with 25 states contained fewer large mispredictions (defined by AE>5 & AE>10) than the values generated on the basis of the EQ-VT. Both designs provided sufficient prediction accuracy, which was below the oft-used minimum important difference (MID) (80, 95). To answer our first research question, the small orthogonal design with 25 states was the most efficient design we identified.

A caveat to the use of the small orthogonal design lies in the large mispredictions in the mild states (VAS value>70) compared to EQ-VT. There are several possible explanations here. First, this could be a consequence of under-representing the mild states in

orthogonal designs compared to the EQ-VT design (note that in a small orthogonal design with 25 states, only 1 or no health state is mild). Thus, to address our second research question, by giving the 5 mildest states more weight in the blocked EQ-VT design or by extending a small orthogonal design with the 5 mildest states, the predictions for mild states improved, at the price of increased mispredictions for the moderate/severe states. Second, we did not take account of the consideration that the mean values for mild states could be seen as censored at 1 (96), and that the main effects model did not capture all effects on valuations (17).

Moreover, the models that we used could introduce further bias, as they do not consider the possible heteroskedastic nature of the data (96, 97), i.e. severe states have more variance than the mild states. It is possible that these issues also affect VAS data differently than they affect TTO data, because VAS data are characterized by relatively low values for mild states, translating into a large intercept. Hence, while awaiting better understanding and modelling of the upper part of the scale in general, consideration could be given to the use of small orthogonal designs extended with the five mildest states if the resulting values are predominantly used for the 'better' half of the health states.

Another important finding was that the performance of the orthogonal designs depended on inclusion of state 11111. Due to the non-additivity of domains in the upper part of the scale, a gap usually exists between 11111 and all other states. In this saturated dataset, the value of 11111 was 90.48, and the next highest value was 83.93 for 11121. Thus, the value of the state 11111 could not be derived from the value impacts of level 1 of the 5 dimensions in non-11111 states, and conversely, the impact of level 1 in general (in non-11111 states) would be mispredicted if it primarily relied on the empirical value of 11111. As the output of design generators like N-gene could permit any translation of the basic permutation scheme, researchers could opt for a variant without state 11111. Additionally, this upper gap issue (11111 effects) of a VAS exercise may have disappeared in TTO data as 11111 is the reference state (no need to value and have a theoretical value of 1) and the gap effect is then translated into the model intercept.

Better performance for statistically efficient designs was similar to the results found in previous EQ-5D-3L studies (60, 69, 90). While we conclude that in using a main effects model, an orthogonal design is stable and efficient, the D-efficient design is a good alternative when an unrestricted orthogonal design is deemed inapplicable. Theoretically, the more the D-efficient design achieves level balance and orthogonality, the more efficient it is (94). Hence, compared to the orthogonal design, which already

optimizes statistical efficiency, D-efficient design may need more states (to compensate for the loss in efficiency) in order to achieve the same prediction accuracy. Similar to EQ-5D, the valuation of other HRQoL instruments such as SF-6D and HUI may also benefit from using statistically efficient designs. Further research is required.

Some general limitations apply. First, we used a saturated VAS dataset to mimic the design choices in EQ-5D-5L valuation studies which use TTO as their elicitation method. Raw VAS values do not have ratio-properties. If we assume a (monotonic) linear relation between VAS and TTO (82, 93), then we would also expect our conclusions to be valid for TTO. Nevertheless, it should be noted that TTO data display more heteroscedasticity between states and more heterogeneity between respondents, and thus we may expect to use more states or observations in a TTO valuation study. Second, there may be a blocking effect as we divided all 3,125 states into 16 blocks when collecting the saturated dataset. While this essentially suggests a two-level analysis, we assumed there was no such effect. Third, we used university students as respondents, who have limited experience in health problems and whose preferences may be more homogeneous. This may have led to smaller RMSE compared to studies using the general public as respondents, but this is a minor issue as the purpose of this study was to test hypotheses rather than to generate value sets.

Our results inspire faith in the design of the EQ-VT for current EQ-5D-5L valuation studies (29, 53). We noted that small orthogonal designs with 25 states performed almost as well as other designs but produced biased estimates for mild states. Further research with respect to this phenomenon, and strategies to avoid it, are warranted because of the potential benefits that can be reaped from adopting small designs. That is, employing a small orthogonal design with 25 states (or 31, if extended to add the 5 mildest states and the pits state) could reduce sample size requirements by over 50%. Future research should also investigate the validity of orthogonal designs utilizing TTO data.

6.5 APPENDIX

6.5.1 Appendix 1

We divided 3,123 health states into 16 blocks using a stratified random selection process so that each block contained around 197 states (11111 and 55555 were presented in all blocks). The stratification was based on health states' 'misery index': the sum of the five digits of a health state, e.g. the misery index for state '12345' is 15. The misery index defines 19 strata (sum score ranges from 5 to 25) and the number of health states in

each stratum differs: e.g. in stratum with misery index 6, there are only 5 health states, 21111, 12111, 11211, 11121, 11112, while in others there might be many more. From each stratum, we randomly selected health states proportionally to the total number of health states in that stratum. Hence each block contained similar numbers of health states from each stratum, ensuring general severity balance across blocks. Using this design, each block has 196 or 197 health states.

For the purpose of acquiring the saturated dataset we condensed the health state description. To achieve this, we pre-tested three different presentations of EQ-5D-5L health states in a group of 10 students, each of those students valued 24 health states in

a. 8 states with normal presentation (e.g. I have no problem walking about);
b. 8 states with bold font for the severity level (e.g. I have **no problem** walking about);
c. 8 states for separating the dimension and severity level (Walking about---------No problem).

Each student saw the three different presentations in a random order. The three presentation styles were selected by several EuroQol scientists. After that, we asked each student to vote which presentation style he/she preferred and not preferred, and stated his/her reasons. The current version (c) received most votes, 6 out of 8 students preferred presentation style. When we introduced the idea of valuing around 200 states for the formal task, all students agreed that the chosen presentation style will make the task simpler as it is easier to capture the information. Figure 1 shows an example of health state '13542' in English (the original questionnaire was in Chinese)

Figure 1. An example of the health states presentation used in this study

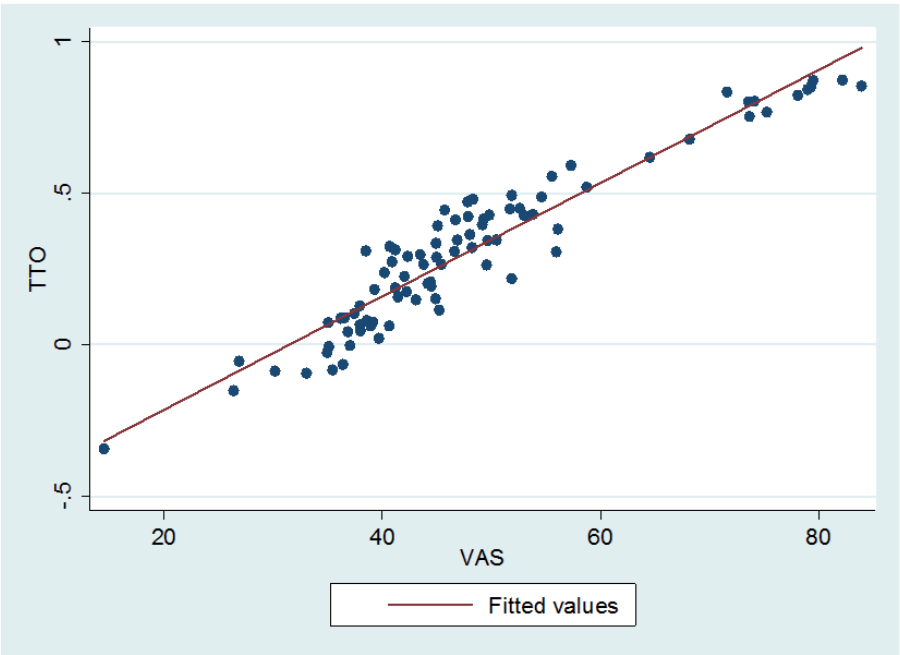
- **Traditional EQ-5D-5L**
- I have no problem in walking about
- I have moderate problems washing or dressing myself
- I am unable to do my usual activities
- I have severe pain or discomfort
- I am slightly anxious or depressed

- **Modified EQ-5D-5L for this study**
- Mobility----------No problem
- Self-care----------Moderate problem
- Usual activity----------Unable to
- Pain/discomfort----------Severe problem
- Anxiety/depression----------Slight problem

During the formal data collection, all students reported their health states using the traditional style of EQ-5D health states before valuing the modified ones. At this step, we briefed the students about the difference in wording between the traditional wording and the modified wording and stated that they are the same. In case of confusing, the students could always turn to the first page of questionnaire to see the traditional wording.

To understand the effect of changing wording, we graphed the values' relationship of 86 EQ-VT design states between 2012 Chinese valuation study and our study. In the figure below, the y-axis is the mean TTO values for 86 states from 2012 Chinese valuation study, which followed the EuroQol protocol of valuation studies; the x-axis is the mean VAS values for the same 86 states from our study. Between those two studies, several differences existed: a. general public from five cities versus. students from one university; b. TTO method versus. VAS method; c. traditional wording versus. modified wording; d. data collected in 2012 versus. data collected in 2016. Despite all the differences, the values from those two studies showed good linear relationship, which indicates the validity of our dataset.

Figure 2. Relationship between 2012 Chinese TTO values and this study's VAS values

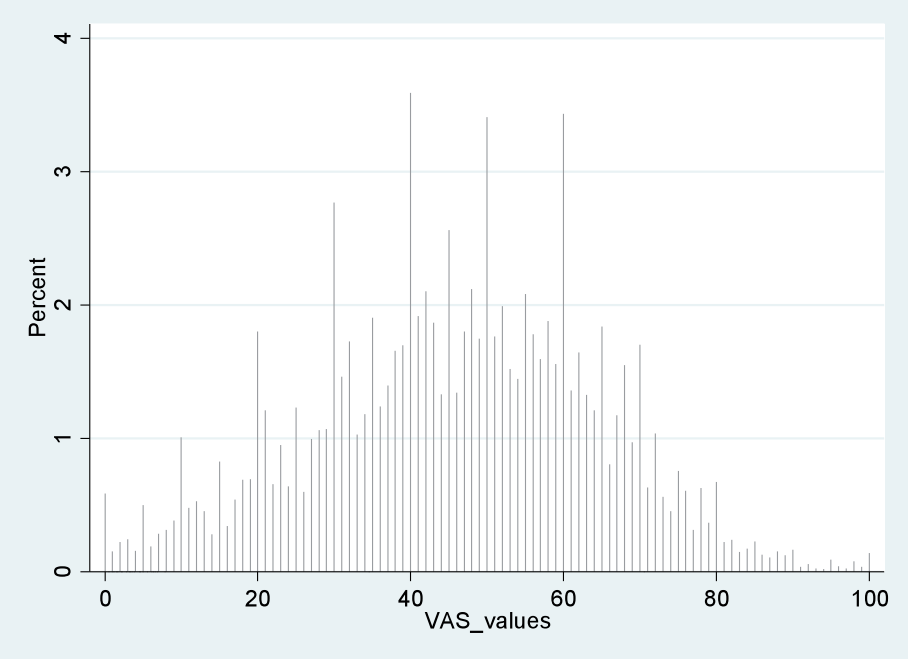


The questionnaires had three sections in the following order: 1) questions to collect background information on respondents, including self-report EQ-5D-5L health status and EQ-VAS of the respondent's own health state; 2) VAS valuations of 11111 and 55555 on one page; 3) VAS valuations of all the other 200 health states in the block, 10 states per page, presented in a random order per block. An EQ-VAS scale was always visible for sections 2 and 3. The EQ-VAS is a 20cm vertical scale from 0 to 100, with endpoints labelled 'the best health you can imagine' and 'the worst health you can imagine'. The respondents were instructed to use the EQ-VAS scale to value each health state by writing down the VAS value beside each health state. Respondents were encouraged to use different numbers/values.

6.5.2 Appendix 2: Details descriptions of the saturated data

The average time to complete the valuation task was 65 minutes (range: 23 to 180 minutes). On average, the students were 21 years old and 62% were female. All students had a health-related education background, such as pharmacy, health law etc.

Figure 3. The count of observed values for all respondents and health states.



As can be seen from Figure 3, which shows the distribution of all observations along the VAS scale, values covered the whole VAS scale, with signs of digit preference (a preponderance of 5s and 10s). The distribution is skewed to the left, with few observations beyond 80. On average, each respondent used 45 distinct values (standard deviation: 12) for the 200 states of the valuation task; the minimum and maximum number of different values recorded were 6 and 80 respectively.

We examined the logical inconsistency for each respondent, and identified all inconsistent observations. In total, 30% of the data had at least 1 inconsistency. To see the effect of removing the illogical observations we made the comparison below. The first Table was used for the manuscript without excluding the illogical observations while the second Table was the results excluding the illogical observations. Nevertheless, we did not exclude any inconsistent observations for several reasons: first, excluding the illogical observations seems increase the prediction errors proportionally across all designs; second, excluding illogical observations affect more for the small design (orthogonal) compared to the large design (EQ-VT). This is reasonable as in the small design, the regression results relied more on each observation; third, similarly like the second reason, some health states may have more illogical observations, as a result, we cannot say that for each health states, there are 100 observations; fourth, as this valuation task is done in a non-standard way (200 states/respondent etc.), it is difficult to find a criteria to exlude illogical observations. Additionally, the current EuroQol valuation protocol did not make suggestions on this issue, different researchers used different criteria to exclude data.

Table 1. Root Mean Squared Error (RMSE) by empirical/validation state set for EQ-VT design and 25 orthogonal design

	Number of health states	RMSE for empirical state set	RMSE for validation state set	RMSE for all 3,125 health states
EQ-VT Protocol (weighted for pits & 5 mildest)	86	2.69	3.69	3.66
EQ-VT Protocol	86	2.65	3.45	3.44
EQ-VT Protocol (excluding pits & 5 mildest)	80	2.39	3.02	3.00
25 orthogonals	25	1.03	3.41	3.40
25 orthogonals (extending pits & 5 mildest)	31	2.61	3.88	3.87

*the italic design was repeated over 100 times

Table 2. Root Mean Squared Error (RMSE) by empirical/validation state set for EQ-VT design and 25 orthogonal design after removing logical inconsistent observation

	Number of health states	RMSE for empirical state set	RMSE for validation state set	RMSE for all 3,125 health states
EQ-VT Protocol (weighted for pits & 5 mildest)	86	4.67	5.01	5.00
EQ-VT Protocol	86	4.67	5.02	5.01
EQ-VT Protocol (excluding pits & 5 mildest)	80	4.68	4.92	4.92
25 orthogonals	25	4.62	5.20	5.20
25 orthogonals (extending pits & 5 mildest)	31	4.66	5.38	5.37

*the italic design was repeated over 100 times

CHAPTER 7

Towards a smaller design for a EQ-5D-5L valuation study

Zhihao Yang, Nan Luo, Mark Oppe, Gouke Bonsel, Jan Busschbach, Elly Stolk

Submitted for Publication

7.1 INTRODUCTION

To estimate an EQ-5D-5L value set, the EuroQol Group developed a standardized EuroQol Valuation Technology (EQ-VT) protocol (28). The design of the EQ-VT protocol includes the selection of 86 different EQ-5D-5L states. These 86 states are arranged into 10 blocks of 10 health states for the composite TTO (cTTO) task. To achieve adequate precision of the mean values, it was decided each block should have 100 observations. The optimal sample size was estimated to be around 1000 respondents =10 blocks x 100 observations/block (29, 53). With this design, EQ-5D-5L value sets for at least 7 countries were established and published (9, 16-22), with more country studies currently underway.

The choice of 86 states for direct valuation through cTTO rested on the considerations of robustness, requiring that each parameter be derived from multiple stimuli. Furthermore, it would enable some interaction terms to be included in the regression model while still allowing for 1 degree of freedom per parameter. However, the latter consideration no longer carries much weight, as practical experience and results from a number of simulations suggest that the 'main effects' model with 21 parameters (5x4 dummy variables + intercept) performs well (16, 96), leaving a surplus of 65 degrees of freedom (DF). The reduced need for a large design has raised the question whether a small design can be used to promote the feasibility of a valuation study. An important criterion in adopting a small design is that the modelling results of a small design should not compromise prediction accuracy at an unacceptable level.

Two recent studies have addressed the impact of design size and the approach to design generation on the accuracy of predicted values, using saturated EQ-5D-5L visual analogue scale (VAS) valuation datasets (30, 98). A saturated dataset contains observed VAS values for EQ-5D-5L states, allowing predictive errors associated with different designs to be quantified. The research strategy taken in these studies was that a subset of health state values was modelled to estimate the value set, and from that value set, predicted values were estimated and compared with the empirical values. The results suggested the current EQ-VT performed well among different design choices, but it is worth noting that the orthogonal design with only 25 health states performed as well as the EQ-VT design assuming the 'main effect' model to be sufficient. One shortcoming of the orthogonal design was that it had larger mispredictions for the mild states compared to the EQ-VT design (12). Similar to the five-level EQ-5D, a study comparing design choices in the three-level EQ-5D VAS saturated dataset also indicated the superiority of the orthogonal design (90).

A limitation of the above-mentioned studies is that VAS was used to collect observed values. Generalizability of the results to valuation data collected using the cTTO method is an open question. VAS values have a different distribution compared to cTTO values. For example, VAS values do not display the relatively linear heteroscedasticity typically observed in TTO values. Also, a property of VAS is 'end of scale aversion'. While the maximum attainable score is 100, many people hesitate to assign such a high score to any state. It is unclear whether this phenomenon partially accounts for the large mispredictions around mild states. In this study, we aim to investigate whether a small design can be used for EQ-5D-5L valuation using cTTO, without increasing mispredictions in any part of the severity scale to an unacceptable level.

7.2 METHODS

7.2.1 Strategy

We collected TTO data for three designs with 500 students. Three designs were divided into five blocks and distributed across the respondents ensuring a minimum of 100 observations per block. Each student valued 1 block of 30 EQ-5D-5L health states. By design, we modelled the observed TTO data to predict values for all possible EQ-5D-5L health states. The predictive accuracy of the design was compared with the observed values in calculating the root mean squared error (RMSE).

7.2.2 Experimental designs

The study comprised three designs that differed in terms of health states included (see the Appendix for the health states in each design):

- The EQ-VT design including 86 health states
- An orthogonal design including 25 states
- A Bayesian-efficient design, including 25 states.

The design of EQ-5D-5L valuation studies has been standardized across countries. It includes a TTO task for 86 EQ-5D-5L states. We refer to this set as the 'EQ-VT design'. Oppe et al provided a detailed description of the TTO task in EQ-5D-5L valuation studies (29). Briefly, the selection of the 80 health states for the protocol was based on Monte Carlo simulations in predicting the prior values obtained from the multi-national pilot study (28). In total 10,000 EQ-VT-like sets of 80-state designs were created. The 'optimal' set was kept as the final design for EQ-VT based on the mean squared error (MSE) between the prior parameters and estimated parameters from a 'main effects' model,

and level balance (53). Six states were manually selected: the worst EQ-5D-5L state and the five mildest states (i.e. state 55555, 21111, 12111, 11211, 11121, 11112) to arrive at a total of 86 states.

An orthogonal design was generated by assigning dimensions and levels defined by EQ-5D-5L to a pre-existing orthogonal array. For EQ-5D-5L an orthogonal main effects design mathematically contains a minimum of 25 states. A theoretical advantage of orthogonal designs is that the absence of correlations between dimensions offers a strong basis for decomposing the observed values to underlying dimension severity levels under the assumption that the model is specified correctly. Orthogonal designs are also level balanced, i.e. each level occurs equally often within each dimension. In this study, we used the best performing orthogonal design from the previous VAS saturated study, in which 100 variants of orthogonal designs were created and tested with respect to their prediction performances for all 3,125 EQ-5D-5L states (98). To account for the above-mentioned issue concerning the prediction of values for mild states in orthogonal designs, the five mild states were added to the orthogonal array. The performance of the design was investigated with and without the five mild states.

In Bayesian-efficient designs, prior information of the model parameter estimates (the coefficients with their uncertainty) guides the selection of states. Unlike the orthogonal design with a fixed number of states, the minimum number of states in the Bayesian design depends on the required degrees of freedom. Furthermore, the design can be generated subject to constraints, for example, to avoid implausible combinations between dimensions or to include the mildest states (4 level 1, 1 level 2) in the design. Hence the Bayesian design is more flexible if constraints are considered. We developed a Bayesian- efficient design using previous Chinese valuation data. For comparison reasons, the number of unique states sampled for the design was set at 25 and the 5 mild states were added for a total of 30 states. Performance criteria were D-error, level balance, mean absolute error (MAE) using the Chinese published value set (9), and MAE on the Chinese VAS saturated dataset (98).

7.2.3 Blocking

The set of 136 (86+ 25+ 25 states) health states was divided into 5 blocks. The EQ-VT design was divided into 3 blocks. To arrive at the same block size of 30, some states were duplicated across blocks: each block containing 55555 and at least 2 of the mildest states. The orthogonal design and the Bayesian design each formed one independent block of 30 states. Each respondent valued one block.

7.2.4 Interviewers & respondents

Following sample size considerations explained by Oppe et al (28), we collected TTO values for the 5 blocks of health states from N=500 university students of Guizhou Medical University, China. A student sample was used in consideration of the large number of health states to be valued (30 states per block) being over-demanding for a general population sample.

Data were collected by 7 interviewers, facilitated by 1 respondent coordinator, who were senior students from Guizhou Medical University, China. Before data collection, all interviewers including the coordinator received 3 days' training regarding background knowledge of the QALY, EQ-5D, and EQ-VT, to help them understand the context of the TTO questions. For the interview location, we rented 4 offices from Guizhou Medical University and set up 2 interview stations in each office. Each interviewer was assigned to a fixed interview station and all interviews were conducted at the interview station. The recruitment advertisement was circulated to university students through e-mails, and interested respondents could contact the coordinator. The coordinator then arranged appointments for the interviewers and respondents. Each respondent was paid 100 RMB (equivalent to 14 euros) upon successful task completion.

All interviews were conducted using EQ-PVT, which is a PowerPoint replica of the EQ-VT software, and was obtained from the EuroQol Research Foundation. There were 3 steps in each interview. First, the background of the study was introduced, and informed consent was obtained orally before the interview, with the respondent being requested to sign a sheet indicating consent to participate in the study. Second, the respondent provided background information and reported his/her health state using EQ-5D-5L. Third, 36 health states including 3 wheelchair examples, 3 practice EQ-5D states and a randomly selected block of 30 EQ-5D states were valued through a face-to-face interview. Following EQ-VT protocol V1.1, each respondent was familiarized with the TTO task using 3 wheelchair example states (wheelchair, a situation worse than a wheelchair, a situation better than a wheelchair), and 3 examples of EQ-5D-5L states (21121, 35554, 15411) (99). The 3 wheelchair examples were used to familiarize respondents not only with the cTTO approach in general, but also to introduce them to the procedure by which they would value a health state as worse than death. The 3 practice states were selected to represent different severity levels (21121=mild, 35554=severe, 15411=moderate) of EQ-5D-5L states. No feedback or debriefing module was provided to the respondents.

Interviews were conducted in accordance with current EQ guidelines for quality control (QC) as reported by Ramos Goni et al (2017). QC reports indicating protocol compliance

and presence of interviewer effects were sent to interviewers every 3 days, with individual suggestions on how to improve interview performance, if necessary.

7.2.5 Data analysis

The data were analyzed by design, using 2 different model specifications plus 2 different standard error specifications, based on the original set (25 states) or the extended set (30 states). Both extending the original set with the 5 mildest states and testing the standard error specifications were explored to resolve the large misprediction issue found in the VAS study. For model specifications, we compared the performance of a 20-parameter additive model (Equation 1) and an 8-parameter multiplicative model (Equation 2). For the 20-parametes additive model, cTTO utility was explained by 20 dummy variables and 1 intercept. For each dimension (MO for mobility, SC for self-care, UA for usual activity, PD for pain/discomfort, AD for anxiety/depression), 4 dummy variables were used to represent the disutility from level 1 to the other 4 levels, e.g. MO_3 took 1 if the health state had a problem in the third level of mobility, and 0 if otherwise (68).

$$\begin{aligned}\text{Utility} = \alpha &+ \beta_1 MO_2 + \beta_2 MO_3 + \beta_3 MO_4 + \beta_4 MO_5 + \beta_5 SC_2 + \beta_6 SC_3 + \beta_7 SC_4 + \beta_8 SC_5 + \\ &\beta_9 UA_2 + \beta_{10} UA_3 + \beta_{11} UA_4 + \beta_{12} UA_5 + \beta_{13} PD_2 + \beta_{14} PD_3 + \beta_{15} PD_4 + \\ &\beta_{16} PD_5 + \beta_{17} AD_2 + \beta_{18} AD_3 + \beta_{19} AD_4 + \beta_{20} AD_5 + \varepsilon. \qquad (1)\end{aligned}$$

The 8-parameter multiplicative model is a variant of the 20-parameter model. It rests on the assumption that the relative distance between levels is the same across all dimensions (34). In this model, 5 dimension parameters were used to represent the disutility of having problems at level 5 on each dimension ($\beta_{MO,}$ $\beta_{SC,}$ $\beta_{UA,}$ β_{PD} and β_{AD}); a set of scalars L2-L4 (i.e. level parameters) was estimated to identify where the cut-offs of the intermediate levels were located, subject to the constraint that the relative distance between levels was constant across all dimensions. Thus, the absolute amounts of the utility of the dimension severity levels were computed by multiplying the relevant scalar L2-L4 with the dimension weights. For example, the disutility of level 4 on anxiety was $\beta_{AD} * L_4$.

$$\begin{aligned}\text{Utility} = \alpha &+ (\beta_{MO} x_{MO2} + \beta_{SC} x_{SC2} + \beta_{UA} x_{UA2} + \beta_{PD} x_{PD2} + \beta_{AD} x_{AD2}) L_2 + \\ &(\beta_{MO} x_{MO3} + \beta_{SC} x_{SC3} + \beta_{UA} x_{UA3} + \beta_{PD} x_{PD3} + \beta_{AD} x_{AD3}) L_3 + \\ &(\beta_{MO} x_{MO4} + \beta_{SC} x_{SC4} + \beta_{UA} x_{UA4} + \beta_{PD} x_{PD4} + \beta_{AD} x_{AD4}) L_4 + \\ &(\beta_{MO} x_{MO5} + \beta_{SC} x_{SC5} + \beta_{UA} x_{UA5} + \beta_{PD} x_{PD5} + \beta_{AD} x_{AD5}) L_5 + \varepsilon \qquad (2)\end{aligned}$$

Rand-Hendriksen et al reported no difference in performance in terms of out-of-sample predictions between the 8- and 20-parameter models on Spanish, Singaporean and Chinese EQ-5D-5L valuation data (34). The 8-parameter model required fewer degrees of freedom compared to the 20-parameter model and perhaps a better model specification for small designs. To address the question with respect to misprediction of the values for mild states, we ran these models by using 3 designs with and without the 5 mildest states, and under either the assumption of homoscedasticity (random effect general linear squared) or of heteroscedasticity (heteroscedastic regression).

To judge design performance, we: (i) predicted health states' values by modelling 3 designs respectively and computed the RMSE within the design/RMSE across designs; (ii) compared the similarity of coefficients between designs; (iii) compared the RMSE along the misery indices to see whether the mispredictions depended on health state severity. The misery index is the sum of 5 digits of an EQ-5D health state (e.g. the misery index of state 13255 = 1 +3 +2 +5 +5 = 16) and was used a proxy for health states' severity levels. Considering that we had multiple models/designs for the results, we reported the RMSE results first and then only the best model for the subsequent analysis. Additionally, we provided the observed values and predicted values of 10 states for reference.

7.3 RESULTS

7.3.1 Raw data

In total 557 interviews were completed. For quality reasons, the first 32 interviews of an interviewer were dropped and after retraining, this interviewer re-conducted 32 interviews with new respondents. Data for the first unqualified 32 interviews were not analysed. On average, respondents spent 46 minutes (SD: 14 minutes) and 7.89 moves (SD: 2.11 moves) on the valuation task, including 6 practice states. The average time to complete a single TTO task was 56.9 seconds for non-practice states. The observed values for all health states can be found in the appendix. The highest mean value was 0.950 for state 12111, and the lowest mean value was -0.719 for state 55555. More severe states had larger standard deviations.

Figure 1: TTO values distribution across three designs

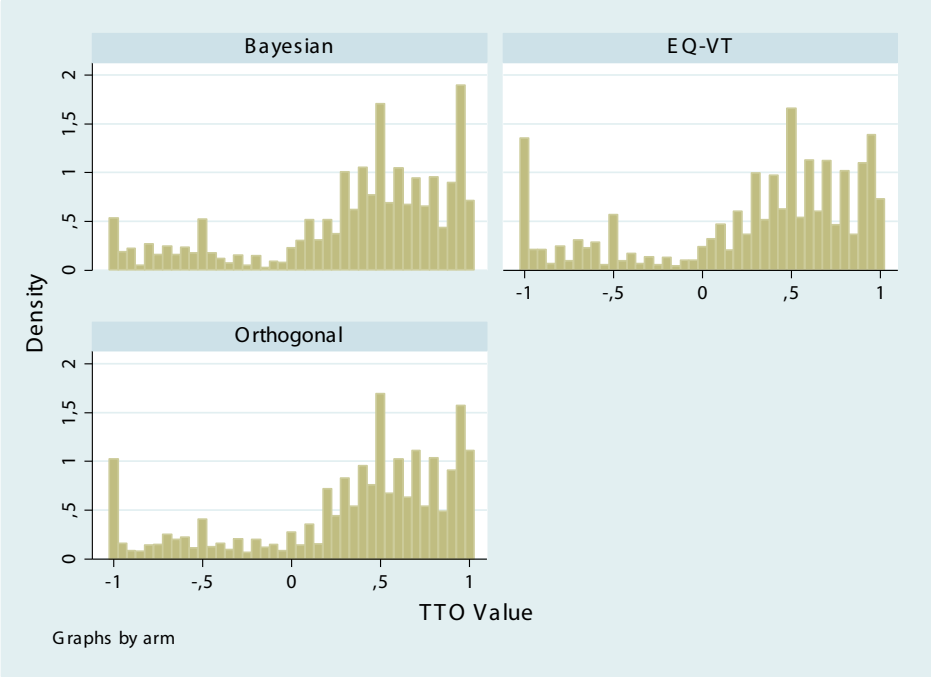


Figure 1 showed similar data distributions for the EQ-VT design and the orthogonal design, while the Bayesian design had fewer observations clustering at -1, but more observations clustering at 0.95.

7.3.2 Modelling & Prediction performance

Table 1: RMSE full results

Model	All	EQ-VT	Orthogonal+5	Orthogonal	Bayesian+5	Bayesian
# Health states	146	86	30	25	30	25
Multiplicative	0.051	0.053	0.066	0.067	0.063	0.067
GLS additive	0.049	0.053	0.069	0.069	0.063	0.095
Hetero additive	0.051	0.054	0.064	0.072	0.065	0.092

Table 1 indicates that the EQ-VT design performed better than the 2 small designs in terms of overall RMSE (0.053). The choice of model specification and standard error specification did not impact much on the RMSE results. Notably, extending the 5 mildest states lowered the overall RMSE for the Bayesian designs (from 0.095 to 0.063), but not

for the orthogonal design (0.066 versus 0.067). Next, we report the results of 3 designs using the 8-parameter multiplicative model with the 5 mildest states extended for 2 small designs.

Table 2: Multiplicative regression model output & RMSE results

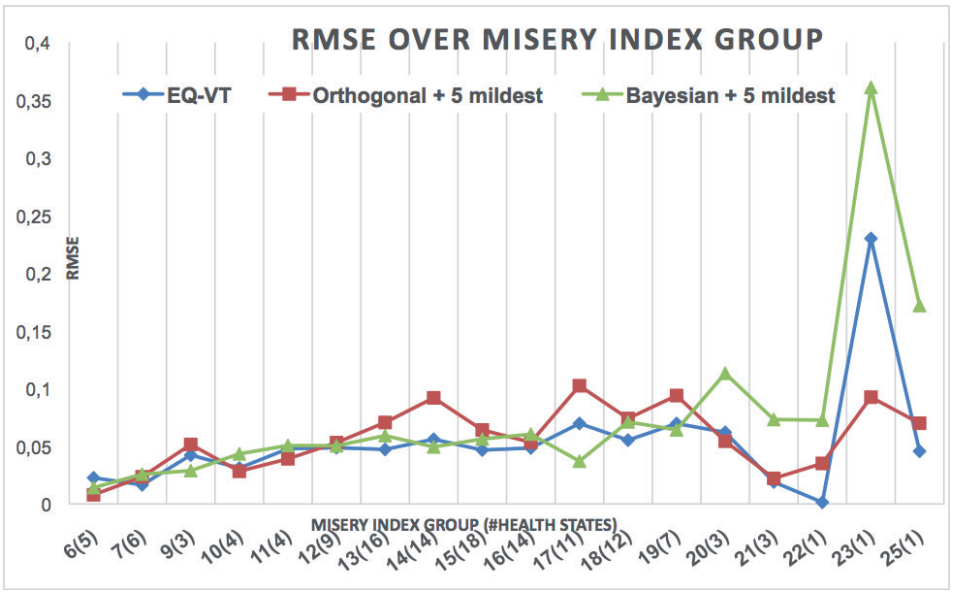
	Bayesian + 5 mildest states	Orthogonal + 5 mildest	EQ-VT
#Health states	30	30	86
Constant	0.961 (0.920, 1.002)	0.943 (0.904, 0.983)	0.992 (0.961, 1.023)
Mobility	-0.280 (-0.335, -0.225)	-0.363 (-0.402, -0.325)	-0.296 (-0.320, -0.272)
Self-care	-0.207 (-0.254, -0.161)	-0.270 (-0.309, -0.232)	-0.244 (-0.269, -0.219)
Usual activities	-0.275 (-0.314, -0.235)	-0.330 (-0.368, -0.291)	-0.294 (-0.319, -0.268)
Pain/discomfort	-0.425 (-0.464, -0.385)	-0.356 (-0.395, -0.317)	-0.404 (-0.430, -0.379)
Anxiety/depression	-0.322 (-0.359, -0.284)	-0.413 (-0.452, -0.374)	-0.428 (-0.453, -0.402)
Level 2	0.127 (0.065, 0.188)	0.036* (-0.017,0.089)	0.115 (0.077, 0.152)
Level 3	0.316 (0.261, 0.370)	0.258 (0.211, 0.304)	0.293 (0.259, 0.326)
Level 4	0.845 (0.778, 0.913)	0.674 (0.621, 0.726)	0.780 (0.748, 0.812)
RMSE for empirical state set	0.024	0.036	0.051
RMSE for validation state set	0.070	0.072	0.054
RMSE for all states	0.063	0.066	0.053

*not significant on 0.05 level

Table 2 shows the regression output of the 8-parameter model. In general, the Bayesian design had the largest level parameters (L2, L3 and L4) and the orthogonal design had the smallest level parameters. 95% confidence intervals did not overlap for the dimension parameters of Mobility (EQ-VT vs orthogonal) and Anxiety (Bayesian vs orthogonal & EQ-VT), and the level parameter of level 4 (orthogonal vs EQ-VT & Bayesian). In terms of RMSE, small designs had better predictions for the empirical state set than the EQ-VT design. EQ-VT had the lowest RMSE for all states (0.053) and RMSEs of two small designs did not differ much (0.069 versus 0.063).

All 3 designs performed similarly in mild states (i.e. with a maximum misery index of 13). The EQ-VT design performed evenly along the scale except for group 23, and in that group, only 1 state presented in our dataset, i.e. 55535, which belonged to the

Figure 2: RMSE of three designs over severity scale



*In the bracket after each misery index group shows the number of health states in that group

orthogonal design. The predicted value for 55535 using the EQ-VT design was -0.343 while the actual value observed was -0.617. The orthogonal design performed worse in the 2 misery index groups 14 and 17, while the Bayesian design did not perform well in the severe part of the scale (misery index>20).

Table 3: Observed values & predicted values for 10 random health states

Health states	Observed value	Standard error	Lower 95%CI	Upper 95%CI	Predicted by EQ-VT	Predicted by Orthogonal +5	Predicted by Bayesian +5
11221	0.915	0.007	0.901	0.930	0.912	0.916	0.847
12324	0.502	0.038	0.427	0.578	0.498	0.557	0.518
24422	0.437	0.035	0.368	0.505	0.443	0.498	0.421
31514	0.362	0.040	0.283	0.441	0.278	0.241	0.324
34234	0.311	0.041	0.230	0.392	0.229	0.286	0.263
43154	-0.017	0.050	-0.115	0.082	-0.048	-0.005	-0.034
43542	-0.009	0.055	-0.117	0.010	0.032	0.045	-0.006
44355	-0.365	0.048	-0.461	-0.269	-0.347	-0.336	-0.271
55424	-0.259	0.056	-0.370	-0.147	-0.157	-0.203	-0.073
55535	-0.617	0.041	-0.699	-0.536	-0.387	-0.524	-0.239

7.4 DISCUSSION

This study built on the previous EQ-VAS study, in which we found that statistically efficient small designs could be used for valuation studies without compromising prediction accuracy (98). Similar to the EQ-VAS study, this study using TTO data showed that the EQ-VT design performed best in terms of prediction accuracy. Smaller designs had higher RMSEs, but the difference was relatively small: 0.01 on a utility scale. It should be noted that in this study the comparison was made only on 136 states instead of all 3,125 states in the EQ-VAS study. As the EQ-VT design accounted for 63% (=86/136) of all states, the overall prediction result was more advantageous towards EQ-VT. Also, using the same model, the parameters of the 3 designs differed considerably. There are 2 possible reasons: (i) a different design (subset of health states) might produce a different set of coefficients; (ii) respondents may have differed in health preferences across the 3 arms. Further study is needed to understand this issue.

It is notable that we did not encounter the large misprediction problem found in the EQ-VAS study, as the RMSE of the mild states (VAS>70) was twice the size of the other states (98). A possible explanation for the large misprediction in the VAS study was the large gap between the value of 11111 and the values of any other states (98). The relative magnitude of the RMSE in this study was larger than the RMSE reported in the VAS study. In the VAS study, the RMSEs for all 3,125 states were 3.44 and 3.87 (on VAS scale: 0 to 100) for the EQ-VT and orthogonal design respectively, while the counterparts were 0.053 and 0.066 (on utility scale: -1 to 1) in the current study. This may have been due to the difference in value distributions between VAS data and cTTO data, e.g. VAS data does not have two parts separated by death as inherent in cTTO data. There was more heterogeneity in the cTTO data: respondents used the scale differently, e.g. some did not enter WTD, some would not go below 0.5, and so on. In contrast, the VAS data showed that most respondents used the values from the same interval of 20 to 80.

With respect to the 20-parameter model, the coefficients of all 5 level 2 dummy variables were not significant for the 2 small designs. One possible explanation is that university students did not trade-off life years for the mild problems in the TTO task, which resulted in negligible effects for the corresponding variables. Increasing the degrees of freedom by extending the small designs with the 5 mildest states and using an 8-parameter model improved this issue for the Bayesian design. This explained the improved prediction performance of the Bayesian design when using the 8-parameter model extended with the 5 mildest states. However, these 2 approaches did not work on the orthogonal design. The different impact of the mild states' extension may be explained by the different basis

on which these two designs were constructed. In the orthogonal design, the extension disrupted orthogonality and level balance. By comparison, the focus of Bayesian design is on standard errors around parameters. Hence, by extending the number of states, modelling results could improve as extra states could provide information on the parameters which are difficult to estimate. Future research could use the small designs approach in the general population to further understand this issue.

From the comparison, the difference in prediction performance between the two small designs was minimal, but to develop a Bayesian design requires prior information, which is not always available in practice. It should be noted that the Bayesian design in this study used the prior information from previous valuation studies conducted in China (9, 15), and hence may not be the optimal design for valuation studies in other countries. An open question is whether using a different design would result in different results in cost-utility analysis. In previous research, the use of different EQ-5D value sets led to different incremental cost-effectiveness ratios (ICERS) (100). In our study, the coefficients for each design differed in the value sets, thus estimating different predicted health state utilities. Future research could evaluate the effect of adopting small designs on cost-utility analysis.

This study had several limitations. First, the TTO task that needed to be completed by each respondent included 30 EQ-5D-5L health states, plus another 6 exercise states. The working load almost tripled compared to the standard EQ-VT protocol. Second, we used university students as respondents. Comparing the value variance between students and the general public, the mean standard deviation of the EQ-VT 86 states from our student sample was 0.416, while the counterpart from the 2012 Chinese general public valuation study was 0.479 (9). It can be observed that students had more homogeneous health preferences than the general public, who had more socio-economic differences. This is a minor issue as the purpose of this study was to test hypotheses rather than generate value sets.

The empirical findings offer support for the use of the current EQ-VT design. It is difficult to tell precisely what the results mean for smaller designs because the level of concern with the change in RMSE essentially rests on an arbitrary evaluation. Considering that the relative increase of RMSE was modest, we do not feel that the results raise a red flag over the use of small designs, especially if one considers that valuation researchers have other options to promote the robustness of their results. An established way to reduce the risk of errors in predicted values involves combining multiple types of valuation data. For instance, the EQ-VT protocol has a DCE task included alongside the cTTO task and a hybrid model can be used to model DCE data and cTTO data together (101). The use of

the hybrid model has proved superior to using the TTO data alone (20, 22, 101). Such an approach offers a way to manage the risks of larger error in predicted values associated with smaller TTO designs, while not undoing the benefits of shrinking the TTO design, since collecting DCE responses is less resource demanding.

7.5 CONCLUSION

The EQ-VT design had the best prediction performance and should be used as the default design for EQ-5D-5L valuation studies. Smaller designs also performed quite well, and may be considered for use in some specified contexts such as for methodological research, in resource-constrained countries, and if data collection is paired with other data collection approaches, e.g. DCE.

7.6 APPENDIX

Table 1. Observed values and standard deviations

EQ-VT						Orthogonal			Bayesian		
HS	Mean	SD	HS	Mean	SD	HS	Mean	SD	HS	Mean	SD
21111	0.933	0.080	23514	0.310	0.438	**21111**	0.922	0.110	**21111**	0.932	0.075
12111	0.950	0.053	31524	0.254	0.537	**12111**	0.927	0.093	**12111**	0.936	0.074
11211	0.932	0.098	53412	0.423	0.343	**11211**	0.945	0.064	**11211**	0.930	0.079
11121	0.926	0.094	52431	0.303	0.502	**11121**	0.934	0.099	**11121**	0.923	0.123
11112	0.940	0.107	24342	0.383	0.408	**11112**	0.932	0.095	**11112**	0.939	0.078
12121	0.917	0.104	22434	0.204	0.533	21331	0.700	0.279	12231	0.741	0.238
11212	0.910	0.091	21444	0.141	0.493	11215	0.536	0.414	31114	0.567	0.374
12112	0.900	0.090	12543	0.193	0.530	24113	0.601	0.271	21125	0.533	0.365
11221	0.915	0.074	35332	0.446	0.355	32511	0.540	0.328	**11235**	0.461	0.411
11122	0.878	0.127	51451	0.052	0.559	42132	0.547	0.318	22324	0.473	0.372
21112	0.876	0.122	45133	0.280	0.463	12324	0.502	0.393	53113	0.495	0.326
11421	0.700	0.283	33253	0.230	0.516	**25222**	0.644	0.263	11551	0.234	0.512
13122	0.837	0.124	44125	0.163	0.487	**15151**	0.329	0.482	44411	0.340	0.400
14113	0.645	0.257	43315	0.145	0.534	53312	0.432	0.367	33341	0.361	0.395
11414	0.393	0.510	35143	0.142	0.536	33125	0.394	0.388	21452	0.200	0.496
13313	0.648	0.295	31525	0.137	0.552	43241	0.290	0.473	11444	0.135	0.505
13224	0.560	0.315	32443	0.292	0.412	13433	0.453	0.346	24422	0.437	0.351
42321	0.516	0.436	45233	0.296	0.422	41523	0.271	0.464	42531	0.278	0.470
11235	0.442	0.473	24443	0.045	0.569	51144	0.109	0.540	32433	0.387	0.377
25122	0.624	0.337	34244	0.045	0.537	31452	0.254	0.502	35223	0.517	0.282
21315	0.386	0.494	53243	0.092	0.557	14542	0.163	0.496	51333	0.357	0.412
12513	0.614	0.245	43514	0.141	0.495	54421	0.192	0.491	42145	0.040	0.522

Table 1. Continued

EQ-VT						Orthogonal			Bayesian		
HS	Mean	SD	HS	Mean	SD	HS	Mean	SD	HS	Mean	SD
11425	0.257	0.543	45413	0.260	0.435	34234	0.311	0.421	43513	0.283	0.426
35311	0.525	0.414	55233	0.163	0.516	52253	0.124	0.514	34352	0.134	0.483
53221	0.587	0.289	43542	-0.009	0.560	22445	0.037	0.534	43154	-0.017	0.504
42115	0.350	0.448	54342	-0.055	0.584	**45414**	-0.044	0.515	24515	0.184	0.463
12334	0.495	0.387	52335	0.031	0.562	35343	0.194	0.520	**45414**	-0.027	0.529
25222	0.583	0.379	54153	0.108	0.479	23554	-0.063	0.508	25542	0.049	0.512
12244	0.272	0.492	34155	-0.104	0.549	44355	-0.365	0.501	31555	-0.105	0.519
23242	0.550	0.277	45144	-0.125	0.543	55535	-0.617	0.424	25445	-0.246	0.549
23152	0.380	0.453	53244	-0.121	0.589						
32314	0.384	0.489	34515	-0.027	0.575						
21334	0.437	0.377	14554	-0.146	0.533						
15151	0.434	0.379	55225	-0.016	0.542						
12514	0.345	0.525	24445	-0.276	0.567						
51152	0.358	0.428	35245	-0.109	0.588						
34232	0.497	0.344	24553	-0.071	0.564						
25331	0.523	0.293	44345	-0.291	0.564						
12344	0.159	0.568	55424	-0.259	0.574						
31514	0.362	0.408	44553	-0.250	0.519						
21345	0.209	0.502	52455	-0.420	0.519						
52215	0.199	0.546	43555	-0.434	0.530						
54231	0.379	0.424	55555	-0.719	0.356						

CHAPTER 8

General Discussion

In this chapter, first the research questions raised in Chapter 1 are answered, then the limitations of the thesis are identified. Subsequently, the implications of the research undertaken in the thesis are discussed and ideas for future research offered.

8.1 KEY FINDINGS: RESPONSE TO THE RESEARCH QUESTIONS

This thesis includes six studies focused on EQ-5D use and health valuation research in China. In Chapter 2, the 2012 Chinese valuation data was used to establish population norm scores for EQ-5D-5L and to examine how demographic factors affected individuals' self-reported health states (research question #1). The norm scores were reported in Chapter 2, in Tables 2 and 3 for males and females respectively. In general, HRQoL outcomes measured by EQ-5D-5L differed over age, gender, education level, health insurance status, employment status, and the residence of origin groups with lower socioeconomic status related to lower HRQoL outcome. There were two noteworthy results. First, the prevalence of reported problems in anxiety/depression decreased with age; second, females reported higher EQ-VAS values than males.

In Chapter 3, the potential cause of individual-level inconsistency in the TTO task was explored (research question #2). It was confirmed that most respondents could use the TTO task to express their health preferences following the EQ-VT interview protocol. Furthermore, using inconsistency as a proxy to compare interviewers' performance, it was found that the Chinese TTO data, which was collected without a quality control procedure, was affected by interviewer effects, i.e. data from certain interviewers had much higher inconsistency rates and larger inconsistency magnitudes. These effects were profound and could partially explain regional differences in the 2012 China valuation data. In that study, Beijing showed higher values than the other four cities. Notably, as different teams of interviewers were used in the different cities it was impossible to disentangle interviewer effects from the effects of regional differences in that study. Hence, it is not clear if the differences in values reflected true differences in health preferences or arose as an artefact of the study design. A safer approach would be to conduct a valuation study using one team of interviewers, to attempt to eliminate interviewer effects which might mask potential regional differences.

From Chapter 4 to Chapter 7, there is an examination of one important design choice in health valuation research: how to select health states for direct valuation (research question #3). First, in Chapter 4, an existing EQ-5D-3L saturated dataset was utilized to

compare prediction performance between designs so as to identify the most important design principle in selecting health states. In Chapter 5, the results of a qualitative study concerning implausible health states were reported and there was an examination with respect to how perceived implausibility affected health valuations. In Chapter 6, based on the experience of Chapter 4, a new EQ-5D-5L saturated dataset was collected to test the generalizability of results found in Chapter 4. Finally, given the difference in data behaviour/distribution between VAS and TTO, in Chapter 7, the most efficient TTO data design identified in Chapter 6 was tested in comparison with the standard EQ-VT design.

For the health states selection, the study showed that the concern for statistical efficiency within the design outweighed the concern with respect to the 'commonness' of the health state (research question #4). In published EQ-5D-3L valuation studies, different designs were used or proposed (60, 67, 68, 75). These design choices were either attempting to ease the valuation burden for the respondents (e.g. excluding uncommon or implausible health states from the design), or to increase the feasibility of the valuation study by valuing fewer states (e.g. using a 17-state subsample of the 42-state MVH study). To test the consequences of these concerns, an empirical approach was taken by imitating different designs' prediction performance in a saturated dataset. A three-step approach was utilized: (i) the empirical values of a design were used to construct a value set, (ii) this value set was used to predict the values of all the health states, (iii) the root mean squared error (RMSE) was computed between empirical values and predicted values. Utilizing this approach, it was found that the use of different designs affected prediction accuracy for the non-valued states' values and by comparison, a statistically efficient design performed better than other designs. Specifically, an orthogonal design performed best among all the comparators.

Simulations performed on the 5L VAS saturated dataset suggested that the standard EQ-5D-5L design (EQ-VT) recommended by the EuroQol Group was valid. Yet, similar to findings in the EQ-5D-3L context, smaller designs with optimal statistical efficiency could achieve similar modelling results (research question #5). The good result for smaller designs was also validated by the TTO data. As another issue for health state selection, the results provided little evidence to exclude implausible health states from the empirical state set/experimental design. First, the observed effect of implausibility on values was quite small (0.91 on the VAS scale, SD:5.58); second, there was no agreement among respondents about exactly which states needed to be classified as implausible.

Overall, the work undertaken here on the design issue suggests that the individual health state selection approach is not appropriate. As an alternative, considering a

selected subset of states as a whole and focusing on its statistical properties would be more advantageous. The selection should lead to a statistically efficient design, which could produce more accurate estimations. The issue of implausible or uncommon health states may still matter for respondents in terms of face validity. Putting these two concerns together, the selection of states for direct valuation is a trade-off between efficient designs with some uncommon/implausible health states, which have some low-value observations, versus a statistically imbalanced design. As the ultimate purpose of a valuation study is to estimate values as accurately as possible, statistical considerations are more important.

8.2 LIMITATIONS

8.2.1 HRQoL instruments other than EQ-5D

This thesis has focused on EQ-5D, which is a preferred instrument for the assessment of HRQoL worldwide and thus a first choice among Chinese health outcome researchers. Alternative HRQoL instruments are available, including SF-6D, 15D, AQoL, and HUI. These HRQoL instruments differ in terms of dimensions and level descriptors. Differences in the description of health states led to different preference weights and in many cases, different utility results (102). For example, EQ-5D has the merit of it being simple to use and its focus on offering an index value, whereas an instrument such as SF-6D has a wider health state descriptive system. While we are aware of these instruments' existence, it is difficult to judge which one is the most suitable to use in health description and health economic evaluation in China. One common characteristic of these instruments is that all were developed and most thoroughly tested in studies in western countries. Yet we do not know whether, for example, the five dimensions in EQ-5D are the most relevant health dimensions for the Chinese in terms of describing HRQoL.

HRQoL instruments are usually first developed for use in one country and then applied elsewhere. While this is practical, it may not always be appropriate to apply an instrument to a culture other than the one for which it was originally developed. Culture influences people's ways of living, thinking, expressing themselves (103), and hence inevitably their ways of conceptualizing and evaluating psychological concepts such as HRQoL (104-106). Indeed, the psychometric evaluation of SF-36 (107-109) and the Center for Epidemiologic Studies Depression (CES-D) Scale (110-112) found that their measurement models were not applicable to some cultures. For example, the relationship between the eight SF-36 concepts and the two SF-36 health components (physical and mental health) in Japan, Singapore, and Taiwan was different from that in the US (107-109). It should be noted that

all non-English versions of the SF-36 questionnaire were developed using a standardized iterative translation protocol (113), which suggests that good translation and/or cultural adaptation may not ensure cross-cultural applicability of HRQoL instruments.
The EQ-5D questionnaire, whether EQ-5D-3L or EQ-5D-5L, has been used to assess the HRQoL or health status of populations all over the world (114). A large body of literature has demonstrated the psychometric properties of the EQ-5D questionnaire in many different cultures. In contrast, qualitative evidence on the content validity of the instrument (i.e. the relevance and adequacy of the health dimensions included in its descriptive system) is limited. To the best of my knowledge, the content validity of EQ-5D has been qualitatively assessed only with UK and Australian research professionals (115), UK patients with diabetes (116), and elderly Dutch people (117). Qualitative research on content validity is an important step in HRQoL instrument development and evaluation which cannot be replaced by quantitative assessments of psychometric properties. It can be noted that the FDA of the United States requires documentation of evidence on content validity for the patient reported-outcomes instrument used to support label claims (118).

China is the world's most populous and rapidly developing country. Driven by rising expectations and population ageing, health outcomes and health policy research in China is growing at a steady pace. A PubMed search suggested a 900% increase (in 2017: 2,038 , and in 2007: 224) in publications using the key words 'China' and 'Health outcome' (119). To date, HRQoL instruments used in the region are predominantly those developed in North America or Europe, due to the lack of locally-developed instruments in China itself. Despite this lack, there are concerns about the appropriateness of using western HRQoL instruments in Chinese populations (120). Possibly, there might be equally or even more important health dimensions in China not included in the five EQ-5D health dimensions, so EQ-5D might not be sensitive to culturally specific health problems important to the Chinese citizen.

All these findings described above suggested that it would be useful to perform a study to assess the appropriateness of the EQ-5D questionnaire for use in Chinese populations. Such a study should aim to ascertain the relevance and adequacy, or content validity, of the EQ-5D descriptive system in the target populations. Specific research questions are: What health concepts or dimensions are most important to the study populations? How many of the health concepts are covered by EQ-5D? How important are the EQ-5D health dimensions to the study populations?

Addressing such questions could strengthen the evidence for the use of EQ-5D in China.

Currently, EQ-5D is the only instrument that has Chinese value sets attached. While it is possible to establish Chinese value sets for SF-6D and HUI, the valuation study designs for other instruments have not been examined as rigorously as has been accomplished for EQ-5D.

8.2.2 Difference data characteristics between VAS and TTO

In Chapters 4 and 6, VAS valuation data was used as a proxy for TTO data. Differences between the two kinds of valuation data should be noted. First, VAS values generally do not display the relatively linear heteroscedasticity typically observed in TTO values; second, VAS values generally display unimodal approaching a bell-curve distribution, particularly around the middle range; third, a property of VAS is 'end of scale aversion'. While the maximum attainable score is 100, many people hesitate to assign such a high score to any state. In summary, VAS values are more 'well-behaved' than their TTO counterparts and are therefore more likely to conform 'nicely' to modelling efforts. In this sense, our VAS 3L and 5L studies were conducted under 'better' conditions, from a modelling point of view, than could be expected even in the best-case scenario for TTO data. As such, designs that may work well using VAS data may not work as well with TTO, and probably require more observations (and/or number of observed states) in TTO-based research.

Due to the distributional difference, the relative magnitude of the RMSE in the TTO study was larger than the RMSE reported in the VAS study. In the VAS study, the RMSEs for all 3,125 states were 3.44 and 3.87 (on VAS scale: 0 to 100) for the EQ-VT and orthogonal designs respectively, while the counterparts were 0.053 and 0.066 (on utility scale: -1 to 1) in the current study. Also, the 'end of scale aversion' property in VAS offered a partial explanation for the large mispredictions around mild states shown in the Chapter 6 where VAS data was used, but not in Chapter 7 where TTO data was used.

8.2.3 Sample issue

In the 2012 Chinese valuation study, the sample employed was collected in five urban areas in China, which was not representative of the whole Chinese population. This posed a limitation to the norm study reported in Chapter 2 as socio-economic differences exist between different areas and, between urban and rural areas the health status of residents may differ by type of area (40). Furthermore, most respondents were recruited in public locations, which may have led to a selection bias towards healthy respondents.

In both 3L (Chapter 4) and 5L (Chapter 6) saturated studies, a student sample was used instead of a general public sample. Students valuing a large number of health states

using VAS are more likely to successfully develop mental shortcuts to ensure consistency and monotonicity than respondents valuing fewer health states using the much more complicated TTO method. This too should be expected to produce more 'well-behaved' data. It is noteworthy that in this study students were used for the sample. For the EQ-VT design, the mean standard deviation across 86 states for the student sample was 0.416 while the counterpart from the 2012 Chinese general public study was 0.479. It can be observed that the students had more homogeneous health preferences than the general public, since the general public had more socio-economic differences. Thus, when adopting the small design for use with the general public, we may expect to use more observations or more health states in order to achieve an acceptable level of prediction accuracy.

8.3 IMPLICATIONS

The norm scores presented in Chapter 2 of the thesis are a valuable asset for EQ-5D users. They provide a reference point for clinical and health economic research outcomes, so that the HRQoL of a given patient group can be compared with that of the general population (5). For example, hypothetically, one study measured a patient group (diabetes) with an average VAS score of 67.9 and an index value of 0.876 for male patients aged between 40-49. From Table 2 of Chapter 2, the corresponding VAS score and the index value of the 40-49 male group from the general public were 85.5 and 0.959 respectively. We could thus calculate the effect of having diabetes on HRQoL, i.e. the effect on VAS (85.5-67.9=17.4), and the effect on the index value (0.959-0.876= 0.083). In such a way, we could estimate the burden of different diseases and use such information to inform policy-making.

This thesis also provides practical guidance for future valuation studies. In terms of sample representativeness, similarly to the effect on self-reported HRQoL, previous studies have shown that demographic factors have also influenced health preferences (26-28). Hence, given the large demographic variance in China, future valuation studies could benefit from improving sample representativeness, e.g. also recruiting respondents from rural areas. Nevertheless, demands to improve sample representativeness came at the price of reduced feasibility for the valuation study. Fortunately, this challenge was moderated by the potential use of small designs. Small designs with optimal statistical efficiency could lower the total number of health states needed to be valued whilst still providing an acceptable prediction level. Thus, the use of small designs paves the way for more cost-effective EQ-5D-5L valuation studies. Following the EQ-VT protocol, each state

needs to be valued by 100 respondents, and with each respondent valuing 10 states we may expect that by using a small design, the minimum sample size requirement for a valuation study could be reduced from 1000 respondents to only 300 respondents. Hence, the resources freed up by adopting small designs could be used to engage the more difficult-to-reach respondents from rural areas in order to improve sample representativeness.

Another implication for health valuation researchers is that the concern with respect to inconsistency in the TTO task has been largely addressed by extending the EQ-VT protocol to employ a quality control (QC) tool (57). Ramos Goni et al reported that the implementation of the QC process was found to lower the inconsistency rate from 11% to 3% in the Spanish valuation study (57). Thus, one important lesson has been that the interviewer selection/training/monitoring process is the key to successful data collection.

To sum up, the work reported in this thesis has identified three important suggestions with respect to the conduct of a successful valuation study. First, use the right design, such as an orthogonal one; second, recruit the right interviewers (who are open for feedback and would like to learn on the job, who apply soft skills to handle difficult topics, who maintain motivation, etc.), train them well, and monitor their performance during data collection; third, improve sample representativeness. The last factor also applies to studies aimed at measuring population health.

8.4 A POSSIBLE MORE FUNDAMENTAL APPROACH?

The identification of small designs for TTO data promotes the feasibility of a valuation study, yet, for a country as large and diverse as China, it is important to improve sample representativeness. Hence, using a less resource-demanding valuation technique may be a more fundamental way to address the sample issue. Discrete Choice Experiment (DCE) valuation has been an integral part of the EQ-5D-5L valuation protocol, and it has been argued that, as both TTO and DCE have provided information for the same object (i.e. health state), they should be used together for modelling purposes (121). As the DCE method only requires a respondent to make a choice between two options, it is relatively easy for respondents to complete and the task can be self-completed, e.g. online or using postal surveys, without requiring assistance from an interviewer. This allows the possibility of estimating a national representative EQ-5D-5L value set for China. For example, 20 regions could be selected based on their level of economic development

and cultural characteristics. Each of the regions could undertake a regional valuation study using the small orthogonal design with 30 states employing the TTO method, plus a large online sample of respondents completing only DCE valuation tasks. Thus, each region could establish its own EQ-5D value set and, more importantly, utilizing the same design and the same valuation techniques (e.g. EQ-VT, QC, standardized interviewer training, etc), a balanced national value set could be established. In addition, research has shown that with some modification of the basic DCE task (e.g. adding time as another dimension), a value set anchored on a QALY scale could be estimated with good validity (122). Either way, it is reasonable and beneficial for China to invest further in developing and understanding novel health valuation methods such as the DCE approach.

CHAPTER 9

Summary

This thesis included 6 studies which reported about various aspects of the EQ-5D, with a focus on its use in China. The population norm study provided the first set of norms based on urban Chinese self-reported health status. This study not only provided insight into Health-Related Quality of Life (HRQoL) variations among subgroups, but also served as a reference point for other disease studies and intervention studies. The subsequent methodological studies of this thesis offered suggestions for improving the design of future valuation studies. These suggestions will strengthen both health technology assessments (HTA) and cost-utility analyses in China and beyond, as the EQ-5D is the most used HRQoL instrument in HTA and cost-utility analyses. At present, the Chinese government has not yet adopted cost-utility analysis as a basis for healthcare coverage decisions, but there are strong signals that this approach will be implemented in the future. For example, in the latest 'China Guideline for Pharmacoeconomic Evaluations', it was stated that cost-utility analysis is preferred over other economic evaluation methods. Policymakers in China need valid instrument to support cost-utility studies in this field. Since EQ-5D is the most widely-used instrument worldwide for this purpose it is also a good candidate for China.

Chapter 2 reported about the norm scores of EQ-5D-5L in the urban Chinese population. Additional analysis was undertaken to test whether self-reported HRQoL varied between different demographic groups. It was found that HRQoL outcomes differed over demographic subgroups: i.e. age, gender, education level, health insurance status, employment status, and the residence of origin groups.

In **Chapter 3**, by analysing the relation of individual level inconsistency in the TTO task with different factors of the interview, it was found that the inconsistencies respondent made in the TTO task varied significantly between interviewers. The results suggested that the valuation process have been influenced by interviewer effects when done without a solid quality control.

Commencing in Chapter 4, possible designs to be used for the EQ-5D valuation study were systematically examined and compared. First, in **Chapter 4**, an EQ-5D-3L 'saturated dataset' was used as a golden standard to compare two often-mentioned design principles in selecting health states for direct valuation: 'commonness of health states' (the prevalence) versus statistical efficiency of a design. By simulating the selection and the modelling process, it was found that the principle of statistical efficiency outweighed the principle of commonness in achieving sufficient prediction accuracy for non-valued states. This result suggested that the selection of health states in previous valuation studies were not optimal, and that future valuation studies could use a smaller design with optimum statistical efficiency.

Chapter 5 described how 1,600 students were invited to value all EQ-5D-5L states using a visual analogue scale and how they judged the implausibility of each state. The results showed that respondents lacked agreement concerning which states were implausible. The mean value of a state valued by respondents who thought it was implausible was lower than counterpart valued by respondents who thought it was plausible, but both mean values were still in good agreement.

Learning from design selection experience with EQ-5D-3L, the research presented in **Chapter 6** aimed to test the selection of health states used in the current EQ-VT design and to identify a possible smaller design for EQ-5D-5L valuation studies. The good performance in using a statistical efficient orthogonal design was confirmed again with EQ-5D-5L, i.e. an orthogonal design with 25 states performed equally well as the EQ-VT design with 86 states in terms of prediction accuracy for all 3,125 states. Despite the favourable outcome, there were some concerns: first, large mispredictions from the orthogonal design in the upper part of the severity scale needed further investigation; second, VAS values normally differ with TTO values in terms of data distribution characteristics, which put the generalizability of the results to TTO data into doubt. Accordingly, in **Chapter 7**, the most efficient design (the orthogonal) was tested in comparison with the standard EQ-VT design using TTO data. The favorable result of 25-state orthogonal design was confirmed in TTO data, in the sense that the overall prediction was as good as the EQ-VT design of 86 health states.

Chapter 8 responded the research questions raised in **Chapter 1** and provided some thoughts for future research. With this thesis, attempts were made to better understand the possible effects of sample and design choices in previous Chinese valuation studies. The findings of this thesis can be generalized to other countries' EQ-5D studies, or to improve valuation studies employing other instruments.

CHAPTER 10

Samenvatting

In dit proefschrift worden 6 studies gepresenteerd die zich richten op de verschillende aspecten van het gebruik van de EQ-5D, met een focus op het gebruik in China. De studie onder de algemene bevolking leverde een eerste normscore op van zelfgerapporteerde gezondheidsstatus in de Chinese stedelijke gebieden. Deze studie geeft niet alleen inzicht in de variantie van gezondheidsgerelateerde kwaliteit van leven (HRQoL) tussen subgroepen, maar geeft ook referentiewaarden voor ziektelaststudies en interventiestudies. De daaropvolgende methodologische studies geven suggesties voor verbetering van toekomstige waarderingsstudies. Die suggesties zullen zowel het evaluatieonderzoek van de gezondheidszorg (HTA) als het kosten-utiliteitsonderzoek in China en daarbuiten versterken, omdat de EQ-5D de meest gebruikte HRQoL vragenlijst is in beide onderzoeksvelden. Op dit moment heeft de Chinese regering kosten-utiliteitsonderzoek nog niet aangewezen als de basis voor allocatiebeslissingen binnen de gezondheidszorg, maar er zijn sterke signalen dat dit in de toekomst wel zal gebeuren. Zo is bijvoorbeeld in de meeste recente Chinese richtlijn voor farmaco-economische evaluaties te vinden dat kosten-utiliteitsonderzoek de geprefereerde methode is voor economische evaluaties in de gezondheidszorg. Chinese beleidsmakers hebben valide instrumenten nodig om kosten-utiliteitsonderzoek te ondersteunen. Aangezien de EQ-5D wereldwijd het meestgebruikte instrument is voor dit doel, is het ook een goede kandidaat voor China.

Hoofdstuk 2 beschrijft het EQ-5D normscore-onderzoek onder de Chinese stedelijke bevolking. Aanvullende analyses werden uitgevoerd om na te gaan of de zelfgerapporteerde HRQoL varieerde tussen demografische groepen. Hieruit bleek dat HRQoL inderdaad verschilde naar leeftijd, geslacht, opleidingsniveau, vorm van ziektekostenverzekering, het beroep en de geboorteplaats.

In **hoofdstuk 3** zijn de inconsistenties in TTO-responsen gerelateerd aan interview factoren. Hieruit bleek dat de inconsistenties gerelateerd zijn aan verschillen tussen interviewers. Deze resultaten suggereren dat het waarderingsproces beïnvloed wordt door interviewereffecten, wanneer er geen solide kwaliteitscontrole wordt toegepast.

Vanaf hoofdstuk 4 zijn verschillende opzetten voor EQ-5D waarderingsstudies systematisch onderzocht en met elkaar vergeleken. Allereerst is in **hoofdstuk 4** een verzadigde dataset als gouden standaard gebruikt om te onderzoeken welk van twee veelgebruikte manieren om gezondheidstoestanden te selecteren voor een waarderingstaak de beste basis geeft voor het voorspellen van de waarde van gezondheidstoestanden die niet in de waarderingstaak zijn opgenomen. Het betrof het selecteren op basis van het 'voorkomen' van de gezondheidstoestand (de prevalentie), versus de statistische

efficiëntie voor het model. Door de selectie en het modelleerproces te simuleren, bleek dat statistische efficiëntie de belangrijkste overweging moet zijn, terwijl in het verleden vooral gelet werd op het voorkomen. Dit resultaat suggereert dat de selecties van gezondheidstoestanden bij eerdere waarderingsstudies niet optimaal waren, en dat toekomstige waarderingsstudies gebruik kunnen maken van een kleinere selectie van gezondheidstoestanden zolang de statistische doelmatigheid geborgd wordt.

Hoofdstuk 5 beschrijft de constructie van een vergelijkbare verzadigde dataset voor de EQ-5D-5L. 1600 studenten kregen ieder een block van 200 EQ-5D-5L gezondheidstoestanden, die zij waardeerden met behulp van een *visual analogue scale*. Daarnaast beoordeelden zij van elke toestand de geloofwaardigheid. Uit de resultaten bleek dat de respondenten geen overeenstemming hadden over de geloofwaardigheid van de verschillende toestanden. De gemiddelde waarde van een toestand beoordeeld door respondenten die de toestand als 'ongeloofwaardigheid' beoordeelden, was lager dan wanneer de respondenten de toestand 'geloofwaardigheid' vonden, maar het verschil was klein.

Gebruikmakend van de ervaring met de EQ-5D-3L, wordt met het onderzoek uit **hoofdstuk 6** de selectie van het huidige EQ-VT protocol getest en wordt er gekeken of er geen kleinere selecties mogelijk zijn voor EQ-5D-5L waarderingsstudies. De goede prestaties van een selectie gebaseerd op statistische overwegingen werd bevestigd in dit onderzoek rondom de EQ-5D-5L: het orthogonale model met 25 gezondheidstoestanden scoorde net zo goed als de EQ-VT selectie met 86 gezondheidstoestanden, wanneer gekeken wordt naar de precisie waarmee de waarden van alle 3.125 gezondheidstoestanden worden voorspeld. Ondanks dit gunstige resultaat waren er ook zorgen; zo laten de orthogonale modellen grote fouten zien bij het schatten van de milde gezondheidstoestanden. Daarnaast zijn de waarden die verkregen zijn met de *visual analogue scale* mooier verdeeld dan de TTO waarden, wat de generaliseerbaarheid van de bevindingen voor de TTO data in twijfel trekt. Derhalve is in **hoofdstuk 7**, het meest doelmatige model (het orthogonale model) getest in vergelijking met het standaard EQ-VT design met behulp van TTO data. De studie bevestigt dat het ook mogelijk is waarderingen te schatten op basis van geobserveerde waarden voor een selectie van 25 toestanden. Dat geeft gezien over alle gezondheidstoestanden gemiddeld een bijna even voorspelling.

In **hoofdstuk 8** worden de onderzoekvragen uit **hoofdstuk 1** beantwoord en worden gedachten over mogelijk toekomstig onderzoek geformuleerd. Met dit proefschrift wordt gepoogd om de effecten van selectie en modelkeuzen bij Chinese waarderingsstudies

beter te begrijpen. De bevindingen van dit proefschrift kunnen ook worden gegeneraliseerd naar EQ-5D studies in andere landen , of om waarderingsstudies voor andere instrumenten te verbeteren.

CHAPTER 11

Acknowledgments

Undertaking this PhD has been a truly life-changing experience for me and it would not have been possible to do without the support and guidance that I received from many people.

First, I would like to thank my promoter Jan van Busschbach, thank you for putting faith in me all the time. I learned a lot from you, not only about scientific research, but also about life, perspective, food and sports etc. Your constant support is greatly appreciated.

Elly Stolk and Nan Luo, you provided me the help tremendously in conducting the studies and in writing the papers. With your help, I could always stay updated with the latest research. Your patient suggestions in writing papers are invaluable and will continually benefit me in the future.

Thank the colleagues and friends from MPP, I greatly enjoyed my stay in the Netherlands because of you. Special thanks to Hetty Gerritse-Kattouw, Martijn Visser and Fredrick Purba, who provided support in both study and life during my stay; Renier Timman, who helped me a lot with statistical analysis.

Thank the colleagues and friends from the EuroQol Office: Bas Janssen, Mark Oppe, Gouke Bonsel, Juan M. Ramos Goni who helped me in designing study, programming codes, writing papers. Also, very special thanks to Richard Brooks, who helped me in editing the language of each chapter of this thesis.

Thank you, my colleagues in Guizhou Medical University, Tang Lei, Yang Xing, Wu Hongyan, Zhou Zhilin who helped me greatly in several times data collection.

Last but not least, I am extremely grateful to my family. My parents: Yang Hui and Pu Chunmei, my parents in law: Zeng Qingcai and Xie Fenghua. Without your support back at home in every possible way, I would not have the privilege to spare four years for the study. Zeng Tianjiao, my wife, in both good days and bad days, I know you always have my back.

CHAPTER 12

Curriculum Vitae

Zhihao Yang (born 1987, Guiyang, China) holds a B.A. in Health Administration (2006-2010) from Sichuan University, China; and an M.Sc. in Health Service Management (2010-2012) from the University of Sheffield, the United Kingdom. Upon graduation from Sheffield, Zhihao worked in the Health and Medicine Management Faculty, Guizhou Medical University (2012-2014). During that time, Zhihao was recruited as an interviewer for the Chinese EQ-5D-5L valuation project.

In 2014, Zhihao obtained a full scholarship from the China Scholarship Council (CSC) and from 2014 to 2018, he worked on his PhD (improving the use of EQ-5D in China) at the Erasmus University of Rotterdam. During this period, he also obtained an M.Sc. in Health Sciences (2014-2016) from the Netherlands Institute for Health Sciences (NIHES) of the Erasmus University Medical Center and became a EuroQol group member in 2018. In 2018, after finishing his PhD thesis, Zhihao was invited to the National University of Singapore as a visiting scholar to work on several EQ-5D-related projects co-funded by the EuroQol Group and the National Natural Science Foundation (NNSF) of China.

Zhihao returned to China after completing his PhD, and continues to work in this field.

CHAPTER 13

Ph.D. Portfolio

Summary of Ph.D. training and teaching

Name Ph.D. student:	Zhihao Yang
period:	2014-2018
Erasmus MC Department:	Psychiatry
Promotor:	Prof. Dr. J.J. van Busschbach
Research School:	Netherlands Institute of Health Sciences NIHES
Co-Promotor:	Dr. Elly Stolk
	Dr. Nan Luo

Ph.D. training	Year	ECTs
Master of Health science, specialized in Public Health, NIHES	**2014-2016**	
- Public Health Research: from Epidemiology to Health Promotion	2014	5.7
- International Comparison of Health Care Systems	2014	1.4
- Courses for Quantitative Researcher	2014	1.4
- Medical Demography	2015	1.1
- Maternal and Child Health	2015	0.9
- Quality of Life Measurement	2015	0.9
- Patient Preferences in Delivery of Health Care	2015	5.0
- Research Integrity	2015	0.3
- Principles of Research in Medicine	2015	0.7
- Introduction to Global Public Health	2015	0.7
- Methods of Public Health Research	2015	0.7
- Methods of Health Services Research	2015	0.7
- Primary and Secondary Prevention Research	2015	0.7
- Social Epidemiology	2015	0.7
- Study Design	2015	4.3
- Biostatistical Methods I: Basic Principles	2015	5.7
- Biostatistical Methods II: Classical Regression Models	2015	4.3
- Research Proposal	2016	2.5
- Presentation research project	2016	1.4
- Research Paper	2016	29.6

Seminars and workshops, short courses	Year	Days
- Bayesian Analysis: Overview and applications, ISPOR, Amsterdam	2014	0.5
- Discrete Choice Experiments in Health Care, Maastricht University, Maastricht	2015	3
- Systematical literature retrieval in PubMed, Erasmus MC, Rotterdam	2015	1
- Introduction to Health Economic/Pharmacoeconomic Evaluations, ISPOR, Milan	2015	0.5
- Transferability of Cost-Effectiveness Data Between Countries, ISPOR, Milan	2015	0.5
- Health Care Systems in Asia, ISPOR, Singapore	2016	0.5

- Introduction to Patient-Reported Outcomes Assessment: Instrument Development & Evaluation, ISPOR, Vienna	2016	0.5
- Cost-Effectiveness Analysis Alongside Clinical Trials, ISPOR, Vienna	2016	0.5
- Budget Impact Analysis I: A 6-Step Approach, ISPOR, Vienna	2016	0.5
- Budget Impact Analysis II: Applications & Design Issues, ISPOR, Vienna	2016	0.5

Presentations	**Year**	**Hours**
- 'Inconsistency in the valuations of EuroQol EQ-5D-5L health states in China was more related to the interviewer and to interview process than to respondents' characteristics', ISPOR, Milan/ LolaHESG, Maastricht/ EuroQol, Krakow	2015/2015 /2015	1/0.6 /0.3
- 'On a design issue of EQ-5D-3L valuation studies: should likelihood of the health states be considered?' ISPOR, Singapore / Wetenschaplunch, Rotterdam	2016/2016	1/1
- 'Using orthogonal design in selecting health states for the construction of EQ-5D-3L value set', ISPOR, Vienna	2016	1
- 'The use of EQ-5D', South China Pharm-economics Forum, Guangzhou	2017	1
- 'How do university students value EQ-5D-5L health states', Wetenschaplunch, Rotterdam	2017	1
- 'Reducing the costs of EQ-5D-5L valuation studies by valuing a smaller set of health states', LolaHESG, Rotterdam	2017	1
- 'How prevalent are implausible EQ-5D-5L health states and how do they affect valuation? A study combining quantitative and qualitative evidence', EuroQol, Barcelona	2017	0.6
- 'Effect of health state sampling method on design size requirements for EQ-5D-5L valuation studies: small designs can suffice', EuroQol, Barcelona	2017	0.3
- 'Comparing the EQ-5D-5L DCE data in seven Asian countries', EuroQol, Barcelona	2017	0.3
-'Towards a smaller design of EQ-5D-5L valuation study', EuroQol, Budapest	2018	1

International Meetings and conferences	**Year**
- 2014 EuroQol Plenary Meeting, EuroQol, Stockholm	2014
- 1st Meeting of the International Academy of Health Preference Research, IAHPR, Amsterdam	2014
- ISPOR 17th Annual European Congress, ISPOR, Amsterdam	2014
- 2015 Low Lands Health Economic Study Groups Conference, LolaHESG, Maastricht	2015
- 2015 EuroQol Plenary Meeting, EuroQol, Krakow	2015
- ISPOR 18th Annual European Congress, ISPOR, Milan	2015
- ISPOR 7th Asian conference, ISPOR, Singapore	2016
- ISPOR 19th Annual European Congress, ISPOR, Vienna	2016
- 2016 EuroQol Plenary Meeting, EuroQol, Berlin	2016
- South China Pharm-economics Forum 2016, Guangzhou	2016
- 2017 EuroQol Academy Meeting, EuroQol, Noordwijk	2017
- 2017 EuroQol Plenary Meeting, EuroQol, Barcelona	2017
- South China Pharm-economics Forum 2017, Guangzhou	2017
- 2018 EuroQol Academy Meeting, EuroQol, Budapest	2018

CHAPTER 14

References

1. Kennedy-Martin T, Mitchell BD, Boye KS, Chen W, Curtis BH, Flynn JA, et al. The Health Technology Assessment Environment in Mainland China, Japan, South Korea, and Taiwan—Implications for the Evaluation of Diabetes Mellitus Therapies. Value in Health Regional Issues.3:108-16.
2. Zhai T, Goss J, Li J. Main drivers of health expenditure growth in China: a decomposition analysis. BMC Health Serv Res. 2017;17(1):185.
3. Versteegh M. Quality of life in economic evaluations of health Erasmus University Rotterdam; 2014.
4. Szende A, Janssen B, Cabases JM. Self-reported population health: an international perspective based on EQ-5D. New York: Springer; 2014.
5. Abdin E, Subramaniam M, Vaingankar JA, Luo N, Chong SA. Measuring health-related quality of life among adults in Singapore: population norms for the EQ-5D. Qual Life Res. 2013;22(10):2983-91.
6. Jin XJ, Liu GG, Luo N, Li HC, Guan HJ, Xie F. Is bad living better than good death? Impact of demographic and cultural factors on health state preference. qual Life Res. 2016;25(4):979-86.
7. Wang P, Li MH, Liu GG, Thumboo J, Luo N. Do Chinese have similar health-state preferences? A comparison of mainland Chinese and Singaporean Chinese. Eur J Health Econ. 2015;16(8):857-63.
8. Karimi M, Brazier J, Paisley S. How do individuals value health states? A qualitative investigation. Soc Sci Med. 2017;172:80-8.
9. Luo N, Liu G, Li M, Guan H, Jin X, Rand-Hendriksen K. Estimating an EQ-5D-5L Value Set for China. Value Health. 2017;20(4):662-9.
10. Li H, Wei X, Ma A, Chung RY. Inequalities in health status among rural residents: EQ-5D findings from household survey China. International Journal for Equity in Health. 2014;13:41.
11. Sun S, Chen J, Johannesson M, Kind P, Xu L, Zhang Y, et al. Regional differences in health status in China: population health-related quality of life results from the National Health Services Survey 2008. Health Place. 2011;17(2):671-80.
12. Zhang T, Shi WX, Huang ZQ, Gao D, Guo ZY, Liu JY, et al. Influence of culture, residential segregation and socioeconomic development on rural elderly health-related quality of life in Guangxi, China. Health Qual Life Out. 2016;14.
13. Dong WL, Li YC, Wang ZQ, Jiang YY, Mao F, Qi L, et al. Self-rated health and health-related quality of life among Chinese residents, China, 2010. Health Qual Life Out. 2016;14.
14. Purba FD, Hunfeld JA, Iskandarsyah A, Fitriana TS, Sadarjoen SS, Passchier J, et al. Employing quality control and feedback to the EQ-5D-5L valuation protocol to improve the quality of data collection. Qual Life Res. 2016.
15. Yang Z, Luo N, Bonsel G, Busschbach J, Stolk E. Selecting Health States for EQ-5D-3L Valuation Studies: Statistical Considerations Matter. Value Health. 2018;21(4):456-61.
16. Augustovski F, Rey-Ares L, Irazola V, Garay OU, Gianneo O, Fernandez G, et al. An EQ-5D-5L value set based on Uruguayan population preferences. Qual Life Res. 2016;25(2):323-33.
17. M MV, K MV, S MAAE, de Wit GA, Prenger R, E AS. Dutch Tariff for the Five-Level Version of EQ-5D. Value Health. 2016;19(4):343-52.
18. Xie F, Pullenayegum E, Gaebel K, Bansback N, Bryan S, Ohinmaa A, et al. A Time Trade-off-derived Value Set of the EQ-5D-5L for Canada. Med Care. 2016;54(1):98-105.
19. Kim SH, Ahn J, Ock M, Shin S, Park J, Luo N, et al. The EQ-5D-5L valuation study in Korea. Qual Life Res. 2016;25(7):1845-52.
20. Shiroiwa T, Ikeda S, Noto S, Igarashi A, Fukuda T, Saito S, et al. Comparison of Value Set Based on DCE and/or TTO Data: Scoring for EQ-5D-5L Health States in Japan. Value Health. 2016;19(5):648-54.
21. Devlin N, Shah K, Feng Y, Mulhern B, Hout B. Valuing health-related quality of life: An EQ-5D-5L value set for England. Office of Economics Research paper. 2016.
22. Purba FD, Hunfeld JAM, Iskandarsyah A, Fitriana TS, Sadarjoen SS, Ramos-Goni JM, et al. The Indonesian EQ-5D-5L Value Set. Pharmacoeconomics. 2017;35(11):1153-65.
23. Rand-Hendriksen K, Augestad LA, Kristiansen IS, Stavem K. Comparison of hypothetical and experienced EQ-5D valuations: relative weights of the five dimensions. Qual Life Res. 2012;21(6):1005-12.
24. Luo N, Li M, Liu GG, Lloyd A, de Charro F, Herdman M. Developing the Chinese version of the new 5-level EQ-5D descriptive system: the response scaling approach. Qual Life Res. 2013;22(4):885-90.

25. Feng Y, Devlin N, Herdman M. Assessing the health of the general population in England: how do the three- and five-level versions of EQ-5D compare? Health Qual Life Outcomes. 2015;13:171.
26. Janssen MF, Pickard AS, Golicki D, Gudex C, Niewada M, Scalone L, et al. Measurement properties of the EQ-5D-5L compared to the EQ-5D-3L across eight patient groups: a multi-country study. Qual Life Res. 2013;22(7):1717-27.
27. Kim TH, Jo MW, Lee SI, Kim SH, Chung SM. Psychometric properties of the EQ-5D-5L in the general population of South Korea. Qual Life Res. 2013;22(8):2245-53.
28. Oppe M, Devlin NJ, van Hout B, Krabbe PFM, de Charro F. A Program of Methodological Research to Arrive at the New International EQ-5D-5L Valuation Protocol. Value in Health. 2014;17(4):445-53.
29. Oppe M, Rand-Hendriksen K, Shah K, Ramos-Goni JM, Luo N. EuroQol Protocols for Time Trade-Off Valuation of Health Outcomes. Pharmacoeconomics. 2016;34(10):993-1004.
30. Golicki D, Niewada M. EQ-5D-5L Polish population norms. Archives of Medical Science. 2015;13(1):191-200.
31. Pan CW, Sun HP, Wang XZ, Ma QH, Xu Y, Luo N, et al. The EQ-5D-5L index score is more discriminative than the EQ-5D-3L index score in diabetes patients. Quality of Life Research. 2015;24(7):1767-74.
32. Scalone L, Cortesi PA, Ciampichini R, Cesana G, Mantovani LG. Health Related Quality of Life norm data of the general population in Italy: results using the EQ-5D-3L and EQ-5D-5L instruments. Epidemiol Biostat Pu. 2015;12(3).
33. Fayers PM, Machin D. Quality of life: The assessment, analysis and interpretation of patient-reported outcomes. England: Wiley; 2007.
34. Rand-Hendriksen K, Ramos-Goni JM, Augestad LA, Luo N. Less Is More: Cross-Validation Testing of Simplified Nonlinear Regression Model Specifications for EQ-5D-5L Health State Values. Value Health. 2017;20(7):945-52.
35. Shiroiwa T, Fukuda T, Ikeda S, Igarashi A, Noto S, Saito S, et al. Japanese population norms for preference-based measures: EQ-5D-3L, EQ-5D-5L, and SF-6D. Qual Life Res. 2016;25(3):707-19.
36. Hinz A, Kohlmann T, Stobel-Richter Y, Zenger M, Brahler E. The quality of life questionnaire EQ-5D-5L: psychometric properties and normative values for the general German population. Qual Life Res. 2014;23(2):443-7.
37. Garcia-Gordillo MA, Adsuar JC, Olivares PR. Normative values of EQ-5D-5L: in a Spanish representative population sample from Spanish Health Survey, 2011. Qual Life Res. 2016;25(5):1313-21.
38. McCaffrey N, Kaambwa B, Currow DC, Ratcliffe J. Health-related quality of life measured using the EQ-5D-5L: South Australian population norms. Health Qual Life Out. 2016;14.
39. Sun S, Chen J, Johannesson M, Kind P, Xu L, Zhang Y, et al. Population health status in China: EQ-5D results, by age, sex and socio-economic status, from the National Health Services Survey 2008. Qual Life Res. 2011;20(3):309-20.
40. Liu GG, Wu H, Li M, Gao C, Luo N. Chinese time trade-off values for EQ-5D health states. Value Health. 2014;17(5):597-604.
41. Visser M, Verbaan D, van Rooden S, Marinus J, van Hilten J, Stiggelbout A. A longitudinal evaluation of health-related quality of life of patients with Parkinson's disease. Value Health. 2009;12(2):392-6.
42. Hajek A, Brettschneider C, Mallon T, Ernst A, Mamone S, Wiese B, et al. The impact of social engagement on health-related quality of life and depressive symptoms in old age - evidence from a multicenter prospective cohort study in Germany. Health Qual Life Outcomes. 2017;15(1):140.
43. Alva M, Gray A, Mihaylova B, Clarke P. The effect of diabetes complications on health-related quality of life: the importance of longitudinal data to address patient heterogeneity. Health Econ. 2014;23(4):487-500.
44. Herdman M, Gudex C, Lloyd A, Janssen M, Kind P, Parkin D, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual Life Res. 2011;20(10):1727-36.
45. Devlin NJ, Hansen P, Kind P, Williams A. Logical inconsistencies in survey respondents' health state valuations -- a methodological challenge for estimating social tariffs. Health Econ. 2003;12(7):529-44.
46. Badia X, Roset M, Herdman M. Inconsistent responses in three preference-elicitation methods for health states. Soc Sci Med. 1999;49(7):943-50.

47. Tongsiri S, Cairns J. Estimating population-based values for EQ-5D health states in Thailand. Value Health. 2011;14(8):1142-5.

48. Ohinmaa A, Sintonen H. Inconsistencies and modelling of the Finnish EuroQol (EQ-5D) preference values. EuroQol Plenary Meeting 1998 Discussion Papers. 1999:57-73.

49. Lamers LM, Stalmeier PF, Krabbe PF, Busschbach JJ. Inconsistencies in TTO and VAS values for EQ-5D health states. Med Decis Making. 2006;26(2):173-81.

50. Dolan P, Kind P. Inconsistency and health state valuations. Soc Sci Med. 1996;42(4):609-15.

51. Mulhern B, Shah KK, janssen B, Longworth L. Valuing EQ-5D-5L using TTO and DCE: Does dimension order impact on health state values? EuroQol Working Paper Series. 2015.

52. Shah KK, Mulhern B, Longworth L, Janssen B. An empirical study of two alternative comparators for use in time trade-off studies. EuroQol Working Paper Series. 2015.

53. Oppe M, Hout B. The 'power' of eliciting EQ-5D-5L values. EuroQol Working Paper Series. 2017.

54. Janssen BM, Oppe M, Versteegh MM, Stolk EA. Introducing the composite time trade-off: a test of feasibility and face validity. Eur J Health Econ. 2013;14 Suppl 1:S5-13.

55. Kularatna S, Whitty JA, Johnson NW, Jayasinghe R, Scuffham PA. Valuing EQ-5D health states for Sri Lanka. Qual Life Res. 2014.

56. Singer JD WJ. Applied longitudinal data analysis - modeling change and event occurrence. Oxford: Oxford University Press; 2003.

57. Ramos-Goni JM, Oppe M, Slaap B, Busschbach J, Stolk E. Quality control process for EQ-5D-5L valuation studies. Value Health. 2017;Epub ahead of print.

58. Craig BM, Ramachandran S. Relative risk of a shuffled deck: a generalizable logical consistency criterion for sample selection in health state valuation studies. Health Econ. 2006;15(8):835-48.

59. Norman R, Cronin P, Viney R, King M, Street D, Ratcliffe J. International Comparisons in Valuing EQ-5D Health States: A Review and Analysis. Value in Health. 2009;12(8):1194-200.

60. Lamers LM, McDonnell J, Stalmeier PF, Krabbe PF, Busschbach JJ. The Dutch tariff: results and arguments for an effective design for national EQ-5D valuation studies. Health Econ. 2006;15(10):1121-32.

61. Ageing. DoHa. Guidelines for preparing submissions to the pharmaceutical benefits advisory committee Canberra2016 [updated September 2016; cited 2017 June 19]. Version 5.0:[Available from: https://pbac.pbs.gov.au/.

62. Gold M. Panel on cost-effectiveness in health and medicine. Med Care. 1996;34(12 Suppl):DS197-9.

63. Claxton K, Sculpher M, Drummond M. A rational framework for decision making by the National Institute For Clinical Excellence (NICE). Lancet. 2002;360(9334):711-5.

64. Taylor R. Using health outcomes data to inform decision-making - Government agency perspective. Pharmacoeconomics. 2001;19:33-8.

65. Green PE. On the design of choice experiments involving multifactor alternatives. Journal of Consumer Research. 1974;1(2):61-8.

66. Fedorov V. Theory of Optimal Experiments Designs. New York: Academic Press; 1972.

67. Bagust A. Improving valuation sampling of EQ-5D health states. Health Qual Life Outcomes. 2013;11:14.

68. Dolan P. Modeling valuations for EuroQol health states. Med Care. 1997;35(11):1095-108.

69. Bonsel G, Oppe M, janssen B. Unexpected large misspecification effects of health profiles selection and interaction analysis to obtain a value function from unsaturated valuation datasets, using the standard EuroQol approach. EuroQol Plenary Meeting 2014 Discussion Papers2014.

70. Busschbach J, McDonnell J, Hout B. Testing different parametric relations between the EuroQol health description and health valuations in students. EuroQol Plenary Meeting 1996 Discussion Papers. 1996.

71. Busschbach JJV, Wolffenbuttel BHR, Annemans L, Meerding WJ, Koltowska-Haggstrom M. Deriving reference values and utilities for the QoL-AGHDA in adult GHD. European Journal of Health Economics. 2011;12(3):243-52.

72. Stolk E, Krabbe P, Busschbach J. Using the Internet to collect EQ-5D norm scores: a valid alternative? EuroQol Plenary Meeting 2009 Discussion Papers. 2009.

73. Badia X, Roset M, Herdman M, Kind P. A comparison of United Kingdom and Spanish general population time trade-off values for EQ-5D health states. Med Decis Making. 2001;21(1):7-16.

74. Shaw JW, Johnson JA, Coons SJ. US valuation of the EQ-5D health states: development and testing of the D1 valuation model. Med Care. 2005;43(3):203-20.

75. Tsuchiya A, Ikeda S, Ikegami N, Nishimura S, Sakai I, Fukuda T, et al. Estimating an EQ-5D population value set: the case of Japan. Health Econ. 2002;11(4):341-53.

76. Chevalier J, de Pouvourville G. Valuing EQ-5D using time trade-off in France. Eur J Health Econ. 2013;14(1):57-66.

77. Sloane NJA. A library of Orthogoal Arrays 2017 [cited 2017 15-02]. Available from: http://neilsloane.com/oadir/index.html.

78. Hedayat A, Sloane NJA, Stufken J. Orthogonal Arrays: Theory and Applications. New York, USA: Springe-Verlag; 1999.

79. Coretti S, Ruggeri M, McNamee P. The minimum clinically important difference for EQ-5D index: a critical review. Expert Rev Pharmacoecon Outcomes Res. 2014;14(2):221-33.

80. Pickard AS, Neary MP, Cella D. Estimation of minimally important differences in EQ-5D utility and VAS scores in cancer. Health Qual Life Outcomes. 2007;5:70.

81. Kim SK, Kim SH, Jo MW, Lee SI. Estimation of minimally important differences in the EQ-5D and SF-6D indices and their utility in stroke. Health Qual Life Outcomes. 2015;13:32.

82. Craig BM, Busschbach JJ, Salomon JA. Modeling ranking, time trade-off, and visual analog scale values for EQ-5D health states: a review and comparison of methods. Med Care. 2009;47(6):634-41.

83. Badia X, Herdman M, Roset M, Ohinmaa A. Feasibility and validity of the VAS and TTO for eliciting general population values for temporary health states: a comparative study. Health Services & Outcomes Research Methodology. 2001;2:51-65.

84. Badia X, Monserrat S, Roset M, Herdman M. Feasibility, validity and test-retest reliability of scaling methods for health states: the visual analogue scale and the time trade-off. Qual Life Res. 1999;8(4):303-10.

85. Johnson FR, Hauber AB, Osoba D, Hsu MA, Coombs J, Copley-Merriman C. Are chemotherapy patients' HRQoL importance weights consistent with linear scoring rules? A stated-choice approach. Qual Life Res. 2006;15(2):285-98.

86. Goodwin E, Boddy K, Tatnell L, Hawton A. Involving Members of the Public in Health Economics Research: Insights from Selecting Health States for Valuation to Estimate Quality-Adjusted Life-Year (QALY) Weights. Appl Health Econ Health Policy. 2018;16(2):187-94.

87. Viney R, Savage E, Louviere J. Empirical investigation of experimental design properties of discrete choice experiments in health care. Health Econ. 2005;14(4):349-62.

88. Bonsel G, Oppe M, Janssen M. Unlikely health states: evidence from healthy and diseased populations. EuroQol Plenary Meeting 2015 Discussion Papers2015.

89. Yang Z, Luo N, Busschbach J, Stolk E. On a design issue of EQ-5D-3L valuation studies: should likelihood of the health states be considered? Value Health. 2016;19(7):A853-A4.

90. Yang Z, Luo N, Busschbach J, Stolk E. Using orthogonal design in selecting health states for the construction of EQ-5D-3L value set. Value Health. 2016;19(7):A386.

91. Robinson A, Dolan P, Williams A. Valuing health status using VAS and TTO: What lies behind the numbers? Social Science & Medicine. 1997;45(8):1289-97.

92. Sun S, Chen J, Kind P, Xu L, Zhang Y, Burstrom K. Experience-based VAS values for EQ-5D-3L health states in a national general population health survey in China. Qual Life Res. 2015;24(3):693-703.

93. Bernert S, Fernandez A, Haro JM, Konig HH, Alonso J, Vilagut G, et al. Comparison of different valuation methods for population health status measured by the EQ-5D in three European countries. Value Health. 2009;12(5):750-8.

94. Kuhfeld WF, Tobias RD, Garratt M. Efficient Experimental-Design with Marketing-Research Applications. J Marketing Res. 1994;31(4):545-57.

95. Coteur G, Feagan B, Keininger DL, Kosinski M. Evaluation of the meaningfulness of health-related quality of life improvements as assessed by the SF-36 and the EQ-5D VAS in patients with active Crohn's disease. Aliment Pharmacol Ther. 2009;29(9):1032-41.

96. Feng Y, Devlin NJ, Shah KK, Mulhern B, van Hout B. New methods for modelling EQ-5D-5L value sets: An application to English data. Health Econ. 2017:16.

97. Devlin NJ, Shah KK, Feng Y, Mulhern B, van Hout B. Valuing health-related quality of life: An EQ-5D-5L value set for England. Health Econ. 2017:16.

98. Yang Z, Luo N, Oppe M, Bonsel G, Busschbach J, Stolk E. The effect of health state sampling methods on model predictions of EQ-5D-5L values: small designs can suffice. Manuscript Submitted for Publication. 2018.

99. Stolk E, Ludwig K, Rand-Hendriksen K, Van Hout B, Ramos-Goni JM. Overview, update and lessons learned from the international EQ-5D-5L valuation work: version 2 of the EQ-5D-5L valuation protocol. Manuscript Submitted for Publication2018.

100. Lien K, Tam VC, Ko YJ, Mittmann N, Cheung MC, Chan KK. Impact of country-specific EQ-5D-3L tariffs on the economic value of systemic therapies used in the treatment of metastatic pancreatic cancer. Curr Oncol. 2015;22(6):e443-52.

101. Ramos-Goni JM, Pinto-Prades JL, Oppe M, Cabases JM, Serrano-Aguilar P, Rivero-Arias O. Valuation and Modeling of EQ-5D-5L Health States Using a Hybrid Approach. Med Care. 2017;55(7):e51-e8.

102. Brazier J, Roberts J, Tsuchiya A, Busschbach J. A comparison of the EQ-5D and SF-6D across seven patient groups. Health Econ. 2004;13(9):873-84.

103. Jiang W. The relationship between culture and language. ELT Journal. 2000;54(4):328-34.

104. The World Health Organization quality of life assessment (WHOQOL): Position paper from the World Health Organization. Social Science & Medicine. 1995;41(10):1403-9.

105. Collings JA. International differences in psychosocial well-being: a comparative study of adults with epilepsy in three countries. Seizure. 1994;3(3):183-90.

106. Kagawa-Singer M, Padilla GV, Ashing-Giwa K. Health-Related Quality of Life and Culture. Seminars in Oncology Nursing. 2010;26(1):59-67.

107. Suzukamo Y, Fukuhara S, Green J, Kosinski M, Gandek B, Ware JE. Validation testing of a three-component model of Short Form-36 scores. Journal of clinical epidemiology. 2011;64(3):301-8.

108. Fuh JL, Wang SJ, Lu SR, Juang KD, Lee SJ. Psychometric evaluation of a Chinese (Taiwanese) version of the SF-36 health survey amongst middle-aged women from a rural community. Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation. 2000;9(6):675-83.

109. Thumboo J, Fong KY, Machin D, Chan SP, Leon KH, Feng PH, et al. A community-based study of scaling assumptions and construct validity of the English (UK) and Chinese (HK) SF-36 in Singapore. Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation. 2001;10(2):175-88.

110. Kim M, Han HR, Phillips L. Metric equivalence assessment in cross-cultural research: using an example of the Center for Epidemiological Studies--Depression Scale. Journal of nursing measurement. 2003;11(1):5-18.

111. Losada A, de los Angeles Villareal M, Nuevo R, Marquez-Gonzalez M, Salazar BC, Romero-Moreno R, et al. Cross-cultural confirmatory factor analysis of the CES-D in Spanish and Mexican dementia caregivers. The Spanish journal of psychology. 2012;15(2):783-92.

112. Zhang B, Fokkema M, Cuijpers P, Li J, Smits N, Beekman A. Measurement invariance of the Center for Epidemiological Studies Depression Scale (CES-D) among Chinese and Dutch elderly. BMC medical research methodology. 2011;11:74.

113. Bullinger M, Alonso J, Apolone G, Leplege A, Sullivan M, Wood-Dauphinee S, et al. Translating health status questionnaires and evaluating their quality: the IQOLA Project approach. International Quality of Life Assessment. J Clin Epidemiol. 1998;51(11):913-23.

114. Rabin R, de Charro F. EQ-5D: a measure of health status from the EuroQol Group. Ann Med. 2001;33(5):337-43.

115. Keeley T, Al-Janabi H, Lorgelly P, Coast J. A qualitative assessment of the content validity of the ICECAP-A and EQ-5D-5L and their appropriateness for use in health research. PLoS One. 2013;8(12):e85287.

116. Matza LS, Boye KS, Stewart KD, Curtis BH, Reaney M, Landrian AS. A qualitative examination of the content validity of the EQ-5D-5L in patients with type 2 diabetes. Health Qual Life Outcomes. 2015;13:192.

117. van Leeuwen KM, Jansen AP, Muntinga ME, Bosmans JE, Westerman MJ, van Tulder MW, et al. Exploration of the content validity and feasibility of the EQ-5D-3L, ICECAP-O and ASCOT in older adults. BMC Health Serv Res. 2015;15:201.

118. Health USDo, Human Services FDACfDE, Research, Health USDo, Human Services FDACfBE, Research, et al. Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims: draft guidance. Health Qual Life Outcomes. 2006;4:79.

119. PubMed. China+Health outcome PubMed2018 [23rd May]. Available from: https://www.ncbi.nlm.nih.gov/pubmed/?term=china+health+outcome.

120. Cheung YB, Thumboo J. Developing health-related quality-of-life instruments for use in Asia: the issues. Pharmacoeconomics. 2006;24(7):643-50.

121. Oppe M. Mathematical approaches in economic evaluations: Erasmus University Rotterdam; 2013.

122. Viney R, Norman R, Brazier J, Cronin P, King MT, Ratcliffe J, et al. An Australian discrete choice experiment to value eq-5d health states. Health Econ. 2014;23(6):729-42.