The Thai population-based preference scores for EQ-5D health states

Dr. Sirinart Tongsiri

August 2009

The fieldwork survey of this project is financially supported by the Burden of Disease (BOD) Program, the Health Intervention and Technology Assessment Program (HITAP) and the International Health Policy Program (IHPP)

Ministry of Public Health, Thailand

Executive summary

Health care expenditures have been increasing rapidly. Economic evaluation can be used to aid decision making on resource allocations to secure a more efficient use of scarce resources. In cost-utility analysis, one method used to measure health outcomes is the Quality adjusted life year (QALY). Given the wide differences in clinical settings, health systems and religious beliefs, "utility" scores should be derived from the local population. This report aims to estimate population-based preference scores for health from the Thai general population. The generic health description EQ-5D is used as a proxy to describe health. This measure was selected because it has been translated officially into Thai and the measure seems to be straightforward to use. A representative sample was randomly recruited using a stratified four-stage sampling method. A series of pilot studies were conducted to develop the interview protocol based on the Measurement and Valuation in Health (MVH) protocol. A group of interviewers were employed and extensively trained to interview the respondents.

A sample of 1,409 Thai respondents was interviewed during May – August 2007 in 17 provinces in face-to-face interviews. Eighty-six health states, classified into twelve sets, were used in the interview. Logical inconsistency was identified when a higher score was given to a poorer state. The greatest number of inconsistent responses was identified in the scores derived using the Time trade-off (TTO) interview. A Random effects model was used to estimate the model to predict the preference scores. The best model was chosen on the basis of logical inconsistency in the predicted scores, model robustness, parsimony and the responsiveness of the predicted scores. The best model is the model using the variables from Dolan 1997 model estimated from the scores given by the respondents with fewer than 11 inconsistencies. The model still suffers from heteroskedasticity, and floor and ceiling effects were identified. The second best score is 0.766 for state 11112 and the poorest score is -0.454 for state 33333.

The report makes a number of contributions. The modelled scores are the first original population-based preference scores on health derived from the Thai general population. An exploratory qualitative interview was undertaken to learn the strategies that respondents employed to cope with the preference interview. Three reasons are identified to explain the high level of inconsistent responses. Respondents may: (1)

have difficulties imagining themselves living in the hypothetical states; (2) use only part of the given information in the health cards or add other information to assist their decisions; and (3) have difficulties in trying to understand the elicitation methods, especially the TTO. Including the inconsistent responses had, to some extent, significant impacts on the model specifications and the modelled scores. Exclusion of the scores from the highly inconsistent respondents was justified because the scores may not represent their preferences towards health. The results from this thesis should be taken into account for future surveys to be successfully administered. Close collaborations with the field coordinators and arrangement of appropriate interview settings contribute greatly to the success of the survey.

Acknowledgments

This report is part of the research I pursued during my PhD at London School of Hygiene and Tropical Medicine, University of London, from 2005 to 2009. I would like to show my gratitude to my supervisor, Professor John Cairns, for making my study at the school to be a precious memory of my life. Doing a PhD with John is a treasure. His vital encouragement, supports, patient and understandings from the beginning to the final stages of my study enabled me to develop a great deal of knowledge and accomplishments in the research.

I would like to acknowledge and extend my heartfelt gratitude to the following persons who have been supporting me through the process of doing a PhD: my family for their moral supports and understandings and their helps in every part of my life, Pawasoot and Nopporn for always being there to listen to all my difficulties.

I am indebted to my research assistants: Sirinya, Artidtaya and the researchers in IHPP, BOD and HITAP, all field coordinators both in the pre-test studies and the fieldwork survey, and interviewers who made extraordinary journey of doing this research possible. I would like to thank especially to Dr. Chulaporn Limwattananon and her students from the Faculty of Pharmacy, Khon Kaen University for untiring helps in conducting the interview in the respondents throughout the Northeast region. I would like to express my appreciations to all respondents and their families for sacrificing their valuable times to participate in the survey. I have learned a lot from them and without them, this thesis would not be accomplished.

I am very grateful not only to the financial support of the fieldwork survey but also intellectual and moral supports from scholars in Thailand: Dr. Kanitta from the Burden of Disease Project (BOD), Dr. Viroj, Dr. Phusit, Dr. Sripen and Dr. Walaiporn from the International Health Policy Program (IHPP) and Dr. Yot from the Health Impacts and Technology Assessment Program (HITAP) and all the staffs. I would like to thank for the helpful collaboration with the staffs of the National Statistical Office (NSO) Thailand for assisting with sharing the sample in this study and the Health and Welfare and the Disabilities Survey 2007.

Regarding my study in the UK, the tuition fees and the monthly expenditures were covered by the Royal Thai Government scholarship. A big thank to all staffs, especially, Khun Jitree, Khun Vorapoj and Khun Natee at the Office of Educational Affairs, Royal Thai Embassy, London. Their helps and hard work enabled me to make the most of my time doing a PhD in London without financially worries and all those supports for essential government documents.

My sincere thanks to all friends and colleagues in the UK, especially, Mariana, Paco, Ayako, Nareerut, Saowaluck and Inthira, for making my life in the UK feels like home away from home and for constant reminders and much needed motivation. I wish to extend my thanks to my friends and colleagues in Thailand- for their friendships, support and inspirations, especially, Weerachai for being the main supplier of articles to which I cannot access and Lertrit for a moral support when I have writing-up crises. Finally, I would also like to extend my gratitude to the PHP IT services: Mick and Caroline, all HSRU staffs and all LSHTM library staffs for their constantly supports.

Table of Contents

List of Tables

List of Figures

List of abbreviations

BOD	The Burden of Disease project
CAPD	Continuous ambulatory peritoneal dialysis
CUA	Cost-utility analysis
ES	Effect size
FE	Fixed effects model
FGLS	Feasible generalised least square
HD	Hemodialysis
HITAP	The Health Impact and Technology Assessment Program
HWS	Health and Welfare Survey and the Disabilities Survey
ICER	Incremental cost-effectiveness ratio
IHPP	The International Health Policy Program
LM	Lagrange Multiplier test
MAD	Mean absolute difference
MAU	Multi-attribute utility function
MVH	The Measurement and Valuation of Health study
NICE	The National Institute for Clinical Excellence
NSO	The National Statistical Office
OLS	Ordinary least squared model
QALY	Quality-adjusted Life Year
RE	Random effects model
RMSE	Root mean squared errors
SI	Statistical inference
SE	Standard errors
SD	Standard deviation
SRM	standardized responses mean
TTO	Time Trade-Off
UC	Universal Coverage scheme
UK	United Kingdom
US	United States
VAS	Visual Analog Scale

Chapter 1 Introduction

There is a wide acceptance that market failure exists in the system of health services. Optimal resource utilisation in health cannot be determined by demand and supply as in a perfectly competitive market(1). There was also an increasing demand for a more transparent and participatory decision making process for the allocation of resources to health (2). One of the tools used to aid efficient resource allocation across different health intervention is economic evaluation which is defined as “a comparative study of alternative interventions in terms of their costs and benefits” (3). Economic evaluations are widely used to aid resource allocation decision making (4). The benefits of economic evaluations have been recognised in several countries. In Australia, economic evaluation is legally required for new pharmaceutical products for listing on the Pharmaceutical Benefits Scheme for being subsidized by the Government(5). Economic appraisal is used by the Dutch Health Insurance Executive Board to decide the health insurance packages (5). Johannesson concluded that economic evaluation is useful in the development of treatment guidelines and reimbursement decisions of medical technologies. (6).

A need for economic evaluation is emerging in Thailand to provide evidence of costs and benefits of medical interventions explicitly for policy makers. Previously, it was mainly the academics in universities who conducted economic evaluation studies. There have eventually been many attempts to conduct economic evaluation by Thai researchers and indeed, a number of economic evaluations have been performed in Thailand (7-10). In 2004, Thailand introduced explicit criteria for decision making on cost and efficiency criteria in the revision of the National List of Essential Drugs in 2004 (11). Chiawchanwattana *et al*. (7) and Teerawattananon (8) performed cost-effectiveness and cost-utility analyses, comparing hemodialysis (HD) and continuous ambulatory peritoneal dialysis (CAPD), with palliative care in Thai patients with end-stage renal failure. It was not until 2006 when the Health Intervention and Technology Assessment Program (HITAP) was established to provide the standard methodological guidelines for economic evaluation in Thailand and the health technology assessment database(12).

In economic evaluation, health benefits are measured in several formats. The evaluation of health benefits includes several types of analyses according to how the outcomes of intervention are measured against their costs. Types of economic evaluation include cost-minimisation, cost-effectiveness, cost-utility and cost-benefit analyses(3). In cost-utility analysis, health outcome is measured in Quality-adjusted life years (QALYs) using both quantitative (years of life living in a particular health state) and qualitative (utility of being in the state) given by individuals on the 0-1 scale where 0 represents death and 1 represents full health (13). In this method, QALYs are assumed to be a cardinal measure and be interpersonally comparable regardless of which type of health interventions are given to an individual. Drummond suggested that given that there are differences in clinical practices and health services organisations in any health settings, for a cost-utility study to be used as a tool for resource allocation decision making in a particular setting, it should be undertaken "using local data" (14). Badia *et al.* also supported this statement (15). The health state valuations should ideally be relevant to the populations under study in order that the results of the analysis are applicable to their own settings. In the UK, the National Institute of Health and Clinical Excellence (NICE) recommends that the valuation of health states should be performed using a generic health outcome measure for which preference scores are elicited using the time trade-off or standard gamble methods from a UK community sample (16). The US panel on Cost-Effectiveness in Health and Medicine recommended that the 'reference case' should be used as a standard methodological practice in order to increase the comparability of economic evaluation results and health outcome should be weighed by a representative, community-based sample using a generic health outcome measure (2).

As previously mentioned a number of economic evaluations have been conducted in Thailand, however, to measure utility of the Thai patients, the researchers have used health outcome preferences obtained either from a group of patients or studies from other countries (7-9). Although there is one study estimating the Thai algorithm to predict preference scores for the EQ-5D health states, the sample used in the study was not representative of the Thai general population (17). Preferences over health outcomes in that study were derived from health professionals, bronchitis and cancer patients in a hospital in Bangkok and a non-probability sample of healthy people in the hospital neighbourhood. Both postal survey and face-to-face interview were conducted using 15 health states including unconscious and death. Preference scores elicited from a representative sample of the Thai general population have not yet been established.

1.1 Research objectives

This study aims primarily to estimate preference scores for health derived from the Thai general population. The scores are expected to be applicable to measure QALYs in cost-utility analysis in Thailand. The followings are the specific objectives to be fulfilled before the primary objective can be achieved.

- Describe an appropriate health descriptive measure and the preference elicitation methods to be used to elicit preference scores
- Plan and carry out a large scale survey of health preferences
- Conduct model specifications to estimate the scores for unobserved health states

1.2 Outline of the report

This document reports the Thai population-based preference scores for EQ-5D health states. The scores can be used to calculate the Quality-adjusted life years (QALYs) in economic evaluation of health interventions in the Thai settings. The fieldwork survey of this project is financially supported by the Burden of Disease (BOD) project and the Health Impact and Technology Assessment program (HITAP). In the next section, a brief report of health state measure and the preference elicitation method used in this study is presented. TheEQ-5D is selected as the most appropriate measure to describe health in this study because the measure is widely used in economic evaluations worldwide and was officially translated into Thai. An additional advantage is that there are a considerable number of countries using the EQ-5D to derive preference scores for health from their own general population. Thus, there is a good opportunity for lessons learnt from conducting the previous surveys to be implemented in the Thai study. The seminal Measurement and Valuation in Health study (MVH) methodology is used as a prototype for the Thai study design. A series of decisions on the appropriate number of health states and the interview props used in the interview had been made based on the feasibility of implementing the methods in the Thai general population. As is the case in the previous studies, it is impossible to ask a Thai respondent to assign scores for all 243 EQ-5D states. The previous model specifications to estimate the scores are reviewed in Chapter 5 and used to estimate scores for the unobserved health states.

The sample size calculation and the random selection of the representative sample were undertaken collaboratively with the National Statistical Office (NSO), Thailand. The fieldwork survey of the Thai study was conducted in parallel with the Health and Welfare (HWS) survey in 2005 which was a good opportunity for the study to share part of the sample with the HWS survey. Several pilot studies were administered to design the feasible fieldwork survey. To familiarise the researcher with the preference elicitation interview, a pilot study was begun in London with Thai PhD students. Another two pilot studies were conducted, firstly, with the staffs of the funding organisations and, secondly, in a convenience sample whose characteristics were similar to those of the Thai general population. The results of these activities and the reasons behind decisions regarding the Thai interview protocol are reported in Chapter 2. A group of interviewers were recruited to help with the preference interview with the representative sample. Because of the complications of the protocol and to control the quality of the interview, the intensive interviewer training programs were organised in parallel with the second pilot study and the interview was performed with the convenience sample. In the same process, the interview props were developed and the health states for use in the interview were selected. A final version of the interview protocol, health states and props are reported in this chapter.

The Thai study involved face-to-face interviews, therefore, the research team planned access to the respondents with the collaboration of the staff of the provincial health office who are very knowledgeable regarding the location of the targeted respondents. Results of the fieldwork survey, including the demographic characteristics of the Thai respondents including, their own health in the past 24 hours, the overall interview duration and the durations according to the specific interview methods, and the mean scores for the health states used in the interview are reported in Chapter 3. By conducting the interview, more insights regarding the increased cognitive overloads of the respondents were gained and the numbers of inconsistent responses were closely related with the interview sites. To realise the feasibility of the interview tasks, the respondents were requested to comment on the difficulties of the tasks and the interviewers were instructed to give their impressions on respondents' performance while participating in the interview. The nature of the actual scores is thoroughly explored and logically inconsistent responses are addressed and further investigations are performed in Chapter 4.

Logical inconsistency arises when higher scores are given to poorer health states. This issue is interesting because to assign values to health outcomes, it is assumed that individuals are the best judges of their own utility and they are assumed to prefer better health. However, logical inconsistency is identified in the actual scores. Logical inconsistency could result from respondents having consistent preferences but who are confused by the complicated tasks. Although determining the cause of logical inconsistency is not the primary objective of this study, it is worth addressing the impacts of including logically inconsistent values on actual mean scores of health states because it may have implications for the possibilities to reduce inconsistency in the future and the need to exclude some respondents when modelling health state valuations.

To explore the effects of logical inconsistency on preference scores, additional analyses of the impact of including the inconsistent responses on the observed scores are reported in Chapter 4. This chapter aims to examine how to treat the inconsistent scores before testing different model specifications because including such scores means that the scores from the respondents who may be unable to understand the tasks are included. This would "dilute" the "quality scores" given by the respondents with a better understanding of the task and as a result the estimated scores may well not represent the Thai preferences. Before a decision can be made on how to exclude the inconsistent scores, the implications of including the scores from logical inconsistent respondents on the models and the estimated scores are thoroughly explored. This is done by classifying the respondents into several groups according to the numbers of inconsistent responses identified. The exclusion of the respondents before including the scores in the model specifications suffers from two concerns; the first is that, to make the most use of the data, all actual scores should be included in the model specifications. However, it is unconvincing to include the scores from those respondents exhibiting extreme logical inconsistency which can be the results of the misunderstanding of the interview process. The second concern is, if some scores have to be excluded, how to find the appropriate number of inconsistent responses to be excluded. The reasons underlying the decision on the appropriate number used to exclude the inconsistent respondents are described in this chapter. Scores from the selected group of respondents are going to be used in the model specifications in Chapter 5.

A number of different model specifications are explored in Chapter 5 in order to find the "best" model to explain Thai preference scores using the scores from the preferred respondent subgroup from the previous chapter. This thesis does not offer a new model to explain the Thai scores, three existing models are explored: Dolan (1997), Dolan & Roberts (2002) and Shaw *et al.* (2005). Different models have their own strengths and weaknesses and different models will generate different scores. Therefore, to select a model to estimate the scores, criteria to select the "best" model are generated and, following the criteria, the "best" model is chosen. To test the performance of different models, the scores from the chosen subgroup were randomly divided into a modelling sample and a validation sample. The best model should predict scores which differ from the actual scores as little as possible. Different specifications of the model are produced using the modelling subgroup and the estimated coefficients are used to predict the preference scores in the validation sample. The predicted scores are then compared with the actual scores. The best model is chosen and the impact of the choice of respondent subgroup on the models is explored to reassure the appropriateness of the selected respondent subgroup. The Thai model is presented at the end of this chapter. The Thai scores for all 243 EQ-5D states are presented in the appendix at the end of the report.

The thesis ends with Chapter 6 where the overall contribution of this study is discussed. The study can fill the gap in economic evaluations in Thailand by providing the first set of Thai population-based preference scores for health. By thoroughly exploring the actual Thai scores, it is clear that the respondents gave a considerable number of logically inconsistent responses. One cause of the generation of inconsistent responses could be that some of the respondents may have had difficulties trying to understand the interview tasks. By closely examining the inconsistent responses, additional issues can be highlighted regarding the impact of including inconsistencies on the estimated scores. The fieldwork upon which the statistical analysis was based was successfully executed. However, there are useful lessons for future surveys of health state preferences. Close collaboration with the field coordinators is one of the key enabling factors for the identification of respondents.

Although, it is certain that the research was successful, a number of limitations emerged and should be documented to be used as a guide to reduce the same kinds of limitations in future studies. Limitations of this study include the modifications of the MVH protocol, the exclusion of the directly observed scores in the model specifications, the model analysis, the interviewer-related difficulties and the arrangements of interview site settings, illness experiences of the Thai respondents and cognitive overloads occurring in the respondents when engaged in the preference interview. The Thai model is not much different from the models estimated for other countries in that although the "best" model was selected, it still suffers from misspecification and heteroskedasticity. Additional relevant variables might be included and different functional forms could be examined and may probably improve the model performance. The new version of EQ-5D which is expected to be available in the near future may offer new potential methods to capture preferences on health of the Thai population. However, it appears likely that the Thai preference scores estimated in this study represent the preferences over health of the Thai population and are applicable to decisions over resource allocation in the health sector.

1.3 EQ-5D health states

The EQ-5D measure has been developed by a multidisciplinary group of experts since 1987 aiming to establish a generic health outcome measure which is easily self-completed (18, 19). The measure is widely used in the measure of population health status, for example, a survey in six European countries to measure population health status (20). Development history of the instrument and the preference elicitation procedures are thoroughly documented (21, 22). The measure is composed of five dimensions including: mobility; self-care; usual activity; pain or discomfort; and anxiety or depression. Each dimension has three levels of severity: no problem; some problems; and severe problems (19). The EQ-5D is to be preferred for use in the Thai study. The measure consists of only five dimensions which is in accordance with the recommendations by Froberg & Kane that less than nine health state attributes should be used to describe health (23). It is likely that by simultaneously processing only five pieces of information, respondents should encounter fewer cognitive difficulties in assigning scores to health states. Other supporting reasons are that the psychometric properties of the EQ-5D are highly acceptable and work considerably well in differentiating the respondents with or without clinical conditions. Users of this measure would benefit from this aspect when using it to measure health outcomes. From the study by Brazier *et al.*, it is easy to self-complete and can be used to discriminate health status of the patients with chronic obstruction of pulmonary disease (COPD), rheumatoid arthritis from general population (24). It is "found to be correlated moderately well with other generic and condition-specific measures". It is reported that the EQ-5D was highly acceptable to the general population (more than 95% response rate), with good reliability and good construct validity (25).

The EQ-5D is ready to implement in this study because the measure was officially translated into Thai and the translation approved by the EuroQoL group. The measure is free of charge for non-profit use. In fact, the measure has already been implemented in a number of studies in Thailand, for example, Misajon *et al.* (26) and Sakthong *et al* (27). Moreover, it is used to measure outcomes in economic evaluations worldwide. It is most frequently used in the UK, the US, Canada and the Netherlands (28). Brauer *et al.* reported that the number of studies using EQ-5D increased from 5.7% in 1997 to 11.5% in 2001(29). Rasanen *et al.* reviewed the economic evaluations published during 1966-2004 and reported that the measure was most commonly used in the QALY estimation of health outcomes (46.8% out of 81 studies) (28). This is in line with the report by Richardson & Manca who reviewed QALY measurements in randomised controlled trials (RCTs) during 1995-2002 where health in 70% of 23 papers were measured using the EQ-5D (30). The measure is most commonly used in the Industry submissions requesting listing by the Australian Pharmaceutical Benefit schemes reviewed during 2002-2004 (31) and in economic evaluation reports to NICE (32). It is recommended in the 2008 NICE methods guide that health effects are preferably measured by the EQ-5D (16). Recently, the EQ-5D has been recommended to be used in health outcome measurement in the Thai Health Technology Assessment Guideline (33).

One may question the feasibility of implementing the measure in the Thai population. Although, to the best of my knowledge, no studies have reported the psychometric properties of the EQ-5D in Thailand, studies conducted with other health outcome measures are potentially relevant. Lim *et al.* reported that the Thai-SF36 has satisfactory psychometric properties (34). Given that there is evidence that the "health concepts" embodied in the SF-36 are "applicable to Thais", the "health concepts" of the EQ-5D could be assumed to work relatively well in Thais because the health attributes encompassed by the SF-6D, are to some extent, similar to the attributes of the EQ-5D. This argument is supported by a study of quality of life dimensions relevant to Thai respondents (60). The health concepts embodied in the EQ-5D can, to some extent, be identified with some of these quality of life dimensions, which include spiritual life, family life, self, personal health, social life and work life (35).

Although the EQ-5D health state descriptions are available in Thai, there is no official report on the extent of the understanding by Thais of these descriptions and of the psychometric properties of the measure in the Thai general population. Fox-Rushby and Hunt *et al*. suggested that users of the EQ-5D should be aware of conceptual (un)equivalence or cross-cultural adaptation between the English language version and those of other languages (36, 37). Cheung and Thumboo also stated their concerns over the translation of an English health outcome measure in Asia, indicating that the quality of translation and the investigations of semantic equivalence may not be sufficient (38). However, the Thai EQ-5D is going to be used in this study although the issues of descriptions are yet to be solved. After all, the Thai EQ-5D has been successfully implemented in several studies and the issues of translation issues have not emerged. The issues of translations are concerned but they are out of the scope of this study.

In short, the EQ-5D is selected for use in the Thai study because the measure seems to be easy to comprehend, and already officially translated in Thai and available from the EuroQol Group. The measure is used worldwide both in economic evaluation and in the measurement of quality of life if the patients in some clinical conditions. A number of organizations recommend that health outcomes should be measured using EQ-5D. In other countries, the EQ-5D is widely accepted by respondents and the psychometric properties of the measure are considerably high in the construct and concurrent validities. The responsiveness of the measure is fairly high. Several dimensions of the EQ-5D are identified with the Thai Quality of life dimensions. Although the psychometric properties of the Thai EQ-5D have yet been examined in the Thai general population, it is likely that the properties are fairly high and the measure is highly acceptable by Thai population.

1.4 The MVH protocol

The Measurement and Valuation of Health (MVH) protocol which was developed by a group in the Centre for Health Economics, University of York, has been used to elicit preferences for EQ-5D health states in a number of countries. The protocol aims to elicit the "valuations that ordinary people attach to different (multi-dimensional) health states" (39). The interview protocol was reported in several papers (40-42). Countries which have estimated preference scores from their general population are presented in Table 1.1.

Table 0.1 **Countries with population-based preference scores for EQ-5D**

Country	**Elicitation methods**	**Sample size**	**No. of health states interviewed**	**Model**	**Authors**
UK	VAS and TTO	2,997	42	Random Effects	Dolan 1997 Dolan & Roberts 2002
Finland	Postal VAS	1,634	42	OLS	Murti *et al.* 1997
US	Postal VAS	1,025	42	OLS	Johnson *et al.* 1998
Slovenia	Postal VAS	370	42	OLS	Rupel&Rebolj 2000
Spain	TTO	975	42	Random Effects	Badia *et al.* 2001
Japan	TTO	621	17	Plain main effects	Tsujiya *et al.* 2002
New Zealand	Postal VAS	1,360	13	Random Effects	Devlin *et al.* 2003
Zimbabwe	TTO	3,395	72	Fixed Effects	Jelsma *et al.* 2003
US	TTO	4,048	42	Random Effects	Shaw *et al.* 2005
Germany	VAS and TTO	339	42	Not stated	Greiner *et al.* 2005
The Netherlands	VAS and TTO	309	42	Random Effects	Lamers *et al.* 2006
Latin America (Using only the scores from Spanish-speaking respondents from the US scores)	TTO	1,115	42	Random Effects	Zarate *et al* . 2008
South Korea	TTO	488	42	Random Effects	Jo *et al.* 2008

(15, 43-55)

All previous studies derived actual values for up to 42 states with the smallest number being 13 states in the New Zealand study. It could be argued that the different numbers of health states may influence the Time trade-off (TTO) scores, however, the study by Kok, Stolk and Vusschbach reported that the resulting TTO scores were unlikely to be influenced by number of health states used in the interview or to have "response spreading" (56). The authors also advocated the implementation of a flexible interview protocol in different settings in which number of health states interviewed could possibly have a significant influence on the preference elicitation using the TTO method. Both postal survey and face-to-face interview have been administered. A ranking exercise was used as a warm-up and preferences were elicited using the Visual analog scale (VAS) and TTO methods and the scores were estimated from the two methods. One of the following three statistical modelling methods were used to estimate the scores: Ordinary least square (OLS), Fixed effects and Random effects models. The MVH protocol is going to be adapted in this study because the protocol has been used in a considerable number of countries. The experience from these countries can be useful. However, it should be noted that the MVH protocol has mostly been implemented in developed countries where the general population tend to be better educated than the Thai general population. Before implementing the MVH protocol in Thailand, a pilot study should be conducted to test the feasibility of the protocol especially the cognitive burdens imposed on Thai respondents participating in the interview. Results of the pilot studies and the overall study design are reported in the next chapter.

1.5 Conclusion

The study aims to estimate preference scores for health from the Thai general population. The scores are to be used to calculate QALYs in cost-utility analysis of health technology in the Thai settings. The wide use of EQ-5D in Thailand gives the first clue leading to the selection of health outcome measure to be presented to the Thai respondents. Additional support for the use of the EQ-5D are that it has been officially translated into Thai and can be used free of charge. The measure is recommended by several international organisations to be used in health outcome measurement and a number of countries have valued the EQ-5D health states. Regarding the preference elicitation methods, because the methods have never been utilized in Thailand, the experiences from other countries are reviewed and adapted. The MVH protocol is commonly used worldwide, thus the protocol can be used as a guide to shed light on how the preference elicitation interview should be conducted in the fieldwork. Information obtained from the review in this chapter is used to plan the study design which is reported in the next chapter.

Chapter 2 Fieldwork survey

2.1 Introduction

As stated in the previous chapter, preferences for health are going to be estimated from general population. A sample size is needed to be estimated and a method used to recruit a representative sample of the Thai general population is reported in the first section of this chapter. The Thailand National Statistical Office (NSO) was invited to collaborate with respondent recruitment. The NSO and the IHPP have had collaborative projects on a number of national surveys, one of them being the Health and Welfare Survey and the Disabilities Survey (HWS) in 2007. It was the respondents to this survey who were expected to be recruited for the preference elicitation interviews. Next, the survey instruments including recording form, health cards and the TTO boards are described. The TTO boards were adapted from the props used in the original MVH protocol. After the pilot studies were conducted in the Thai PhD students in London and the researchers in the funding organisations, it was suggested that Thais may be able to give scores for eleven health states within one to one and a half hour of interview duration. This section reports the selection method of health states to be used in the interview. The third section describes identification of respondents recruited from the HWS survey and preparation of interview sites. To cope with the number of respondents, a group of interviewers were employed and trained extensively on the interview protocol. To realise the cognitive burden which would have occurred to the respondents taking part in the interview, an exploratory qualitative study was also conducted. The conclusion section is presented at the end of this chapter to summarise the fieldwork survey before proceeding to the results of the interview and data analysis in the next chapter.

2.2 Sample size and the sampling method

As stated in the previous sections, the aim was to conduct preference interviews in a nationally representative sample. To calculate a sample size, following the suggestions of O'Brien and Drummond (41), the minimal meaningful difference between health state values was determined to be 0.1, the sample size determined to detect a difference between the means of two health state values was calculated using the following formula.

$$N = \frac{2\sigma^2(\varepsilon\,(U+V))^2}{(\mu-\mu_0)^2}$$

Where:

N = sample size

U = a desired power of the test

V = a desired significance level

σ = standard deviation

ϵ = a function of the desired power and significance level

$\mu - \mu_0$ = difference between two means

(41).

Table 2.1 shows the determination of sample size for given significance levels and the differences between the means of two health state values to be detected at 80% power.

Table 2.1 Sample size determination

Difference to be detected between two means (power = 80%)	Significance level	
	0.01	0.05
0.025	4,827	3,235
0.05	1,207	809
0.10	302	200

Applied from Gudex *et al.* (41).

As described in the sample size calculation for the UK MVH study, the difference between two means was expected to be 0.025 at 80% power and the significance level was at 0.05, therefore, the size of a sample needed was 3,235. This implies that 3,235 observations are needed for each health state and every respondent is given the same health states. This size of observation was unlikely to be manageable in the Thai study, given the limited time and budgets. Number of observations was then changed to the lowest number, so the difference to be detected between two means and the significance level were changed to be 0.10 and 0.05, respectively. As a result, at least 200 observations per health state were obtained in this study and the respondent is expected to be given the same sets of health states to be valued. Alternatively, the suggestion by Williams as suggested in the pilot study of the MVH protocol application in the population survey could be used as a guideline (39). Williams recommended that to avoid respondent cognitive overloads, each state could be valued by at least 35 people. This latter recommendation could be applied if the first recommendation is not feasible in the fieldwork survey in Thailand.

A stratified four-stage sampling method was implemented. In the HWS 2007, the principal frame was taken from the 2000 Population and Housing Consensus (57). The principal frame was divided into 2 residential areas: Municipal (urban) and Non-municipal (rural). To cover the geographical distributions of the respondents, all 76 provinces in Thailand were divided into 5 regions: North, Northeast, Central, South and Bangkok. The primary sampling units were provinces, the secondary sampling units were blocks in municipal areas and villages in non-municipal areas. The third stage sampling units were private households and the final stage sampling units were persons aged 20 years and over from the randomly selected households. This group of respondents was recruited because at this age, the respondents were assumed to be mature and possibly capable of expressing their own preferences towards health states. Random selection was used in every sampling unit and the number of selected households was based on the probability proportional to size. Ten households per block or village were randomly selected by the NSO then the lists of households and persons, with the information including age and gender, were given to the researcher and the household members were randomly recruited to be interviewed.

Seventeen provinces were randomly selected from five regions. The three provinces in the South region were Chumporn, Nakorn Srithammarat and Trang. The six provinces from the Northeast region were Khon Kaen, Kalasin, Mahasarakham, Chaiyaphum Buriram and Roiet. The three provinces from the North region were Phitsanulok, Lampang and Payao. The four provinces from the Central region were Supanburi, Chainat, Prachuab-Kirikhan and Chanthaburi. The capital city: Bangkok was treated as a separate region. The sample recruited for this study is presented in Table 2.2 according to regions, provinces, block/villages and number respondents. Lists of blocks/villages according to the provinces are presented in Appendix 1.

Table 2.2 Numbers of respondents selected from the chosen provinces according to residential areas (urban/rural)

Regions and provinces	Total		Urban area		Rural area	
	block/village	number of respondents	block	number of respondents	village	number of respondents
Bangkok	**16**	**160**				
Central region	**35**	**350**	**12**	**120**	**23**	**230**
Supanburi	12	120	3	30	9	90
Chainat	6	60	1	10	5	50
Chantaburi	10	100	5	50	5	50
Prachuab-Kirikhan	7	70	3	30	4	40
North	**25**	**250**	**5**	**50**	**20**	**200**
Lampang	10	100	3	30	7	70
Payao	6	60	1	10	5	50
Phitsanulok	9	90	1	10	8	80
Northesast	**44**	**440**	**9**	**90**	**35**	**350**
Kalasin	7	70	2	20	5	50
Khonkaen	9	90	3	30	6	60
Roiet	7	70	1	10	6	60
Mahasarakham	6	60	1	10	5	50
Buriram	7	70	1	10	6	60
Chaiyaphum	8	80	1	10	7	70
South	**17**	**170**	**4**	**40**	**13**	**130**
Nakorn-Srithammarat	8	80	2	20	6	60
Trang	5	50	1	10	4	40
Chumporn	4	40	1	10	3	30
Total		**1370**				

2.3 Survey instruments

The survey instruments, including the recording forms, were redesigned according to the recommendations from the pre-test studies. The information sheet and consent form were initially prepared in London and some changes were made after the pre-tests. All survey instruments included:

- Recording form with the EQ-5D questionnaire
- Health cards
- TTO boards

Information sheet and consent form

Information about the study was presented in the information sheet. Respondents were asked to give their consent in the consent form.

Recording form with the EQ-5D questionnaire

The original Thai EQ-5D questionnaire was adapted and integrated in the recording form by the researcher. In the final version, respondent background information: name; address; age; marital status; and number of children were located in the upper half of the first page. The lower-half included date of the interview, the start and finish time, the health set used in the interview and the interviewer name. Fuller respondent information was not requested in the interview because, as stated before, the respondents were part of the HWS 2007 survey. It was planned to link the database of respondents in this study to the NSO database when the survey was completed. To be able to link the respondent database with the NSO, the bridging codes included the codes for respondent's address were added to the front page of the recording form.

The second page was the self-completed EQ-5D of own health in the past 24 hours, followed by the "thermometer scale" for the respondent's own health on the third page. Results from the Ranking and VAS methods were presented in pages 4-5 and those for the TTO method in pages 6-9. The final sections were the self-completed questionnaire asking the respondent to comment on any difficulties encountered during the interview. The self-completed questionnaire asking for the interviewer's comments on respondent performance is included in Appendix 2.

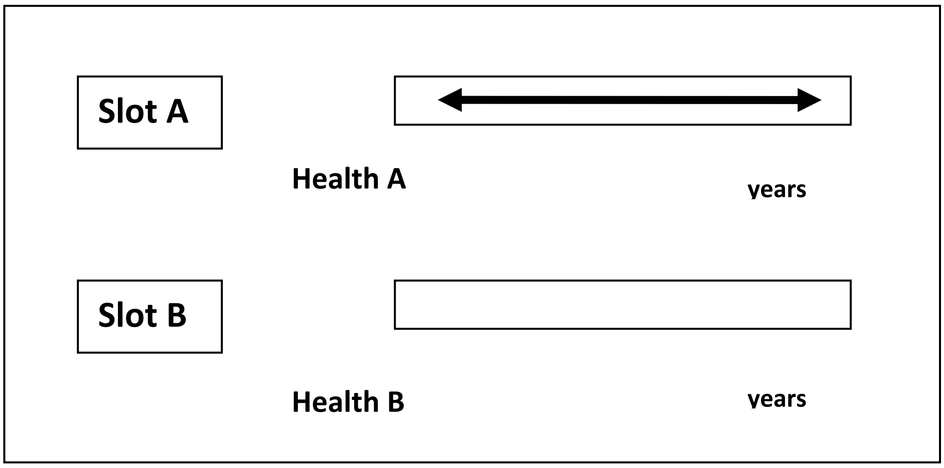
Health cards

Descriptions of health states in Thai were printed on white paper sized 12 x 18 centimetres, one line representing each dimension ranging from mobility, self-care, usual activities, pain/discomfort and anxiety/depression respectively. All health cards were laminated to make them durable for any possible damages, for example, water or scratching. Eleven cards for each health set were put in an envelope with a distinctive label indicating the number of health set. Examples of health cards are presented in Appendix 3.

TTO boards

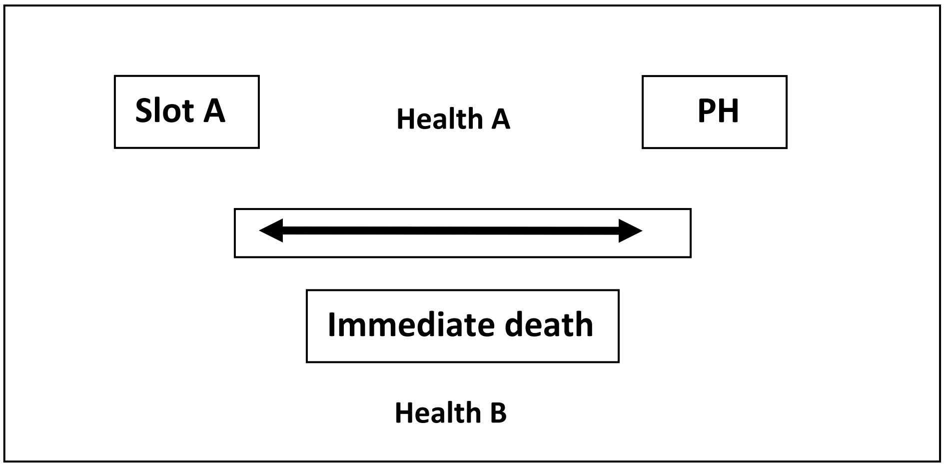
There were two TTO boards, one was used in the interview with health states respondents regarded as better than death and the other side was with health states viewed as worse than death as shown in Figures 2.1 and 2.2, respectively. The boards were made of blue rectangular cardboard sized approximately 70 x 50 cm. The boards were intended to be larger than the original TTO boards to assist with the visualisation by the respondents.

Figure 2.1 TTO board for state better than death (TTO board 1)



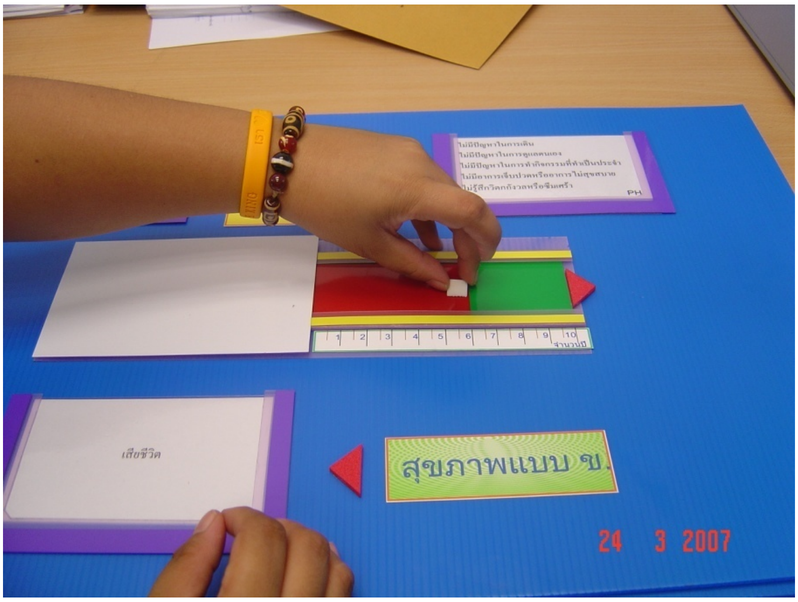
Slot A and B were the transparent envelopes prepared for the insertion of health states used in the TTO interview. Next to Slot A, a moving indicator was attached to identify years of life that the respondent would like to sacrifice. Note that the duration between 9 years and a half and 10 years was divided into single months to allow respondents to choose the duration of 9 years 7 months, 9 years and 8 months until 9 years and 11 months before choosing 10 years. These slots were used to allow respondents to sacrifice a very short duration of living in a very mild state. The indicator beside Slot B was fixed at 10 years.

Figure 2.2 TTO board for state worse than death (TTO board 2)



In this board, Slot "PH" was a permanently fixed card description of state 11111 and "Immediate Death" was a permanently fixed card described as immediate death. In the centre of the board, a sliding indicator was used to indicate the number of years the respondent would like to sacrifice. A picture of the actual TTO board for state worse than death is presented in Picture 2.1.

Picture 2.1TTO board for state worse than death



2.3.1 The selection of health states to be used in the interview

The pre-test studies suggested that Thai respondents might be able to assign scores to eleven health states in a single interview and that the average interview duration ranged from one hour to one hour and a half. Around the time that the health states were to be selected for the fieldwork interview, Prof. Paul Kind was invited to be a consultant of the preference elicitation survey by the IHPP. Prof. Kind suggested the method of health state selection be as follows. The states selected for the interview should cover the full range of severity. All 243 health states were divided into 3 groups: mild; moderate; and severe. Mild states were those without level 3 in any dimension and with level 2 for up to 3 dimensions. Severe states were those without level 1 in any dimension and at least with two of level 3. The states fulfilling neither of these criteria were regarded as moderate states. Distances of the states from state 11111 were calculated. Results of the differences were then summed up and used as a guide to categorise the states into Distance Group ranging from 1 to 9. Mild states were then those in Group 1-3 (five states in Group1, ten states for Group2 and ten states for Group 3), Group 4-6 were moderate and Group 7-9 were severe states (ten states for Group 7, ten for Group 8 and five for Group 9). For example, the sum of the difference between state 11212 (mild state) and state 11111 was 2 (0+0+1+0+1). This state was therefore, classified in Distance Group 2. One state from each group was randomly selected, without replacement, to form the health sets.

After assigning twenty-five states into the mild group and twenty-five states into the severe group, thirty states from Distance group 2 and 3 and twenty-five states from Distance group 7 and 8 were removed from moderate group, leaving 136 states in moderate group. Forty states (twelve from Distance Group 4, sixteen from Group 5 and twelve from Group 6) were randomly selected from this moderate group. The health states categorised into mild, moderate and severe groups are presented in Table 2.3.

Eleven health states were used in each interview. These eleven states included two anchor states (11111 and 33333), three mild states, three moderate states and three severe states. Three states (one from the distance group 4, one from group 5 and one from group 6) were selected from the moderate group without repetition, twelve sets were formed with all states from the distance group 4, two states were left out from distance group 5 and one state was left out from distance group 6. Three of health states in the mild group were randomly chosen to combine with the twelve sets previously prepared. One health state from the distance group 1, one from the distance group 2 and one from the distance group 3 were randomly selected. However, since there are only twenty-five states in the mild group and only five states from the distance group 1, therefore, after the states in the distance group 1 were used up (from set 1 to set 5), the states in this group were repeatedly allocated to set 6 to set 10. Thus the three mild states in set 11 and set 12 were similar to those in set 1 and set 2. The selection of severe health states was similar to the selection of mild states. One state was randomly chosen from the distance group 7, one from the distance group 8 and one from the distance group 9. In total, eighty-six health states (including state 33333) were used in the Thai study. Details of health sets in all twelve sets are presented in Table 2.4.

Table 2.3 EQ-5D states in the mild, moderate and severe groups

Mild group					
EQ-5D states	**Distance group**	**EQ-5D states**	**Distance group**	**EQ-5D states**	**Distance group**
11112	1	11122	2	12122	3
11121	1	11221	2	12212	3
11211	1	11212	2	12221	3
12111	1	12112	2	11222	3
21111	1	12121	2	21122	3
		12211	2	21212	3
		21112	2	21221	3
		21121	2	22112	3
		21211	2	22121	3
		22111	2	22211	3
Moderate group					
EQ-5D states	**Distance group**	**EQ-5D states**	**Distance group**	**EQ-5D states**	**Distance group**
3 1 3 1 1	4	23311	5	33122	6
11223	4	11332	5	32123	6
31131	4	13123	5	21332	6
21312	4	31213	5	13232	6
21231	4	23131	5	31313	6
11313	4	21313	5	22232	6
11232	4	12331	5	23222	6
22113	4	33211	5	22313	6
12123	4	13222	5	33221	6
12312	4	21133	5	23132	6
21123	4	12313	5	23321	6
22221	4	31222	5	23231	6
		33121	5		
		11323	5		
		21331	5		
		23113	5		
Severe group					
EQ-5D states	**Distance group**	**EQ-5D states**	**Distance group**	**EQ-5D states**	**Distance group**
22233	7	23323	8	23333	9
22323	7	23332	8	32333	9
22332	7	22333	8	33233	9
23223	7	23233	8	33323	9
23232	7	32233	8	33332	9
23322	7	32323	8		
32223	7	32332	8		
32232	7	33223	8		
32322	7	33232	8		
33222	7	33322	8		

Table 2.4 Twelve sets of health states used in the study

Set 1	Set 2	Set 3	Set 4
11211	11112	21111	11121
12112	21121	21211	22111
22112	12221	11222	12212
31131	11313	12123	22221
13123	21313	13222	31222
21332	22232	33221	23231
23223	22323	32223	32232
23323	32323	33232	23233
33233	23333	33332	32333
11111	11111	11111	11111
33333	33333	33333	33333

Set 5	Set 6	Set 7	Set 8
12111	11211	11112	21111
11212	21112	11122	12211
12122	22121	21122	22211
21312	22113	31311	11232
21331	11332	12331	12313
13232	22313	33122	23222
32322	23232	33222	22233
32332	22333	33223	32233
33323	33233	23333	33332
11111	11111	11111	11111
33333	33333	33333	33333

Set 9	Set 10	Set 11	Set 12
11121	12111	11211	11112
11221	12121	12112	21121
21212	21221	22112	12221
11223	21123	21231	12312
23113	23131	33121	31213
32123	23132	23321	21332
23322	22332	23223	22323
33322	23332	23323	32323
32333	33323	33233	23333
11111	11111	11111	11111
33333	33333	33333	33333

English letters were used as a code for each health state and located at the lower right corner of the cards. The objectives of coding were to minimize the interviewers' workloads and to reduce the possibilities of incorrectly recording the scores given to the states. The coding also facilitated data entry. The first letters were A, B, C,…, L for health set 1,2,3,…12 respectively. The second letters were randomly assigned. As a result, E, J and O were assigned for the states selected from the mild group; T, Y and D for the states from the moderate group and K, P and V for the states from the severe group. The codes: XT and PH were assigned for state 33333 and 11111 respectively.

2.3.2 Preparation of the sample and the interview sites

The field coordinators were recruited to help with the identification of and arranging access to the respondents. To recruit the field coordinators, official letters were sent from the IHPP, to invite the provincial health offices to collaborate with the research project. Enclosed with the letter were the research proposal, the list of respondents' names and addresses, the interview schedule (date and time-slot) specified for the targeted areas in the provinces, an outline of the interview process and arrangements for the interview site (number of tables and chairs and the lay-out of the interview sites to ensure the peaceful environments) and props. Refreshment for the respondents and interviewers, and meals for interviewers were requested.

The field coordinators were appointed by the provincial health offices. They were nurses who worked in the village health offices (if the respondents were in rural areas) or nurses in the provincial hospitals (if the respondents were in urban areas except Bangkok), health volunteers, teachers and community leaders, i.e. heads of the villages. The field coordinators were assigned to locate, contact and make an appointment with the respondents to participate in the interview according to date and time indicated from the research team. The appointments were made via postal mails, telephone and oral communications. The costs of locating the respondents were covered by the project.

The interview sites were arranged in a variety of places. As seen from the pilot studies, good interview sites should be peaceful and, if possible, free from distractions. However, it depended on the feasibility of the areas and whether the field coordinators could find appropriate venues. The interview sites ranged from, meeting rooms in the hospitals or the provincial health offices, community centres, community leaders households and the respondents' households (if the respondents could not come to the provided sites).

2.4 Recruitment and training of interviewers

There were 48 interviewers participating in the survey. The interviewers were recruited from:

- Master degree students from the Faculty of Pharmacy, Mahidol University.
- Bachelor degree students from the Faculty of Pharmacy, Khon Kaen University.
- Master degree student from the Institute of Population and Social Research, Mahidol University
- Physiotherapists and occupational therapists from the Sirindhorn National Medical Rehabilitation Centre, Ministry of Public Health.
- Staffs from the BOD, HITAP and IHPP offices
- Recently graduated bachelor degree students from Lampang Rajabhat University.

A three-day interviewer training workshop was arranged during 23-25 March 2007 to inform the interviewers regarding the research objectives and what was expected from the interview. The overall interview processes were demonstrated and the interviewers had opportunities to practice the interview by interviewing their colleagues. To familiarise the interviewers with the survey environment and the respondents with whom they were going to interview, mock interviews with a group of respondents were also arranged in a public school in Bangkok. The school was selected because of the location and the availability of space including the numbers of tables and chairs which could accommodate the large number of interviewers and respondents. A convenience sample selected from the urban area of Nonthaburi province was invited to take part in the interview by the interviewers. This group of respondents was selected because of the proximity of their households to the school. Although the sample was chosen from the urban area, the characteristics of this group of respondents were closely matched with the respondents to be interviewed in the data collection phase. This gave the interviewers the sense of what the "real" interview would be like. Regarding quality control of the interview, the performance of the interviewers was observed by the researchers and the assistants. Recording forms were checked after the end of the interview on the first day and mistakes were addressed and the interviewers were informed. The researcher had a good opportunity using this workshop to practice the preparation of interview sites and the management of the problems that could arise in the fieldwork survey. The most common pitfalls were analysed and discussed with the interviewers. A final version of the interview manual was developed after the interviewer training workshop. The interview manual was distributed to all interviewers. After the training workshop was completed, the fieldwork survey schedule was planned.

During the fieldwork, five interviewers including the researcher and the research assistants formed the research team. The respondents in the Northeast region were interviewed by the interviewers from Khon Kaen University. The respondents in other regions were interviewed by the rest of the interviewer team. The researcher accompanied the interviewer teams on all fieldwork trips.

2.5 The interview process

The interview process including the respondent screening procedure, the three preference elicitation interview methods and the criteria used to terminate the interview, are described in this section.

2.5.1 Respondent screening

Inclusion criteria for the respondents were that they were: [1] included in the list given by the NSO; [2] literacy and no extreme hearing problem; and [3] able to communicate with the interviewers without assistance from their family members (to avoid any influence on the respondent's answers). Respondents were first, screened by the research assistant to check whether they were included in the list of potential respondents. Second, they were asked to verify whether their full names were correct and to sign if their personal details were correct. The name verification was used as a means to screen the literacy ability as well as the visual and hearing abilities of the respondents. The respondents were then asked to read an example of health state cards. If the respondents needed reading glasses, they were asked to go back to their households and bring their reading glasses. If the hearing ability was poor (at the communication level), the researcher decided to exclude the respondents at this point because they would have caused communication problems with the interviewers.

2.5.2 The overall interview process

To ensure that each health set was used and received relatively equal numbers of observations, the sets to be used each interview day were pre-defined by the researcher. The overall interview process was as follows.

Introduction of the research project

The interviewers introduced general information about the project according to the information sheet and described the interview procedure to the respondents. If they agreed to participate in the interview, they were asked to indicate their consent on the consent form.

Background information

The respondent was asked to fill in their name, address, age, marital status and number of children. The respondents' addresses were then used to merge with the database of NSO in which other personal information such as educational level and incomes, were already available. This would help minimise the overall interview duration. Then the respondent filled in the Thai EQ-5D questionnaire and rated her own health status in the past 24 hours, using a VAS thermometer scale

Ranking exercise

The health set was given to the interviewer by the researcher. All eleven health states in the set were presented to the respondent. The interviewer asked the respondent to rank all eleven health states according to her preference of being in each health state for 10 years followed by death. The state at the top of the rank was the best state and at the bottom of the rank was the worst state.

Visual Analog Scale (VAS) exercise

Next, a 20-cm scale with the lowest score of 0 identified as "the worst health state imaginable" and the highest score of 100 as "the best health state imaginable" was introduced. The respondent was requested to place all 11 health states on the scores according to her preferences over the states. Before moving to the next process, the respondent was allowed to re-visit the rank and the scores given to the states, she was allowed to rearrange the rank or change the scores if she wished to do so.

Time trade-off (TTO) exercise

The interviewer randomly selected a health state from all ten states from a given health set (state 11111 was used as a reference state) and asked the respondent to imagine herself being in this state for ten years. The respondent was asked whether she regarded the given health state as a better than death or worse than death. This first question was used because the interview dialogue, as well as the TTO board, for a state regarded as better than death differed from those for a state regarded as worse than death. The time sacrificed in order to live in a better state was gradually changed until the respondents were indifferent between the two states.

After the respondent completed all steps in the interview, she was given a remuneration of 200 baht (£4). If the respondent could not complete the interview, the remuneration was reduced to 100 baht (£2). The interviewer is given 100 baht (£2) per interview. The respondent was requested to sign a receipt after receiving the money. The interviewer was required to check the completeness of the recording form before returning the forms to the researcher. At the end of all interview days, the researcher re-examined the completeness of all recording forms to check and identify any mistakes occurring in the interview. In the following interview day, the mistakes were demonstrated without blaming the interviewers, to remind the interviewers before starting the interview. The mistakes were mostly found in the direction of moving the time indicators in the TTO boards.

Criteria to terminate the interview

There are a number of factors that contribute to an unsuccessful interview. The respondents may be unable to understand the task after the interviewers explained the overall procedure. Alternatively, the respondents may understand the tasks but cannot differentiate between health states or imagine themselves in the hypothetical states described in the cards. As a result, they may have difficulties expressing their preferences over health states. It is possible for respondents to become stressed in response to the interview questions, especially when they successfully assigned scores using the Ranking and VAS methods, but not with the TTO method. If this was the case, the interview was considered to be terminated if the respondents were willing to terminate the interview and if the interviewers were uncertain about the respondent's level of stress. The researcher was notified and the decision on terminating the interview was dependent on the researcher's decision.

2.6 A qualitative study

Although the investigation of the cognitive burden on elderly respondents was not one of the main objectives of the study, it could be seen through the data collection process that more of the elderly respondents from the rural areas tended to have more difficulties with the elicitation interview. The opportunity of conducting an exploratory qualitative interview emerged when the preliminary results were reported to the funders in Thailand in July 2008. The researcher took this opportunity to conduct some exploratory qualitative interviews with elderly respondents. The main objective of this qualitative interview was to understand the coping mechanisms employed by the elderly respondents when taking part in the interview, and to explore the cognitive burden of the interview on the respondents. Note that the qualitative interview in this study is intended to be an "exploratory" exercise. The results of this interview could be used to generate the preliminary assumptions to explain the cognitive burdens of the respondents. A future "proper" qualitative interview can be conducted building on the findings generated from these exploratory interviews. The qualitative interviews were conducted in Khon Kaen province in August 2008. A convenience sample of ten respondents was recruited by the health volunteer working in a village where the researcher lives. The respondents included were those aged 60 years and older with primary level education. The interviews were conducted in the health volunteer's household which was in the respondents' neighbourhood. The respondents were screened, by the researcher, for literacy ability and the ability to participate in the interview. The preference interview method used in the qualitative study was similar to the interview in the data collection phase, except the number of health states used in the qualitative study was reduced to six states. The scores were not going to be used in the preference scores estimation. To allow the researcher to explore further on the coping mechanisms employed by the respondents, they were encouraged to "think aloud" while engaging in the elicitation tasks. A remuneration of 100 baht (£2) was given to each respondent at the end of the interview. The expenses incurred to the health volunteer to make appointments with the respondents were covered by the researcher.

2.7 Conclusion

This chapter reports the preparations for the fieldwork survey including the sample size calculation and sampling method, the recruitment of respondents, the health state selection and the preference interview procedure. A representative sample was part of the HWS 2007 recruited by the NSO. A stratified four-stage sample method was implemented in this study, and potential respondents were randomly selected from seventeen provinces to ensure that the respondents were recruited to cover all geographical areas of the country. The MVH protocol was redesigned to minimise the cognitive workload that would be incurred by the Thai respondents. A total of eighty-six health states organised in twelve sets of eleven states were used in the interviews. The interview included Ranking, VAS and TTO methods. Details of the interview props and the interview process were described in this chapter. Forty-eight interviewers were recruited and extensively trained to interview a representative sample of 1,370 respondents, aged 20 years and older from both urban and rural areas. Field coordinators were assigned to locate the respondents and arrange the interview sites in the respondents' neighbourhoods. An exploratory qualitative study was conducted in a convenience sample to explore the coping mechanisms, and evaluate cognitive workload, in respondents who were likely to have difficulties participating in the interview.

Chapter 3 Results of the interview and data analysis

This chapter reports the results of the fieldwork survey which was conducted during May-August 2007. As described in the previous chapter, to minimise the complexities of the interview tasks and to maintain the level of concentration of the respondents on the tasks, the MVH protocol was redesigned and appropriate interview environments were provided. The results in this chapter shed light on whether the interview tasks used in the survey are still to some extent problematic for the respondents, given the adaptations of the original MVH protocol. Cognitive workloads are monitored relatively using interview duration, comments from the respondents and from the interviewers, and the qualitative study. The natures of the scores elicited from the TTO interview are explored in this chapter before being used in model specifications in the next chapter. To examine the influence of respondent characteristics on the scores, those derived from different groups of respondents according to age-group and gender are compared. There are two parts to this chapter: the results of the fieldwork survey and the qualitative study; and the analysis of the TTO scores. In the first part, a brief overview of the fieldwork management and the numbers interviewed in each province are reported. The number of respondents, their demographic characteristics, and the self-reported EQ-5D as well as the VAS scores of the respondents' own-health are included in this section. Then, the actual scores given to the health states, including the mean and standard deviation (SD), are reported in the second part. The chapter ends with a comparison of the scores across different groups of respondents and the chapter conclusion.

3.1 Fieldwork managements

Two parts of the fieldwork management are explained in this section. The access to the respondents by the field coordinators and the problems arising in the process are reported. Next, the arrangements for interview sites are explained.

3.1.1 Locating the respondents

There were two stages to the selection of respondents. Ten households per block (urban areas) or village (rural areas) were randomly selected by the National Statistical Office (NSO) from the sample of the Health and Welfare 2007 survey before being sent to the researcher. The lists included names, ages and gender of all household members.

Because the numbers of the respondents given by the NSO exceeded the numbers expected to be interviewed, the respondents in each block were randomly chosen again by the researcher. Gender and age proportions of the respondents in the second selection were considered to be similar to those of the general population. The final lists of respondents were then sent to the field coordinators. Reminders were sent to the coordinators if there was no reply a couple of weeks after the lists were sent.

It was often the case that some respondents were unable to be located by the coordinators. They were those who worked or studied in other cities or had moved out from the household without reporting to the registration offices. Some addresses were invalid or the houses were demolished. In these cases, the researcher was notified by the coordinators and new lists of the respondents from the same blocks were sent to the coordinators until the target number for that block was met. Some respondents could not be located by the coordinators but were known to other respondents. They were invited to the interview by the respondents who recently completed the interview.

3.1.2 Interview sites arrangements

There were two types of interview site arranged for the survey. The respondents in the Bangkok region were interviewed in their households because it was difficult to arrange the interviews in one place and invite the respondents to travel to the arranged site given their tight schedule and heavy traffic in Bangkok. All of the interviews in Bangkok region were scheduled on weekends or on holidays. The NSO staffs were contacted and assigned to help locating the expected households because they have interviewed the same respondents in a number of NSO surveys. This was also an advantage because the respondents who were previously interviewed were well acquainted with the NSO staffs and allowed the interview team to conduct the interviews in their households.

For the respondents in other regions, the interview sites were arranged by the field coordinators and varied ranging from hospital meeting rooms, out-patient departments (OPD) of village health offices, schools, temples, community centres, shops and the households of the respondents, the field coordinators and the community leaders. Examples of the interview sites and the interviews with the respondents are presented in Picture 3.1-3.3.

Picture 3.1The respondent was given the instructions of the ranking interview

Picture 3.2 Interview at a small canteen in the respondent neighbourhood in Bangkok

Picture 3.3 Interview in a temple

Many interviews took place in unplanned interview sites. Some respondents were engaged with their day jobs, for example, those who were teachers or shop keepers, and were not available to travel to the interview sites. The interviewers were then sent to interview them at their workplaces. It was not uncommon that the respondents had to take a break from the interview to see their customers before coming back to the interview. There were also cases where miscommunications between the field coordinators and the respondents arose, for example, some respondents misunderstood that they were invited to an annual health check-up, rather than the interview, so they decided not to travel to the interview sites. In this case, the interviewers were also sent to interview them at their households. To ensure the security of the interviewers if they were dispatched to interview away from the site, at least, two interviewers were sent to the respondents' households or work places.

3.2 The Thai respondents

This section reports the numbers of respondents interviewed in the survey, demographic characteristics, numbers of interviews per interviewer and the numbers of respondents per health set. Next the respondents' own EQ-5D health states and VAS scores, overall interview duration and time for each interview method, are reported. Comments of the respondents and the interviewers regarding the interview procedures are also included at the end of this section.

3.2.1 Numbers and demographic characteristics of the respondents

A total of 1,409 respondents were interviewed. The numbers of the respondents interviewed compared with the target numbers are shown in Table 3.1. Respondent demographic characteristics; mean age; gender; and numbers according to residential areas are also presented in the Table.

The average age of the respondents was 44.6 (SD 12.5) years old. The highest number of respondents interviewed is from the Northeast region. Number of respondents interviewed was slightly higher than the target numbers except for the South region where one hundred and seventy respondents in the South region were expected to be interviewed and only 161 respondents were actually interviewed. Mean ages of the respondents from each region were lower than 50 years except in Chumporn province. The proportions of the female respondents were higher than those of the male respondents except in Chaiyapoom province. Note that only the respondents in rural areas were interviewed in Buriram province. To see the national representativeness of the sample, demographic characteristics of the sample and the Thai general population are compared in Table 3.2.

Table 3.1 The target and interviewed numbers of the respondents and their demographic characteristics

Province	Target no. of resp.	Accomplished no. of resp.	age (years) mean (SD)	gender		residential area	
				male	female	urban	rural
Bangkok	**160**	166	42.8 (13.6)	70	96	100	0
North	**250**	259					
Lampang	100	102	45.7(12)	41	61	30	72
Payao	**60**	67	43.4 (11.3)	30	37	12	55
Phitsanulok	90	90	45.3(13.2)	39	51	10	80
Northeast	**440**	442					
Kalasin	70	82	44.2(12.1)	39	43	29	53
Khonkaen	90	78	45(12.8)	35	43	17	61
Roi-Et	70	57	48.1 (14.6)	28	29	10	47
Mahasarakhan	60	70	46.3 (11.8)	33	37	12	58
Buriram	70	62	45 (10.5)	29	33	-	62
Chaiyapoom	80	92	45.2 (11.5)	47	45	26	66
Central	**350**	382					
Supanburi	120	139	42.1(12.6)	67	72	41	98
Chainat	60	60	42.9(10.8)	27	33	8	52
Chanthaburi	100	111	44.8(13.8)	54	57	56	55
Prachuab-Kirikhan	70	72	43.4(12.7)	35	37	32	40
South	**170**	161					
Nakorn-Srithammarat	80	72	47.4(13.3)	30	42	16	56
Trang	50	49	43.5(13.3)	21	28	10	39
Chumporn	40	40	50.8(13.6)	19	21	13	27
Total	**1,370**	**1,409**	**44.6(12.7)**	**644**	**765**	**422**	**921**

resp. = respondent

Gender, age, education level and residential areas are compared in Table 3.2. Note that the primary education level indicates that the respondents attended formal schooling for up to 6 years, secondary level for between 6 and 12 years, and university level for more than 12 years. Compared with the Thai general population, the mean age of the respondents in the sample was higher than the general population and the proportions of female respondents, adult age-group and those living in urban areas were greater. The proportion of elderly respondents and of respondents with secondary and university education in the sample were lower than those of the general population.

Table 3.2 Demographic characteristics of the sample compared with those of the Thai general population

Respondents characteristics		Thai general population** (x 1,000,000)	%	The samples no.	%
Number		62.80	100	1,324	100
Gender					
	Male	31.01	49.30	553	45.40
	Female	31.82	50.67	665	54.60
	proportion		1.00		0.83
Mean age(yrs.)	(SD)	32.8		44.6	(SD 12.7)
Age-group					
	Adult (20-59)	37.30	85.00	1,162	87.76
	Elderly (60+)	6.60	15.00	162	12.24
	proportion		5.67		7.17
Education*					
	Primary	20.48	58.00	841	63.52
	Secondary	9.78	27.80	264	19.94
	University	5.01	14.2	151	11.4
Residential	area				
	Urban	19.60	30.70	454	34.29
	Rural	45.40	69.30	870	65.71
	proportion		0.4		0.52

* Education data of some respondents were missing
** Source: The Key Statistics 2007, National Statistical Office, Bangkok, Thailand

Interview experiences of the interviewers may influence, to some extent, the overall interview duration. To examine the workload of the interviewers, the numbers of interviews per interviewer are provided in Table 3.3. Interviewer no. 12 had the greatest number of interviews. Interviewer no.20 and 21 had the least number of interviews. One interviewer was dismissed because she was not competent at the interview and repeatedly making mistakes in following the interview manual especially in the TTO method. The reason of the highly differences in the numbers of interview was the short of interview number. The problem arose when all the interviewers were unavailable because the interviewers who had full-time jobs (as physiotherapists, and occupational therapists) in the rehabilitation centre were unable to leave to participate in the fieldwork because they were unable to take any more leave from work. Given that most of the interviewers were master degree students, they sometimes were unavailable to participate in the interviews if the interviews were scheduled in their term-times. The problem was solved by recruiting new interviewers from the university located near the provinces where the interview took place. As a result, the recently recruited interviewers were likely to have lower numbers of interview compared to the current ones. The researcher and the research assistants also conducted some of the interviews if the interviewers were fully engaged because some of the respondents had to leave the interview sites soon.

Table 3.3 Numbers of respondents interviewed per one interviewer

Interviewer no.	no.of resp. interviewed	Percent	Interviewer no.	no.of resp. interviewed	Percent
1	66	4.68	26	21	1.49
2	32	2.27	27	28	1.99
3	34	2.41	28	12	0.85
4	33	2.34	29	25	1.77
5	58	4.12	30	22	1.56
6	35	2.48	31	25	1.77
7	75	5.32	32	24	1.7
8	59	4.19	33	18	1.28
9	33	2.34	34	31	2.2
10	71	5.04	35	11	0.78
11	28	1.99	36	23	1.63
12	77	5.46	37	24	1.7
13	21	1.49	38	19	1.35
14	57	4.05	39	23	1.63
15	36	2.56	40	21	1.49
16	32	2.27	41	26	1.85
17	71	5.04	42	22	1.56
18	50	3.55	43	11	0.78
19	24	1.7	44	16	1.14
20	2	0.14	45	13	0.92
21	2	0.14	46	9	0.64
22	17	1.21	47	5	0.35
23	19	1.35	48	3	0.21
24	25	1.77			
25	20	1.42			
			Total	1,409	100

3.2.2 Number of respondents per health set

The number of respondents for each set ranged from 101 to 129. The greatest proportion (9.16%) of interview was conducted using Health Set 2 whereas the smallest proportion using Health Set 10-12. Table 3.4 shows numbers of the respondents according to health sets.

One reason for the unequal numbers of respondents in the sets is because of the poor management of distributing the health sets to the interviewers in the fieldwork survey. The health sets were planned in each interview day starting from set 1 to set 12 respectively according to the number of respondents planned to be interviewed on that day, for example, set 1 to set 12 were planned to be used in the interview with twelve respondents. If the number of respondents interviewed was lower than twelve, for example, only nine respondents were interviewed, the health sets used on that day were set 1 to set 9. On the following day, the health sets used for that day were started from set 1, rather than from Set 10 which would have been used on the previous day. This practice was used in the first half of the fieldwork survey (the South and North regions). It was not until the number of respondents per health set that used previously in the survey was re-checked before the researcher realized about this problem. After that, the health set distribution plan was changed in order to enable the number of respondents for the health sets to be equal, especially those of the last 3 health sets.

Table 3.4 Number of respondents according to health sets

Health Set	No.of respondents	Percent
1	126	8.94
2	129	9.16
3	122	8.66
4	125	8.87
5	125	8.87
6	115	8.16
7	117	8.30
8	120	8.52
9	115	8.16
10	105	7.45
11	109	7.74
12	101	7.17
Total	1,409	100

3.2.3 Self EQ-5D health states

Before assigning scores to health states, the respondents were asked to rate their own health in the last 24 hours using the EQ-5D health states. Out of the total of 1,409 respondents, only 320 (22.71%) rated their own health as full health (11111). Almost 32% of the respondents asserted that they had some problems in pain/discomfort and anxiety/depression. Eight percent of the respondents or 113 respondents assigned level 3 to at least one dimension, two of them assigned level 3 to four out of the five dimensions. Table 3.5 summarises the overall problems over the five dimensions of health. It should be noted that some respondents may have some or severe problems in more than one dimension.

The smallest proportion of severe problems in health was seen with respect to mobility and the second smallest was with respect to self-care. More than half of the respondents had some problems in pain/discomfort followed by having some problems in anxiety/depression at almost half of the respondents. The highest proportion of no problems was seen in self-care followed by usual activities.

Table 3.5 EQ-5D given to their own health in the last 24 hours

EQ-5D dimension		no.	%
mobility			
	No problem	1038	73.63
	Some problem	364	25.80
	Severe problem	7	0.57
self-care			
	No problem	1287	91.38
	Some problem	104	7.36
	Severe problem	18	1.26
usual activities			
	No problem	1089	77.34
	Some problem	281	19.92
	Severe problem	39	2.74
pain/discomfort			
	No problem	493	35.05
	Some problem	885	62.77
	Severe problem	31	2.18
anxiety/depression			
	No problem	741	52.64
	Some problem	633	44.90
	Severe problem	35	2.46

To examine differences in characteristics of the respondents who considered themselves as having "good health" and those as having "fair or poor health", respondents with "good health" were defined as those who rated themselves as having 11111 or only one dimension with level 2 (11112, 11121, 11211, 12111 and 21111). Having some problems in one of any dimensions was considered as having "good health" because having some problems in only one dimension was unlikely to prevent the respondents from performing full functions in their everyday activities. The rest of the respondents were categorised as having "fair or poor health". Demographic characteristics of the respondents in both groups are presented in Table 3.6.

Table 3.6 Summary of characteristics of the respondents in “good health” and “fair or poor health”

Characteristics		Good health		Fair or poor health		Total
number		645	45.8%	764	54.2%	1,409
gender						
	male	320	49.7%	324	42.4%	644
	female	325	50.4%	440	57.6%	765
residential area						
	urban	242	37.5%	246	32.2%	488
	rural	403	62.5%	518	67.8%	921
education*						
	primary	328	56.5%	347	68.2%	675
	secondary	157	27.0%	107	21.0%	264
	university	96	16.5%	55	10.8%	151
	Total	581		509		

* some of education data are missing

Note that some of education data are missing. Therefore, out of 645 respondents who were classified as having “good health”, only 581 respondents had the data of education level whereas, out of 764 of those classified as having “fair or poor health”, only 509 respondents had the education data. Six hundred and forty five respondents (45.8%) have been classified as having “good health”. Half of the male respondents were classified as having “good health”. Regarding the respondents classified as having “good health”, 242 respondents lived in urban area (37.5%) and 403 (62.5%) in rural area. Regarding the educational attainment level, 56.5% had primary education level, 27.0 % and 16.5 % had secondary education and university education level, respectively. Note that the proportions of the respondents with secondary and university level who regarded themselves as having better health (good health) are higher than those with fair or poor health.

3.2.4 VAS scores representing health of the respondents

Two respondents considered their health to be in “the worst imaginable state as possible”. More than half of the respondents assigned scores of more than 80 to their own health with the greatest proportion of the respondents (26%) rated the score 80.

Only 15% rated their own health as "the best imaginable state as possible". Table 3.7 summarizes the scores of the respondents' own health using VAS method.

Table 3.7 VAS scores for own health

ownvas score	Freq.	Percent	Cum.
0	2	0.14	0.14
10	2	0.14	0.28
20	2	0.14	0.43
30	3	0.21	0.64
40	12	0.85	1.49
50	122	8.66	10.15
55	2	0.14	10.29
56	1	0.07	10.36
60	88	6.25	16.61
65	5	0.35	16.96
70	210	14.9	31.87
72	1	0.07	31.94
75	13	0.92	32.86
79	1	0.07	32.93
80	360	25.55	58.48
84	1	0.07	58.55
85	24	1.7	60.26
86	1	0.07	60.33
88	1	0.07	60.4
89	1	0.07	60.47
90	320	22.71	83.18
95	15	1.06	84.24
96	2	0.14	84.39
98	1	0.07	84.46
99	2	0.14	84.6
100	217	15.4	100
Total	1,409	100	

3.2.5 Interview duration

Interview durations according to the interview method are presented in Table 3.8. To examine the differences of the interview duration between adult and elderly respondents, the average interview durations according to age-group and interview method are also reported. On average, the overall interview duration was 56 minutes. Note that the mean duration includes only the respondents who completed all three elicitation methods. The minimum overall interview duration was 15 minutes and the maximum was 130 minutes. According to the type of interview method, the shortest interview duration was seen in VAS method and the longest was in TTO method. Compared with the elderly, the adult respondents tend to have shorter interview durations in all interview types although the differences were not statistically significant.

Table 3.8 Mean durations of the overall interview and each interview method according to age-group

		Mean (min)	SD	Min	Max
Overall		56.04	20.0	15	130
	adult	55.03	19.7	15	130
	elderly	63.44	20.8	29	123
Ranking		12.14	8.2	1	67
	adult	11.81	7.9	1	58
	elderly	14.57	10.0	3	67
VAS		6.76	4.3	1	57
	adult	6.60	4.2	1	57
	elderly	7.94	4.8	1	28
TTO		29.81	12.2	3	103
	adult	29.22	12.0	3	103
	elderly	34.04	12.9	12	79

3.2.6 Self-completed questionnaire

The respondents and the interviewers were asked to fill in their opinions regarding the interview process at the end of the interview. Sixteen and twenty percent of the respondents expressed that Ranking and VAS methods were difficult. The reasons were they could not understand the health states descriptions, and were thus unable to imagine themselves being in the health states. Half of the respondents (51%) admitted that the TTO method was difficult, 20% of them think so because of the same reasons as given to Ranking and VAS methods and 10% expressed that they cannot understand the method of trading-off time to live in full health. From the interviewer perspectives, they stated that 605 respondents (45%) were confident with the tasks, 627 respondents (47%) were confident after participating with the interview and only 8% were not confident with the interview at all.

Comments from the respondents from the open-ended question

In addition, the open-ended questions were used to allow the respondents and the interviewers to independently express their concerns about the interview methods. More reasons for the difficulties can be identified using this type of question. In general, the respondents were concerned that they could not differentiate between the health cards. They also expressed difficulty imagining themselves in the states as described in the cards because they had no previous experience of them. Some levels on the dimensions presented in the cards were thought to contradict each other. Some admitted that they cannot understand the interview tasks for the first two or three states, eventually, after assigning values for a couple of states, their level of understanding seem to increase. Some respondents were confused after reading "a number of" health cards and they were unable to compare the states with the previous ones because they cannot remember the scores previously given. Some reported feeling intimidated by the interviewers forcing them to choose only one state from the two.

Comments from the interviewers

Many of the interviewers reported that the respondents were confused with the tasks in the beginning, then the respondents gained more understanding of the tasks and eventually, the level of confidence in assigning values for health states increased. Although the respondents recruited for the interviews were, to some extent, literate, many of them needed a considerable time to assign scores to health states. Some misinterpreted the health states and they took into account the extra-information, for example, family members, to assist their decision making. Some respondents learned the "trick" of assigning values to health states. To complete the interview as soon as possible, some chose the answer "indifferent between the two states" without carefully compared the two states in the TTO method. They learned that, by considering two health states as "indifferent", the new states were introduced and they would complete the tasks shortly.

3.3 Data analysis

After the survey was finished, the next process was to transform the results from the recording forms preparing for the analysis. Three sub-sections of data management were reported in this section: data entry, the transformation methods of the TTO scores and numbers of the respondents excluded from the analysis. To examine nature of the actual scores, mean scores of each health state are calculated and the distributions of the actual scores for each state are tested for normality. Note that only the TTO scores were examined because the preference scores are modelled from the TTO scores. Influences of age and genders on the actual scores are explored at the end of this section.

3.3.1 Data entry

One research assistant was assigned to enter the results of the interview using the Microsoft Excel 2007 program. Codes were generated for the respondents and the interviewers. All data for each respondent were entered in the same rows. The data were then transferred using the Stat Transfer program, version 9, ready to be analysed using the Stata 10/SE program. The data were then rearranged, to prepare for the analysis. The scores from each respondent were converted from wide form into long form according to health states and separately categorised according to the elicitation methods. THe raw TTO scores were transformed using the following formulae.

For states better than death, the scores were:

$$\frac{X}{10}$$

For states worse than death, the score are:

$$\frac{-X}{(10-X)}$$

Where: X =number of years being in perfect health (43)

The lowest score for a state worse than death was -39. This score was assigned when the respondent preferred to die immediately over living in an inferior state for 6 months followed by living in perfect health for 9 years and 6 months. Therefore, the duration of living in an inferior state was 3 months (or 0.25 years) followed by living in perfect health for 9 years and 9 months (9.75 years). Therefore, the TTO score for this state was $\frac{-9.75}{(10-9.75)}$ which is -39.

3.3.2 TTO scores transformations

TTO scores for states worse than death in this study were transformed using the monotonic and linear transformation. The scores transformed using the monotonic transformation were prepared for the analysis using the Dolan 1997 and the Dolan & Roberts 2002 models. The linear transformation was used to prepare for the analysis using the Shaw *et al* 2005 model. Regarding the monotonic transformation, the lowest score was bound at -0.975 (58). The equation used in the transformation is as follows:

$$\boldsymbol{U' = \frac{U}{1-U}}$$

Where *U'* = the transformed TTO scores for states worse than death, *U* = the untransformed scores from the raw data where the scores for state worse than death (59).

To the best of my knowledge, the US is the only country where the linear transformation was used in the utility score estimation model; the USD1 model (51). The equation for the linear transformation is as follows.

$$\boldsymbol{U' = \frac{U}{39}}$$

Where *U'* = the transformed TTO scores for states worse than death, *U* = the untransformed scores from the raw data where the scores for state worse than death (51).

3.3.3 Numbers of respondents excluded from the data

The exclusion criteria were following those used in the MVH protocol (43). Those excluded were: (1) the respondents with completely missing values for every state; (2) those who assigned values for fewer than 3 states; (3) those who assigned the same values for every health state; and (4) those who assigned scores for all states as worse than death. Out of 1,409 respondents, the numbers of respondents excluded according to the elicitation method are presented in Table 3.9.

Table 3.9 Numbers of the respondents excluded from the data and the causes of the exclusion

Causes	Ranking	VAS	TTO
Completely missing values	13	1	4
Give values to fewer than 3 states	0	2	9
Same values for all states	0	1	8
Value all states as worse than death	NA	NA	2*
No. of respondents after the exclusion	1,396	1,392	1,371

* these 2 respondents also assigned same values for all states

NA = Non applicable

Thirteen respondents were unable to assign any value to the health state in Ranking, one in VAS and seven in TTO. Note that the respondents who had completely missing values in Ranking are higher than that of the VAS and TTO methods because the interviews were terminated in the middle or after the Ranking method. Therefore, of all 1,409 respondents, thirteen respondents had completely missing values in the Ranking method and were excluded from the interview. Of all 1,396 respondents who completed the Ranking interview and moved to undertake the VAS task, one respondent had completely missing scores in the method, two respondents gave values to fewer than three states and one respondent gave same values for all states. There were 1,392 respondents completed the VAS interview and moved to the TTO method. Of those, four respondents had completely missing values, nine respondents gave the scores to fewer than three states and eight respondents gave the same values for all states. Of all these eight respondents, two respondents considered all health states to be worse than death. As a result, 1,371 respondents completed the TTO interview. It should be noted that there was no state identified as worse than death in the Ranking and VAS methods because "immediate death" was not used in the two methods.

3.3.4 Mean actual TTO scores

Mean TTO scores of all 86 states are shown in Table 3.10. Note that the monotonic transformation was used to transform the TTO scores in this table.

Table 3.10 Mean TTO scores of the states used in the interview

State	Mean TTO scores from the respondents				
	n	Mean	SDs	Min	Max.
11112	314	0.705	0.302	-0.525	1.000
11121	237	0.684	0.308	-0.775	1.000
11122	109	0.674	0.342	-0.725	1.000
11211	312	0.667	0.322	-0.975	1.000
11212	118	0.570	0.398	-0.975	0.996
11221	111	0.641	0.321	-0.425	1.000
11222	112	0.476	0.474	-0.975	1.000
11223	114	0.428	0.438	-0.975	0.996
11232	120	0.583	0.363	-0.975	1.000
11313	115	0.360	0.483	-0.975	1.000
11332	95	0.354	0.486	-0.975	1.000
12111	222	0.645	0.321	-0.675	1.000
12112	211	0.602	0.379	-0.925	1.000
12121	106	0.478	0.398	-0.675	1.000
12122	119	0.478	0.444	-0.875	1.000
12123	111	0.391	0.482	-0.975	0.997
12211	118	0.582	0.366	-0.975	1.000
12212	113	0.483	0.438	-0.975	0.996
12221	209	0.515	0.396	-0.975	0.996
12312	98	0.397	0.481	-0.800	1.000
12313	120	0.280	0.521	-0.925	0.996
12331	107	0.247	0.511	-0.975	1.000
13123	92	0.277	0.503	-0.975	1.000
13222	112	0.196	0.516	-0.975	0.996
13232	119	0.106	0.506	-0.975	0.996
21111	232	0.667	0.334	-0.975	1.000
21112	96	0.628	0.323	-0.525	1.000
21121	209	0.594	0.365	-0.975	1.000
21122	108	0.572	0.391	-0.625	1.000
21123	104	0.340	0.507	-0.975	0.996
21211	111	0.604	0.410	-0.975	1.000
21212	113	0.570	0.386	-0.975	1.000
21221	106	0.421	0.504	-0.975	1.000
21231	104	0.278	0.513	-0.975	1.000
21312	119	0.455	0.437	-0.875	0.996
21313	115	0.253	0.481	-0.975	0.996
21331	116	0.175	0.488	-0.975	0.996
21332	213	0.250	0.525	-0.975	1.000

Table 3.10 Mean TTO scores of the states used in the interview (continued)

State	Mean TTO scores from the respondents				
	n	Mean	SDs	Min	Max.
22111	112	0.518	0.440	-0.950	0.996
22112	216	0.472	0.449	-0.975	0.996
22113	97	0.384	0.464	-0.975	1.000
22121	99	0.469	0.482	-0.975	1.000
22211	120	0.492	0.479	-0.975	0.996
22221	112	0.385	0.468	-0.975	0.996
22232	114	0.131	0.527	-0.975	0.996
22233	119	-0.003	0.541	-0.975	0.996
22313	95	0.260	0.473	-0.975	1.000
22323	209	0.167	0.553	-0.975	0.996
22332	102	-0.017	0.574	-0.975	1.000
22333	98	0.056	0.497	-0.975	1.000
23113	112	0.154	0.520	-0.975	0.987
23131	100	0.050	0.531	-0.975	1.000
23132	101	-0.009	0.511	-0.975	0.996
23222	118	0.327	0.507	-0.975	1.000
23223	214	0.078	0.571	-0.975	1.000
23231	112	-0.008	0.530	-0.975	0.996
23232	99	0.020	0.500	-0.925	1.000
23233	114	-0.134	0.509	-0.975	0.996
23321	104	0.126	0.531	-0.975	0.996
23322	112	0.025	0.541	-0.975	0.950
23323	113	0.019	0.573	-0.975	1.000
23332	101	-0.129	0.547	-0.975	0.996
23333	318	-0.119	0.492	-0.975	0.996
31131	112	-0.025	0.529	-0.975	0.996
31213	98	-0.013	0.535	-0.975	0.996
31222	111	0.000	0.546	-0.975	0.979
31311	108	0.160	0.559	-0.975	0.996
32123	113	-0.085	0.519	-0.975	0.971
32223	109	-0.213	0.530	-0.975	0.996
32232	110	-0.134	0.497	-0.975	0.996
32233	121	-0.215	0.488	-0.975	1.000
32322	117	-0.124	0.513	-0.975	0.996
32323	210	-0.192	0.512	-0.975	0.996
32332	120	-0.155	0.513	-0.975	0.996
32333	233	-0.282	0.469	-0.975	0.925
33121	104	-0.131	0.559	-0.975	0.996
33122	108	0.002	0.520	-0.975	0.996
33221	110	-0.178	0.506	-0.975	0.996
33222	109	-0.028	0.513	-0.975	0.996
33223	113	-0.117	0.485	-0.975	0.996
33232	109	-0.303	0.459	-0.975	0.971
33233	314	-0.251	0.475	-0.975	0.996
33322	111	-0.233	0.526	-0.975	0.975
33323	221	-0.268	0.486	-0.975	1.000
33332	226	-0.318	0.441	-0.975	0.996
33333	1313	-0.346	0.454	-0.975	1.000

n=number of observations

The numbers of observations for each state ranged from 95 to 1,313. State 33333 had the greatest number of observations because this state was used in every health set. The highest mean TTO score was 0.705 given to state 11112 and the lowest score was -0.346 to state 33333. Mean scores of almost 30% (27 states) of the total number of health state were negative. Almost all the states had the lowest score of -0.975 including health states without level 3 in any dimensions, for example, states 11211 and 11212. Some respondents assigned score 1 for state 33333 or state 33323 even though these states were theoretically considered to be very extreme states.

3.3.5 Normality test

Shapiro-Francia test was used to test the normality of the TTO scores distribution (60). Stata program was used to calculate the z-statistics to test the null hypothesis of normal distribution. The scores of only six states were normally distributed (p-value > 0.05) whereas those of the other 80 states were skewed. The severity of the skewness of the distribution of the scores was measured and arbitrarily classified into mild, moderate and severe. Numbers of health sates in each category are shown in Table 3.11.

Table 3.11 Degree of skewness and numbers of states in each category

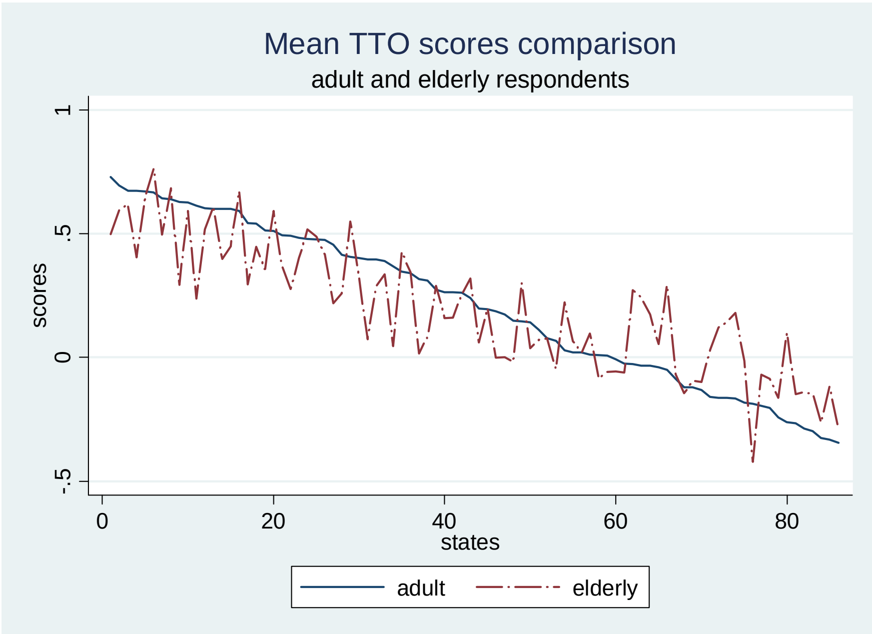
Degree of skewness	z-statistic	p-value	no.of state
Mild	1.667-2.198	0.048-0.014	12
Moderate	2.326-3.091	0.010-0.001	13
Extreme	3.239-7.189	0.0006-0.00001	55

The z-statistics of lower than 2.3 were classified as mild, those between 2.3 and 3.1 were moderate and those higher than 3.1 were extreme. The distributions of more than half of the states were extremely skewed (55 out of 86 states with the Z-statistics greater than 3.1). The states with one of level 2 and level 1 in other dimension (mild states) and those with no level 1 in any dimensions (severe states) tend to have highly skewed distributions. The mild states tend to skew to the left (more states with positive scores) and the severe states tend to skew to the right (more states with negative scores).

3.3.6 Mean TTO scores according to age-group

To examine the influences of age on mean TTO scores, the respondents were classified into 2 groups: adult (< 60 years old) and elderly (60 and older), t-test was used to compare the mean TTO scores of both groups. The comparison of all TTO scores according to age-group is shown in Figure 3.1.

Figure 3.1 Comparison of mean TTO scores according to age-group



The Y-axis represents the actual scores and the X-axis represents health states ranked from the best to the poorest states (using the adult mean scores). A solid line represents mean TTO scores assigned by the adult respondents. Compared with the scores assigned by adults, the elderly tend to assign lower scores for better states and higher scores for poorer states. The scores of 12 health states (15%) were significantly different between the two groups (p-value < 0.05). Those states are shown in Table 3.12.

Table 3.12 TTO scores with significant different between the elderly and adult groups

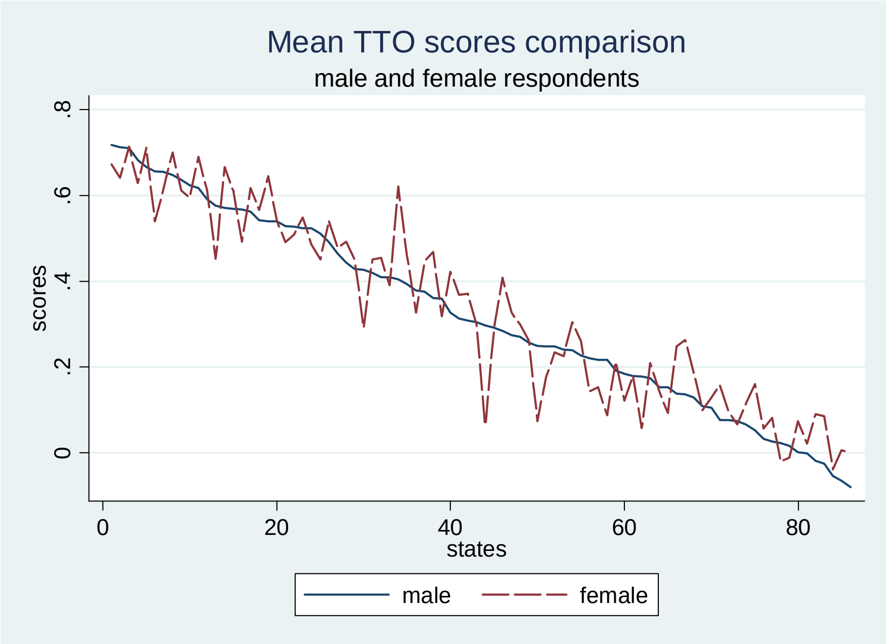
state	adult	elderly	p-value
11112	0.729	0.499	0.000
11232	0.628	0.293	0.001
21212	0.613	0.238	0.001
11221	0.673	0.405	0.004
32223	-0.263	0.100	0.013
23222	0.368	0.045	0.020
11313	0.397	0.073	0.022
12313	0.318	0.016	0.035
12221	0.543	0.296	0.037
21312	0.492	0.276	0.043
11212	0.601	0.398	0.046
23332	-0.166	0.180	0.047

The second column represents the mean TTO scores of adults and the third column represents those of the elderly. There were two states; 32223 and 23332, for which, on average, the elderly assigned higher scores than did the adults. Note that these two states could arguably be considered as the poorer states given that there is no level 1 in any dimension.

3.3.7 Mean TTO scores according to gender

A t-test was used to compare the differences between the mean TTO scores assigned by the male and female respondents. Comparison of all TTO scores according to gender is illustrated in Figure 3.2.

Figure 3.2 Comparison of mean TTO scores according to gender



The Y-axis represents the actual scores and the X-axis represents health states ranked from the best to the poorest states (according to the mean male scores). A solid line represents mean TTO scores assigned by the male respondents. It appears that, the scores assigned by both male and female respondents were similar. Mean TTO scores of four health states (0.05%) were significantly different between male and female respondents (p-value < 0.05). The states with significant differences are shown in Table 3.13.

Table 3.13 Mean TTO scores with significant different between male and female respondents

State	Male	Female	p-value
11222	0.322	0.591	0.003
23323	0.176	-0.129	0.004
31131	0.098	-0.145	0.015
22112	0.551	0.410	0.022
22332	0.102	-0.132	0.039

The second column represents mean TTO scores assigned by the male respondents and the third column by the female respondents. Of the five states shown in Table 3.13, the female respondents, on average, assigned higher scores to only one state (11222), compare with those assigned by the male respondents.

3.5 Results of the exploratory qualitative study

The thematic approach is used to understand how this group of respondents understand the TTO task and how they assigned values to health states. Three major themes were identified. Firstly, the respondents may not be able to imagine themselves living in hypothetical health states. Some refuse to believe that, for states worse than death, after living in a poorer state for some time, they can recover completely and stay in perfect health for some years. As seen in the following quotes, some respondents cannot differentiate the differences between two cards in the beginning of the task and develop the understanding later in the interview. This finding is in line with what Patrick *et al.* reported in 1994.

> *"I was confused at the beginning of the interview, but when I compared this card with that card after carefully reading both of them, now I understood that these two cards, in fact, differ.*

> *"I thought I have made a mistake. I thought this card is more severe than that card"*

> *"(reading aloud) some problems in walking-but I have no problem with my walking. I never use a tricycle (a common mean of commuting in a village). I think walking is one method I can use as an exercise. I have no problem with walking at all"*

> *"I have no experience in this health state. How can I imagine myself being in such a poor state"*

Secondly, the respondents used extra-information meaning that the respondents used the information which may not relate to the health states described in the cards to contribute with their decision making to trade-off time. After reading the cards, the respondents may not only imagine only themselves being in the states, but they also consider their family members into the scenarios. Nobody would have helped them if they are ill.

For example, the respondents care about the family members who are going to take care of them if they have to be confined to bed. So if there is choice involving being confined to bed, they decided to die immediately.

"If I lived in that situation, I would not have money to treat myself. I am poor and have no job. My children live far away. I do not want to be a burden for my children. How can they earn money to pay my medicines?"

"I don't want someone to take care of myself when I stay in bed. By causing burdens on my children, it is sinful"

"If I have a good family, when I fell sick, my grand children would come to take care of me"

On the contrary, some want to live their lives as long as possible because they want to see their children (and grand-children) grown up and live their lives. In contrast, some participants may have used only partial information from health cards. They may consider only the "key" dimension (61). For example, mobility is crucial for participant *i* who is young and energetic. However, for participant *j*, anxiety/depression dimension is a key element for his/her well-being. As long as the key element is in level 1, no matter which levels are for other dimensions, they may assign high values for those states. Some of the respondents used only part of the information to make a decision. Some respondents used only the first line of the health cards.

"*I read only the first line because I'm getting tired when I continue to read the 2nd and 3rd line. So I use only the first line to imagine myself with"*

"I want to live as long as possible no matter how bad my health state is. I want to live even though I am confined to bed because I want to see how my children get on with their lives"

The respondents may have their own "key" dimension that they do not want to suffer if this "key" dimension is at the extreme level (level 3).

"*I don't want to feel anxiety/depression. If it is extreme in this anxiety/depression, I just want to die immediately even though it is only some problem in mobility"*

The third theme is that the respondents cannot understand the TTO question for states worse than death. It is not uncommon to see the respondent immediately choose the perfect health without taking into account that she has to accept the situation that she has to stay in the inferior health states before living in perfect health. Some chose the health states on the basis of the number of years they can live without taking into account the type of health states they are going to stay.

3.5 Discussion

This chapter reports the results of the fieldwork survey. The survey was successfully administered and a broadly representative sample of the Thai general population was interviewed. One reason for this success was the provision of interview sites that were easy to travel to and, where necessary, allowing interviews to be conducted in the respondents' household or workplace. The target numbers of respondents from all regions were reached except in the South region where the actual number of interviews was slightly lower than expected. One reason for the lower number could be because although the exact number of respondents was identified and all of them were successfully contacted by the field coordinators, when it came to the interview dates, some of the respondents failed to present at the appointed interview sites because of unexpected engagements. This problem was evaluated to prevent the same problem happening again in the interview in the other regions. The researcher identified a slightly higher number of expected respondents before sending names to the field coordinators. The increased numbers of respondents were carefully considered because there would be an impact on the budget if more respondents appeared than were expected.

Compared with the Thai general population, a larger proportion of female respondents and respondents living in urban areas were seen in the Thai sample. On average, the mean age of the respondents was higher than that of the Thai general population. This is in line with the findings from the US, UK and Spanish studies where larger proportions of the respondents were female.

Not all respondents rated their own health as full health (11111) or assigned the score 100 to the VAS. Almost half of the respondents were classified as having "good health". The respondents living in rural areas and those having primary education tended to report some problems in their health. The highest proportion of respondents had no problem in self-care followed by usual activities. Given that most of the respondents were requested to travel to the arranged interview sites and one of the inclusion criteria of the respondents was that the respondents were able to communicate with the interviewers, it would be almost impossible to identify respondents with severe problems in these dimensions because if they had they would not have been able to travel to the interview sites. However, most of the respondents had some problems in one or more dimensions. The respondents in the Thai sample may have had experience with some degree of sickness. As a result, they may have had some background understanding of the difficulties described in the health states when they assigned scores to the health states.

By reducing the number of health states to eleven, the interview could be conducted, on average, within one hour. As expected, the longest part of the interview was for the TTO interview because of the complexities of the task. A considerable number of respondents could not understand the task in the beginning but they gained more understanding and confidence with trading-off time. This is in line with what the interviewers observed and comments made in response to the open-ended questions. The burden was still presumably high for Thai respondents although the interview protocol was redesigned from the original.

Note that the numbers of respondents per health set presented in Table 3.4 are smaller than that expected from the sample size calculation in Chapter 2. From the calculation, at least 200 observations per health state are required to give the meaningful differences of 0.1 between two health states at the significance level of 0.05. The smaller number of observations is justified in this study for the following three reasons. Firstly, given the results from the pilot studies that Thais may not be able to cope with the preference interview using more than 11 health states (including state 11111 and 33333), to be able to achieve 200 observations per *health set*, at least 1,720 respondents or, that is, 300-400 more respondents would have been needed to be interviewed. Secondly, it was decided to include a larger number of health states in the Thai study, namely 86 health states. Given the limited budget, availability of interviewers and time available for the fieldwork, approximately 1,400 respondents was the best that could be achieved. In practice, not all health states had less than 200 observations, as can be seen in Table 3.10, there are twenty percent of the total health states with more than 200 observations. This is because some health states were included in more than one health set and state 33333 was used in all health sets. Only eight states have less than 100 observations.

Thirdly, some authorities have argued that a minimum of thirty-five observations per health state is acceptable(39). This study achieved almost three times that number. Furthermore, the numbers of observations per health state in the Thai data are not greatly different from the Korean study (154 observations per health state) and the Dutch study (167 observations per health state). However, the number of health states included in the Thai study was twice that of the Korean study and five times that of the Dutch study.

An initial understanding of the coping mechanisms employed by the respondents was revealed by conducting the exploratory qualitative study. It is possible that the Thai respondents applied extra information or may have used only partial information about health states in making their decisions on sacrificing time in the TTO. The elderly respondents with primary education tend to have difficulty imagining living in the health states given in the interview. This may result from the lower level of reading competency in this group of respondents. If they have difficulty reading the health cards, it is likely that they would have had problems in reading and understanding the explanation of the trading-off of time on the TTO boards, especially in the complicated questions for states worse than death. Some respondents reported that they "learned" to respond to the questions after answering the first couple of questions. From this finding, an interesting question would be whether the "learning effects" plays some role in the elicitation interview.

Results from this study suggest a number of ways in which it might be possible to reduce cognitive workloads on respondents in future preference studies. The age and education findings provide a justification for adjusting the amount of information collected from individual respondents belonging to different sub-groups. That is, studies could recruit a larger number of older respondents but ask them fewer questions. To reduce the influence of individual interviewers, a computer-assisted interview with a prompt when logical inconsistency is identified, may reduce the number of logical inconsistent pairs. If a face-to-face paper-based interview is going to be conducted, interviewers could be more extensively trained. Experienced interviewers are favoured. The health states used in any one interview should be carefully selected, with a view of avoiding combinations of states that might be particularly hard for respondents to differentiate between. Plausibility of health states should be taken into account.

The next chapters will explore the effect of exclusion of respondents with different numbers of inconsistent pairs on the mean actual scores, the predicted scores for the health states and the coefficients of the model predicting utility scores for the EQ-5D health states from the Thai data.

3.6 Conclusion

A total of 1,409 respondents were interviewed during May – August 2007. The mean age of the respondents was 44.6 years and the proportion of female respondents was slightly higher than that of male respondents. Compared with the Thai general population, females, adults and those living in urban areas seem to be over-sampled. The overall interview duration was about 56 min with the longest time being taken over the TTO interview. Elderly respondents and those with primary level education tend to have longer interview durations. The distributions of the actual TTO scores of almost all health states interviewed were skewed. The elderly tend to assign lower scores for mild states and higher scores to poorer states. Gender has no effect on preference scores for almost all of the health states. What has been learned in the survey can guide future studies regarding the number of health states that Thai respondents can cope with in preference interviews and the types of interviews that can be conducted with Thai respondents.

Chapter 4 Logical inconsistency and the selection of scores from a respondent subgroup to estimate preferences on health

4.1 Introduction

The mean TTO score for each health state were calculated in the previous chapter. It is shown that some scores were not consistent with the severity of health states and mean scores of some better states are lower than some poorer states. The extent of inconsistent responses is examined in this chapter. Health states are inconsistently valued if a higher score is assigned to a worse health state. This implies that, to detect logical inconsistency, a pair-wise comparison between two scores is needed. Not all pairs of health states can be used to identify logical inconsistency. An eligible pair is the two health states with at least, one dimension, is lower or better than a corresponding dimension in the other state, given other dimensions being equal. For example, considering 2 scenarios:

<u>Scenario 1</u>: the pair-wise comparison between state 1**21**21 and 1**12**21

In these two states, mobility, pain/discomfort and anxiety/depression dimensions are at the same levels. Differences are in self-care and usual activities in that self-care in the first is worse (level 2) than the second state but usual-activities is better (level 1). In the second state, self-care is better but usual-activities is worse. A respondent would assign higher value for the first state if he/she prefers better level in usual activities. Others may prefer better self-care, thus higher value is assigned to the second state. Therefore, this pair cannot be used to detect logical inconsistency because different respondents may have different preferences.

<u>Scenario 2:</u> the pair-wise comparison between state 1122**1** and 1122**2**

Difference between the two states is only at anxiety/depression. Logically, the first state is better because there is no problem in the anxiety/depression in the first state. A respondent is assumed to prefer the first state and assign higher value than the second state. If a respondent assign higher value to the second, these two values are labelled as logical inconsistency.

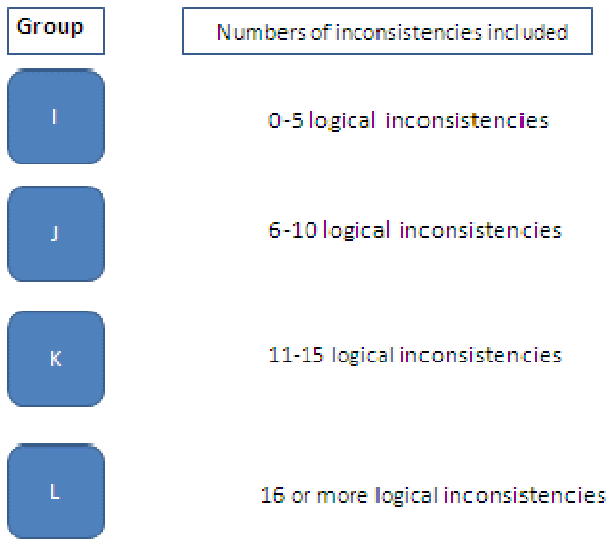
In this report, the term "logical inconsistency" is singular and used to identify inconsistent values given to a pair of health states. If there are two pairs of health states with inconsistent values in each pair, they are labelled as two logical inconsistencies. The Dolan and Kind approach is followed here. Dolan and Kind estimated the inconsistency rate as the number of pairs of health states with inconsistent responses expressed as a proportion of the number of pairs of health states that could have been inconsistently valued (62). Badia *et al*. and Devlin *et al*. also identified inconsistency in this way (63).

To develop a Thai tariff for the EQ-5D requires the estimation of a model using "valid" scores, or scores which to some extent represent preferences over health states. However, logically inconsistent scores were identified and at least some of these may not be a valid representation of the preferences of the respondent. The extent to which inconsistent respondents should be excluded from the model estimations will depend partly on the impact of these inconsistencies on the mean scores. This chapter sheds light on these effects. The outline of this chapter is as follows. Firstly, the methods used to examine the effect of excluding the data from respondents exhibiting differing levels of inconsistency are described. Secondly, the effect of the exclusion of logical inconsistencies is thoroughly explored. The most appropriate group of respondents to use when estimating the Thai tariff is chosen at the end of the chapter.

4.2 Examination of the validity of the scores

The TTO scores are used in the analysis of the effects of exclusion because the Thai preference scores are to be estimated from the TTO data. The validity of the scores is examined based on the assumption that the respondents with fewer numbers of inconsistencies are those who can assign TTO scores according to the severity of health states and the higher scores are assigned to better states and lower scores to poorer states. To examine the validity of the scores, the respondents are arbitrarily divided into four groups according to the numbers of logical inconsistent responses as presented in Figure 4.1. Group I consists of the respondents with 0-5 inconsistent responses. Groups J, K and L consist of the respondents with 6-10, 11-15 and 16 or more inconsistencies, respectively.

Figure 4.1 Four respondent groups with various numbers of inconsistencies

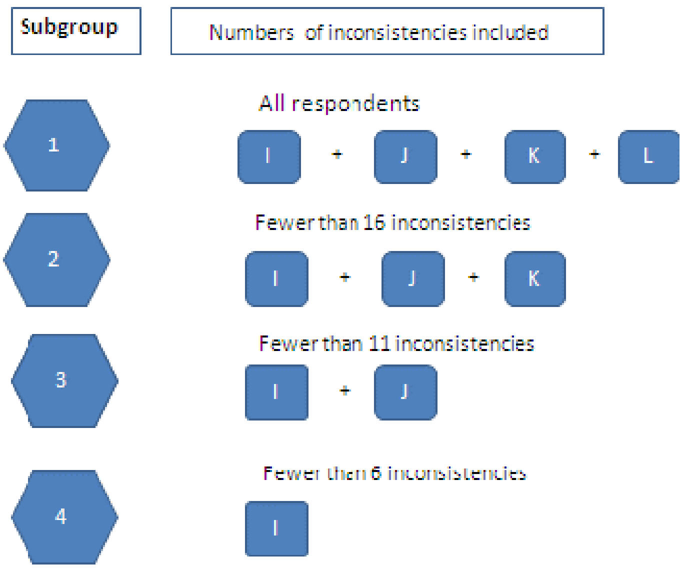


The respondents in Group I are assumed to be the ones who are most likely to have assigned "valid" scores reflecting their preferences over health states because this group consists of the respondents who assigned scores characterised by the least inconsistency. These scores are assumed to be a potentially robust basis upon which to base the EQ-5D tariff for Thailand.

4.3 Examination of the impacts of excluding data from inconsistent respondents

To explore the impacts of excluding the scores from the inconsistent respondents on the mean scores, all respondents formed the respondent subgroup 1. Therefore, subgroup 1 comprises the Groups I, J, K and L. The respondents from Group L who had more than fifteen inconsistencies were excluded to form the respondent subgroup 2. Those who had more than ten inconsistencies (Groups K and L) are excluded to form subgroup 3 and only the respondents in Group I formed subgroup 4. The four subgroups of respondents thus generated are shown in Figure 4.2.

Figure 4.2 Four respondent subgroups and numbers of inconsistencies included



Subgroup 1 is regarded as a reference group because they are the scores from all respondents who participated in the study. Mean scores for Subgroups 2-4 are compared with those in Subgroup 1 and the differences in mean scores from each subgroup are explored. Spearman rank correlation coefficients are used to investigate the correlations between the ranks of the mean scores of different subgroups. As shown in the previous chapter, elderly respondents and those who have primary level education exhibit more logical inconsistencies. Demographic characteristics and average interview durations are reported for the different subgroups. These factors could be used when trying to justify the exclusion of the scores from inconsistent respondents.

4.3 Results

4.3.1 Demographic characteristics of the respondents in all four subgroups

Demographic characteristics of the respondents in the four subgroups (subgroup 1-subgroup 4) are compared with those of the Thai general population in Table 4.1. Compared with the general population, all subgroups have higher proportions of females and respondents living in urban areas, and a lower proportion of the elderly. Note that the respondents in subgroup 4, have lower proportion of the respondents with primary education level compared with that of the general population. Almost ten percent of all respondents had more than fifteen inconsistencies and were excluded from subgroup 1. Twenty percent and fifty percent of all respondents had more than ten and more than five inconsistencies respectively.

4.3.2 Mean scores of the respondents with various numbers of inconsistencies

The number of observations and the mean TTO scores for the eighty-six health states in the respondents from Group I-L are reported in Table 4.2. The mean scores were ranked from highest to lowest according to the scores in Group I. The mean scores of health states in Group J, Group K and Group L are compared with the corresponding states in Group I because Group I consists of the respondents with fewest number of inconsistencies. As a result, three comparisons are made; Group J-Group I, Group K-Group I and Group L-Group I. A t-test is used to investigate the statistical significance of differences between the groups. An asterisk indicates the health states where the difference in means is statistically significant difference at $p < 0.05$.

Only 24 per cent (21 states) of the mean scores of the respondents in Group J are significantly different from the mean scores of the corresponding health states given by the respondents in Group I. Almost 40 per cent (34 states) of the scores from the respondents in Group K, and 45 per cent (39 states) of the scores from Group L, are significantly different from the scores of the respondents in Group I.

Only the respondents with more than fifteen inconsistencies (Group L) assigned a positive score to state 33333. The score for this state was also higher than the score for states 33323 and 11222. The score assigned to state 11112 by the respondents in Group L is half of the score assigned by the highly consistent respondents (Group I). The lowest score was assigned to state 33333 and the highest score to state 11112 by the respondents in Group I.

Table 4.1 Demographic characteristics of the respondents in all four subgroups

Participant characteristics		Thai general population** (x 1,000,000)	%	The samples Subgroup 1 number	%	Subgroup 2 number	%	Subgroup 3 number	%	Subgroup 4 number	%
Number		62.80	100	1,324	100	1,218	91.99	1,074	81.12	632	47.73
Gender											
	Male	31.01	49.30	553	45.40	600	46.01	498	46.37	299	47.31
	Female	31.82	50.67	665	54.60	704	53.99	576	53.63	333	52.69
	proportion		1.00		0.83		0.85		0.86		0.90
Mean age(yrs.)	(SD)			44.2	(12.50)	43.85	(12.45)	43.4	(12.27)	42.56	(11.81)
Age-group											
	Adult (20-59)	37.30	85.00	1,162	87.76	1,082	89.00	962	89.57	575	91.00
	Elderly (60+)	6.60	15.00	162	12.24	136	11.96	112	10.43	57	9.00
	proportion		5.67		7.17		7.44		8.59		10.11
Education*											
	Primary	20.48	58.00	841	63.52	753	61.82	646	60.15	353	55.85
	Secondary	9.78	27.80	264	19.94	254	20.85	235	21.88	155	24.53
	University	5.01	14.2	151	11.4	144	11.82	132	12.29	91	14.4
						missing data=67		missing data= 61		missing data = 33	
Residential	area										
	Urban	19.60	30.70	454	34.29	423	34.73	372	34.64	231	36.55
	Rural	45.40	69.30	870	65.71	795	65.27	702	65.36	401	63.45
	proportion		0.4		0.52		0.53		0.53		0.58

** Source: The Key Statistics 2007, National Statistical Office, Bangkok, Thailand

Note: Subgroup 1 = all respondents, Subgroup 2 = those with 0-15 inconsistencies, Subgroup 3= those with 0-10 inconsistencies, Subgroup 4=those with 0-5 inconsistencies

Table 4.2 Mean scores assigned by the respondents with various numbers of inconsistencies

State	Mean TTO scores from the respondents in							
	Group I		Group J		Group K		Group L	
	n	Mean	n	Mean	n	Mean	n	Mean
11112	133	0.815	118	0.689*	43	0.511*	20	0.491*
12111	97	0.768	83	0.602*	18	0.430*	24	0.457*
11121	121	0.766	70	0.682*	19	0.528*	27	0.427*
11122	47	0.765	44	0.642*	14	0.466*	4	0.699
11211	144	0.742	111	0.628*	33	0.575*	24	0.515*
21111	116	0.742	66	0.642*	30	0.637	20	0.364*
21112	49	0.736	31	0.548*	3	0.715	13	0.390*
21121	87	0.725	76	0.602*	30	0.222*	16	0.545*
12211	62	0.721	29	0.427*	18	0.429*	9	0.427*
11212	50	0.719	41	0.564*	12	0.314*	15	0.298*
21211	54	0.692	39	0.645	13	0.446*	5	-0.240*
21122	47	0.685	43	0.537	14	0.345*	4	0.425
11221	74	0.684	29	0.618	3	0.233*	5	0.394
12112	94	0.680	78	0.592	29	0.453*	10	0.380*
11232	62	0.659	29	0.532	17	0.461*	12	0.489
22211	62	0.633	29	0.274*	18	0.437*	11	0.363*
12212	46	0.631	41	0.478	16	0.199*	10	0.273*
22121	50	0.627	31	0.255*	4	0.729	14	0.301*
12121	47	0.625	42	0.350*	6	0.369	11	0.401
11222	54	0.614	39	0.527	13	0.006*	6	-0.092*
22111	46	0.614	40	0.562	16	0.369	10	0.138*
12221	87	0.608	75	0.493*	30	0.318*	17	0.485
21212	75	0.603	30	0.601	3	-0.058*	5	0.254*
21221	46	0.584	42	0.423	6	0.258	12	-0.127*
22112	95	0.573	79	0.493	30	0.091*	12	0.483
21312	49	0.554	42	0.424	12	0.275*	16	0.371
12312	39	0.543	37	0.286*	13	0.212*	9	0.488
12122	50	0.534	39	0.520	12	0.339	18	0.323
22221	46	0.485	39	0.435	16	0.184*	11	0.084*
11223	75	0.452	30	0.427	3	0.153	6	0.279
12123	53	0.450	39	0.446	13	0.042*	6	0.262
22113	50	0.439	30	0.302	5	0.309	12	0.393
11313	47	0.420	39	0.390	17	0.207	12	0.248
21123	46	0.376	42	0.268	6	0.276	10	0.519
13123**	48	0.347	44	0.200	NA	NA	NA	NA
13222	54	0.329	39	0.216	13	-0.287*	6	-0.088*
31311	47	0.314	43	0.041*	14	0.016	4	0.134
12313	62	0.297	29	0.213	17	0.300	12	0.327
11332	50	0.291	30	0.366	4	0.630	11	0.503
23222	62	0.288	29	0.315	17	0.394	10	0.489
22313	49	0.285	31	0.148	5	0.240	10	0.500
21313	48	0.269	39	0.239	17	0.172	11	0.360
21231	47	0.268	35	0.248	11	0.266	11	0.431
21332	86	0.256	81	0.183	28	0.330	18	0.406

Table 4.2 Mean scores assigned by the respondents with various numbers of inconsistencies (continued)

State	Mean TTO scores from the respondents in							
	Group I		Group J		Group K		Group L	
	n	Mean	n	Mean	n	Mean	n	Mean
12331	47	0.238	43	0.327	13	0.037	4	0.181
23321	46	0.222	35	-0.078*	13	0.255	10	0.231
23113	75	0.200	29	0.057	4	-0.175	4	0.335
22323	87	0.193	75	0.150	31	0.006	16	0.420
21331	49	0.131	39	0.264	12	0.323	16	-0.019
31222	46	0.106	40	-0.119	15	-0.110	10	0.153
23223	96	0.103	79	-0.028	28	0.137	11	0.472*
33122	47	0.102	43	-0.194*	14	0.073	4	0.680*
31131	47	0.052	43	-0.203*	17	0.180	5	0.075
23131	45	0.050	41	-0.007	7	0.125	7	0.314
13232	50	0.035	41	0.082	13	0.316	15	0.226
23322	75	0.015	29	0.008	3	0.225	5	0.150
23231	46	-0.002	41	-0.082	16	-0.049	9	0.372*
22232	46	-0.011	42	-0.031	5	0.045	9	-0.017
22332	46	-0.011	42	-0.031	5	0.045	9	-0.017
33222	47	-0.028	44	-0.099	14	-0.014	4	0.705*
31213	39	-0.032	37	-0.084	13	0.006	9	0.338
23132	46	-0.046	41	-0.034	6	-0.167	8	0.452*
23323	48	-0.052	44	0.039	16	0.091	5	0.295
23232	50	-0.053	31	-0.032	5	0.809*	13	0.121
22233	62	-0.068	28	-0.026	17	0.223*	12	0.070
22333	50	-0.078	31	-0.008	4	0.617*	13	0.549*
32123	75	-0.100	29	-0.075	4	0.206	5	-0.140
33221	54	-0.109	39	-0.291	12	-0.269	5	0.189
33223	47	-0.129	44	-0.243	14	0.016	8	0.405*
23233	46	-0.149	41	-0.280	16	-0.019	11	0.309*
32232	45	-0.151	40	-0.203	16	-0.105	9	0.203
33121	47	-0.168	35	-0.239	13	0.114	9	0.127
32223	54	-0.179	37	-0.242	13	-0.270	5	-0.210
23333	134	-0.225	120	-0.147	43	0.033*	21	0.408*
32322	50	-0.252	41	-0.168	12	0.224*	14	0.161*
23332	46	-0.268	42	-0.057	4	-0.106	9	0.238*
32323	87	-0.275	76	-0.237	31	-0.192	16	0.476*
33322	75	-0.290	29	-0.156	2	-0.238	5	0.190*
33232	54	-0.315	38	-0.330	12	-0.302	5	0.040
32333	122	-0.338	70	-0.355	19	-0.172	22	0.166*
33233	147	-0.368	110	-0.279	34	0.017*	23	0.244*
32233	62	-0.371	29	-0.236	17	0.056*	13	0.217*
33332	116	-0.389	68	-0.360	31	-0.142*	11	0.197*
33323	98	-0.405	83	-0.193*	18	-0.182*	22	-0.013*
32332	51	-0.429	40	-0.083*	13	0.171	16	0.271*
33333	614	-0.480	454	-0.333*	142	-0.139*	90	0.297*

* statistical significant difference from Group I (p-level <0.05)** the number of inconsistencies of this state ranges from 0-10.Note: Gr. I = resp. with 0-5 incons, Gr.J =resp. with 6-10 incons, Gr.K=resp.with 11-15 incons, Gr.L=resp. with 16 or more incons. n=numbers of observations.

4.3.3 Mean scores of the four respondent subgroups

Numbers of observations, mean actual scores of all 86 health states in Subgroup 1-4 are estimated and shown in Table 4.3. Mean scores were ranked from highest to lowest scores according the scores from Subgroup 1. In Subgroup 1, numbers of observations are ranged from 92 to 1,313, 83 to 1,214 in Subgroup 2, 80 to 1,074 in Subgroup 3 and 45 to 614 in Subgroup 4. There is no observation from state 13123 in Subgroup 2.

Note that the respondents in subgroup 4 are the same as those in Group I. Health states with significantly different mean scores (95% CIs not overlapped) are shown with an asterisk. Out of eighty-six states, mean score of only one health state (33333) from Subgroup 2 (with fewer than sixteen inconsistencies) significantly differs from that of Subgroup 1. Four states in Subgroup 3 (with fewer than eleven inconsistencies) significantly differ from those of Subgroup 1 and twenty-five states in Subgroup 4 significantly differ from those of Subgroup 1. The highest mean score for state 11112 and the lowest score for state 33333 were seen from the highly consistent respondents (subgroup 4). The number of states with negative scores is smallest in subgroup 1.

Table 4.3 Mean scores of health states after excluding scores from the inconsistent respondents

State	Mean TTO scores from the respondents in							
	Subgroup 1		Subgroup 2		Subgroup 3		Subgroup 4	
	n	Mean	n	Mean	n	Mean	n	Mean
11112	314	0.705	294	0.720	251	0.756*	133	0.815*
11121	237	0.684	210	0.717	191	0.735	121	0.766*
11122	109	0.674	105	0.674	91	0.705	47	0.765
21111	232	0.667	212	0.696	182	0.706	116	0.742*
11211	312	0.667	288	0.679	255	0.693	144	0.742*
12111	222	0.645	198	0.667	180	0.691	97	0.768*
11221	111	0.641	106	0.653	103	0.665	74	0.684
21112	96	0.628	83	0.665	80	0.663	49	0.736*
21211	111	0.604	106	0.644	93	0.672	54	0.692
12112	211	0.602	201	0.613	172	0.640	94	0.680*
21121	209	0.594	193	0.598	163	0.667*	87	0.725*
11232	120	0.583	108	0.594	91	0.619	62	0.659
12211	118	0.582	109	0.595	91	0.627	62	0.721*
21122	108	0.572	104	0.578	90	0.614	47	0.685*
11212	118	0.570	103	0.610	91	0.649	50	0.719*
21212	113	0.570	108	0.584	105	0.603	75	0.603
22111	112	0.518	102	0.555	86	0.590	46	0.614
12221	209	0.515	192	0.518	162	0.554	87	0.608*
22211	120	0.492	109	0.505	91	0.519	62	0.633*
12212	113	0.483	103	0.503	87	0.559	46	0.631*
12121	106	0.478	95	0.487	89	0.495	47	0.625*
12122	119	0.478	101	0.505	89	0.528	50	0.534
11222	112	0.476	106	0.508	93	0.578	54	0.614*
22112	216	0.472	204	0.471	174	0.536	95	0.573*
22121	99	0.469	85	0.496	81	0.485	50	0.627*
21312	119	0.455	103	0.468	91	0.494	49	0.554
11223	114	0.428	108	0.436	105	0.445	75	0.452
21221	106	0.421	94	0.491	88	0.507	46	0.584*
12312	98	0.397	89	0.388	76	0.418	39	0.543
12123	111	0.391	105	0.398	92	0.448	53	0.450
22221	112	0.385	101	0.418	85	0.462	46	0.485
22113	97	0.384	85	0.383	80	0.388	50	0.439
11313	115	0.360	103	0.374	86	0.406	47	0.420
11332	95	0.354	84	0.334	80	0.319	50	0.291
21123	104	0.340	94	0.321	88	0.324	46	0.376
23222	118	0.327	108	0.312	91	0.297	62	0.288
12313	120	0.280	108	0.275	91	0.270	62	0.297
21231	104	0.278	93	0.260	82	0.259	47	0.268
13123	92	0.277	NA	NA	92	0.277	48	0.347
22313	95	0.260	85	0.232	80	0.231	49	0.285
21313	115	0.253	104	0.242	87	0.256	48	0.269
21332	213	0.250	195	0.236	167	0.220	86	0.256

Table 4.3 Mean scores of health states after excluding scores from the inconsistent respondents (cont.)

State	Mean TTO scores from the respondents in							
	Subgroup 1		Subgroup 2		Subgroup 3		Subgroup 4	
	n	Mean	n	Mean	n	Mean	n	Mean
12331	107	0.247	103	0.250	90	0.281	47	0.238
13222	112	0.196	106	0.212	93	0.282	54	0.329
21331	116	0.175	100	0.206	88	0.190	49	0.131
22323	209	0.167	193	0.146	162	0.173	87	0.193
31311	108	0.160	104	0.161	90	0.184	47	0.314
23113	112	0.154	108	0.147	104	0.160	75	0.200
22232	114	0.131	103	0.102	86	0.091	48	0.139
23321	104	0.126	94	0.115	81	0.093	46	0.222
13232	119	0.106	104	0.089	91	0.056	50	0.035
23223	214	0.078	203	0.057	175	0.044	96	0.103
22333	98	0.056	85	-0.020	81	-0.051	50	-0.078
23131	100	0.050	93	0.030	86	0.023	45	0.050
23322	112	0.025	107	0.019	104	0.013	75	0.015
23232	99	0.020	86	0.005	81	-0.045	50	-0.053
23323	113	0.019	108	0.007	92	-0.008	48	-0.052
33122	108	0.002	104	-0.024	90	-0.039	47	0.102
31222	111	0.000	101	-0.015	86	0.001	46	0.106
22233	119	-0.003	107	-0.011	90	-0.055	62	-0.068
23231	112	-0.008	103	-0.041	87	-0.040	46	-0.002
23132	101	-0.009	93	-0.049	87	-0.041	46	-0.046
31213	98	-0.013	89	-0.048	76	-0.057	39	-0.032
22332	102	-0.017	93	-0.017	88	-0.021	46	-0.011
31131	112	-0.025	107	-0.030	90	-0.070	47	0.052
33222	109	-0.028	105	-0.056	91	-0.063	47	-0.028
32123	113	-0.085	108	-0.082	104	-0.093	75	-0.100
33223	113	-0.117	105	-0.157	91	-0.184	47	-0.129
23333	318	-0.119	297	-0.156	254	-0.188	134	-0.225*
32322	117	-0.124	103	-0.163	91	-0.214	50	-0.252
23332	101	-0.129	92	-0.164	88	-0.167	46	-0.268
33121	104	-0.131	95	-0.156	82	-0.199	47	-0.168
23233	114	-0.134	103	-0.181	87	-0.211	46	-0.149
32232	110	-0.134	101	-0.164	85	-0.176	45	-0.151
32332	120	-0.155	104	-0.221	91	-0.277	51	-0.429*
33221	110	-0.178	105	-0.195	93	-0.185	54	-0.109
32323	210	-0.192	194	-0.247	163	-0.257	87	-0.275
32223	109	-0.213	104	-0.213	91	-0.205	54	-0.179
32233	121	-0.215	108	-0.268	91	-0.328	62	-0.371*
33322	111	-0.233	106	-0.253	104	-0.253	75	-0.290
33233	314	-0.251	291	-0.290	257	-0.330*	147	-0.368*
33323	221	-0.268	199	-0.296	181	-0.307	98	-0.405*
32333	233	-0.282	211	-0.329	192	-0.344	122	-0.338
33232	109	-0.303	104	-0.319	92	-0.322	54	-0.315
33332	226	-0.318	215	-0.344	184	-0.379	116	-0.389
33333	1313	-0.346	1214	-0.386 *	1074	-0.419 *	614	-0.480 *

*significant difference from subgroup 1 (95% CIs not overlapped)

Subgr1=all respondents, Subgr2=those with fewer than 16 incons, Subgr3=those with fewer than 11 incons, Subgr4=those with fewer than 6 incons.

4.3.4 Spearman rank correlation coefficients

As stated previously that Subgroup 1 is regarded as a reference subgroup. To see whether the rank of mean scores of the health states are changed after excluding the scores from the respondents with various numbers of inconsistently values, Spearman rank correlations are used. The scores from all subgroups are ranked from the highest to the lowest scores. The coefficients are examined to see correlation between the rank of scores from Subgroup 2 and Subgroup 1, Subgroup 3 and Subgroup 1 and Subgroup 4 and Subgroup 1. The results are presented in Table 4.4.

Table 4.4 Spearman rank correlation coefficients between mean scores of the four subgroups

Correlation between the ranks	**Spearman rank corre coeff.**	**95% confidence interval**	
		lower limit	upper limit
Subgr.1 - Subgr.2	0.997	0.995	0.998
Subgr.1-Subgr.3	0.992	0.988	0.995
Subgr.1-Subgr.4	0.985	0.978	0.990

After excluding the respondents with various numbers of inconsistent responses, the ranks of the three subgroups are highly correlated with the rank of mean scores from subgroup 1. The correlation coefficients between the ranks of mean scores from Subgroup 1 and 4 and Subgroup 1 and 3 are significantly lower than that between Subgroup 1 and 2 at $p<0.05$ (95% CIs are not overlapping). Whereas the correlation coefficients between the ranks of mean scores from Subgroup 1 and 3 and from Subgroup 1 and 4 are not significantly different from each other.

4.4 Discussion

This chapter explores the extent of logical inconsistency in the TTO scores and the mean scores for health states from respondents displaying different levels of inconsistency in their responses. The exclusion of responses from 'inconsistent' respondents has significant effects on the mean scores of some health states. Highly inconsistent respondents tend to give higher scores for poorer states and lower scores for better states (compared to the more consistent respondents).

Excluding the respondents with inconsistent scores changed the mean scores of health states. By excluding the scores from the respondents with more than five inconsistencies (subgroup 4), almost one-fourth of the health states are significantly different from the scores estimated from all respondents. There are four respondent subgroups and one of this subgroup is expected to be used in the model specifications. By learning that excluding the scores from the inconsistent respondents has significant effects of the mean actual scores, the question would then be which subgroup is an appropriate choice. It is seen that the respondents with fewer than six inconsistencies can be assumed to be those who give “valid” or “high quality” scores to represent their preferences on health states. However, if the scores from all respondents are chosen to be used to estimate that Thai scores, at least 25% of the health states have “invalid” or “low quality” scores. Therefore, the scores from all respondents (Subgroup 1) are not favourable.

Based on the score validity, the scores from subgroup 4 respondents should be selected because the scores are the most “valid” scores. The respondents in this subgroup tend to understand the task and likely to assign the scores according to their preferences. However, if this subgroup were to be selected, almost fifty per cent of the respondents would be excluded and the scores from only 632 respondents would be used to estimate the Thai tariff. This number is not small compared with the other studies, except for the UK and the US studies, since 621 respondents participated in the Japanese study, 339 in the German study, 370 in the Slovenian study, 309 in the Dutch study and 488 in the South Korean study. However, the exclusion of fifty per cent of the sample is unacceptable in this study because it is certainly not making the best use of the available data and would result in a substantial number of valid responses being discarded without a robust justification.

There are then two subgroups available for the selection: subgroup 2 (fewer than sixteen inconsistencies) and subgroup 3 (fewer than eleven inconsistencies). Only one state (33333) in subgroup 2 was significantly different from the corresponding health state in subgroup 1 whereas, in subgroup 3, four states are different from subgroup 1. This implies that after excluding the respondents with more than fifteen inconsistencies (subgroup 2), the remaining scores are not much different from those in subgroup 1. In the comparison between the mean scores of subgroup 1 and subgroup 3, more than one state is significantly differed. This makes subgroup 3 more favourable than subgroup 2. Moreover, only twenty per cent of the respondents are excluded to form this subgroup. This number is the minimum number that could be offered on the basis of the availability of four subgroups generated in this study. By this ground, the respondents with fewer than eleven inconsistencies (subgroup 3) are chosen to estimate the Thai scores.

4.5 Conclusion

Excluding the scores from the inconsistent respondents changes the mean scores of the health states. When selecting the data set from which to estimate the Thai preference scores, the data should be a valid representation of health preferences and any exclusion of respondents should be kept to a minimum. Demographic characteristic, mean scores from the respondents with various numbers of inconsistencies and mean scores of four respondent subgroups are compared. The scores from the respondents with fewer than eleven inconsistencies are chosen for the estimation of the Thai tariff in an attempt to balance the twin concerns of excluding as few respondents as possible while maintaining a degree of confidence in the validity of the data. The high correlation between the ranks of the scores assigned by all respondents and the more consistent respondents could be used to justify the exclusion of the scores from highly inconsistent respondents. It is likely that the highly consistent respondents may have fewer inconsistencies at the second half of the task, presumably because of the learning effect. This issue needs to be further explored.

Chapter 5 Model analysis

5.1 Introduction

The previous chapter has shown that inclusion of responses from the inconsistent respondents systematically changes the mean scores given to different health states. The scores from the respondents with fewer than eleven inconsistencies are chosen to be used in the model analysis according to the reasons stated in the previous chapter. This chapter reports the results of model analyses using the chosen subgroup to estimate the Thai scores. To see the differences if the scores from other subgroups would have been used to estimate the Thai scores, the scores from the other three subgroups are also used to estimated the models.

The outline of the chapter is as follows. Firstly, the analysis plan is described including the criteria used to choose the “best” model to estimate the Thai scores, the models and the variables. The results of the analysis using the scores from the respondents with fewer than eleven inconsistencies are presented. Performances of the scores are compared and the “best” model is chosen. Impacts of choice of subgroups on the “best model” are examined using the scores from the respondents of the other three subgroups. The Thai algorithm for determining scores for EQ5D health states is presented at the end of the chapter as well as the comparison of the Thai preference scores compared with those estimated from the UK and the Japanese models.

5.2 Analysis plan

From the outset it was decided that an algorithm for valuing EQ-5D health states would be developed using existing models. The Dolan (1997), Dolan & Roberts (2002) and Shaw e*t al*. (2005) models are explored in this study (43, 44, 51). The Dolan (1997) model is selected because it was the first model used to estimate preference scores for EQ-5D health states for the UK and the model has been used as a reference model in the estimation of preference scores for many countries. The Dolan & Roberts (2002) model is chosen because the model offers an alternative way of estimating preference scores and the model’s performance (in UK data) was better than that of Dolan (1997). The Shaw e*t al* (2005) model is also chosen to model the Thai scores because in an analysis of US data it performed better than the Dolan (1997) model.

5.2.1 Criteria to select the best model

As a consequence of estimating a number of different models a means of identifying which model is to be preferred must be found. The "best model" is chosen based on four criteria: logical inconsistency, model robustness, parsimony and the responsiveness of scores to changes in health. To conform with " a utility maximization model" that "If a specific health care program improves the health of some persons, they will move to a higher level of health sooner than they would have otherwise, and the amounts of his health improvement can be readily calculated in terms of index days (health days)" (64). This statement implies that a higher level of health has a "higher score" and the amount of the difference between the higher and lower levels of health indicates how much "better–off" a person is after receiving health care. This amount is used to determine whether the effect of a health care program justifies its costs. Therefore, first priority is given to the models estimating higher scores for better health states, that is, that produce logically consistent scores. The remaining models are then considered following the next criterion.

The second most important criterion is the robustness of the model. This will be assessed by randomly assigning two-thirds of respondents to a modelling sample, and the remaining one-third to a validation sample. The coefficients estimated from the modelling sample are used to predict scores in the validation sample and these predicted scores are compared with the mean actual scores for the corresponding health states in the validation sample. Small R-squared and large root mean squared error (RMSE) and mean absolute differences (MAD) indicate poor model performance. An additional method to reassure model robustness is the number of states with the absolute difference between the predicted and the actual scores larger than 0.1 (43, 65). The better performing model is expected to estimate scores closer to the actual scores.

The third criterion is parsimony. The simplest model or the model with the smallest number of independent variables is preferred. The fourth criterion is that the scores are sensitive to changes in health states. Cohen effect size is used to compare the responsiveness of scores across different models. Also, other things equal, the model estimating the highest score for the best ill health state and lowest score for the worst state is favoured.

5.2.2 Statistical analysis

Utility scores are assumed to depend on the levels of the five dimensions of the EQ-5D health state. Initially the responses from individuals are assumed not to be correlated and all observations are pooled and analysed using the Ordinary Least Square model (OLS). The explanatory variables are the dummy variables indicating whether a particular dimension was at level 2 (some problem), or at level 3 (severe problems).

A general model is: $\boldsymbol{y_{ij}} = \propto + \boldsymbol{x'_{it}\beta} + \epsilon_{ij}$

Where: y_{ij} = a score for state *j* observed from the respondent *i* (*i*=1,2,3,…n and *j* = 1,2,3,…10). y_{ij} is a continuous variable, $\boldsymbol{x'_{it}}$ = the explanatory variables, $\boldsymbol{\beta}$ = a vector of coefficients, ϵ_{ij}= an error term

And:

$$E[\epsilon_{ij} \,|x_{i1}, x_{i2}, x_{i3}, \dots, x_{i10}] = 0$$

$$Var[\epsilon_{ij} \,|x_{i1}, x_{i2}, x_{i3}, \dots, x_{i10}] = \sigma_{\epsilon}^{2}$$

$$Cov[\epsilon_{ij}, \epsilon_{ts} \,|x_{i1}, x_{i2}, x_{i3}, \dots, x_{i10}] = 0 \text{ if } i \neq t \text{ or } j \neq s$$

However, the scores from one respondent are likely to be correlated. Biases could arise using the OLS model. The error terms are heteroskedastic. To take the correlation of scores into account, the data are treated as panel data. Given that the number of scores assigned by respondents is unequal, the panel is unbalanced. The time-invariant factors (e.g. age, gender and race) are reported to have effects on utility scores for health states (66).

A model for panel data is:

$$y_{ij} = x'_{it}\beta + z'_{i}\alpha + \epsilon_{ij}$$

where: y_{ij}= a score for state *j* observed from the respondent *i* (*i*=1,2,3,…n and *j* = 1,2,3,…10). y_{ij} is a continuous variable, $\boldsymbol{x'_{it}}$ = the explanatory variables, $\boldsymbol{z'_{i}}$ = the individual effects on the scores, ϵ_{ij}= an error term (67)

The first model to be used is the Fixed-effects model (FE) where it is assumed that the error terms are correlated with the explanatory variables. The second model is the Random-effects model (RE) where the error terms are assumed to be uncorrelated with the explanatory variables. The Ramsey RESET test is used to test for misspecification and the Hausman test is used to verify the appropriateness of using FE or RE models.

The Breusch-Pagan test is used to test the appropriateness between the OLS and the RE model.

The FE model is estimated by the following equation:

$$y_{ij} = x'_{it}\beta + c_i + \epsilon_{ij}$$

Where c_i is the component of time-variant and time-invariant factors. Utility scores may be affected by age, gender, residential area, interviewer effect, religious belief and the personal beliefs on health. Although some of these factors are observed (age, gender, residential area, interviewer effect), some are not.

$$E[c_i|\mathbf{X}_i] = \mathbf{h}(\mathbf{X}_i)$$

In the FE model, we assume that individuals give scores with the same slope (β) but different intercept α_i where *i* =1-N. The time-variant and time-invariant individual factors are absorbed into ($\propto$). Therefore, the formula can be written as:

$$y_{ij} = x'_j\beta + \alpha_i + \epsilon_{ij}$$

A set of dummy variables (**D**) is established to identify the respondent *i*. **D**=$[d_1, d_2, \ldots, d_n]$

$$y = X\beta + D\alpha + \epsilon$$

This model can be treated as the Least Square model, the estimator β is:

$$[\ \sum_{i=1}^{N}\sum_{j=1}^{N}(x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)']^{-1}\,[\sum_{i=1}^{N}\sum_{j=1}^{N}(x_i - \bar{x})\,(y_i - \bar{y}_i)]$$

Where:

$\bar{y}_i = \frac{1}{N}\sum_{j=1}^{N} y_{ij}$, $\bar{x}_i = \frac{1}{N}\sum_{j=1}^{N} x_{ij}$ (67, 68)

The RE model is estimated by the following equation:

$y_{ij} = x'_{ij}\beta + (\propto + u_i) + \epsilon_{ij}$ (12)

Where u_i= the random heterogeneity of the individual respondents that is constant through time and

$$E[\epsilon_{ij}|\mathbf{X}] = \mathrm{E}[u_i|\mathbf{X}] = 0$$

$$E[\epsilon_{ij}^2|\mathbf{X}] = \sigma_\epsilon^2$$

$$E[u_i^2|\mathbf{X}] = \sigma_u^2$$

$$E[\epsilon_{ij}\,u_t|\mathbf{X}] = 0 \text{ for all i, and j and t}$$

$$E[\epsilon_{ij}\epsilon_{ts}|\mathbf{X}] = 0 \text{ if } j \neq s \text{ or } i \neq t$$

$$E[u_i u_t|\mathbf{X}] = 0 \text{ if } \; i \neq t$$

5.2.3 The variables

The variables specified in Dolan (1997), Dolan and Roberts (2002) and Shaw *et al.* (2005) used in this analysis are as follows (43, 44, 51). Eleven dummy variables are included in the Dolan model. Two dummy variables are generated for each dimension. The first variable takes value one for level 2, two for level 3 and zero otherwise. The second variable takes value one for level 3, zero otherwise. The final variable takes the value one if any dimension is at level 3, zero otherwise. The dependent variable is the difference between perfect health (1) and the score estimated from the model. Details of the definitions of the variables are presented in Table 5.1.

Table 5.1 Variables and definitions of Dolan 1997 model

variable	definition
cons.	constant
mo	1 if mobility is at level 2, 2 at level 3, 0 otherwise
sc	1 if self-care is at level 2, 2 at level 3, 0 otherwise
ua	1 if usual activities is at level 2, 2 at level 3, 0 otherwise
pd	1 if pain/discomfort is at level 2, 2 at level 3, 0 otherwise
ad	1 if anxiety/depression is at level 2, 2 at level 3, 0 otherwise
m2	1 if mobility is at level 3, 0 otherwise
s2	1 if self-care is at level 3, 0 otherwise
u2	1 if usual activities is at level 3, 0 otherwise
p2	1 if pain/discomfort is at level 3, 0 otherwise
a2	1 if anxiety/depression is at level 3, 0 otherwise
N3	1 if any dimension is at level 3, 0 otherwise

(43)

The Dolan & Roberts model also includes 11 dummy variables. Two dummy variables are generated for each dimension. The first variable takes the value one if the difference between state 33333 and the corresponding dimension at the state of interest is one. The second variable takes value one if the difference is two. The final variable (*ANY13*) takes the value one if at least one dimension is at level 1 and at least one dimension is at level 3. The dependent variable is the sum of the mean actual score for state 33333 and the scores estimated from the model. Details of the definitions of the variables are presented in Table 5.2.

In Shaw *et al*. model, two types of variables: dummy variables and ordinal variables are generated. Two dummy variables are created for each of the five dimensions. The first variable takes the value one for level 2, zero otherwise. The second set takes the value one for level 3, zero otherwise. Five ordinal variables are created: d_1 is the number of dimensions moving away from level 1, minus one; i_2 is the number of dimensions at level 2, minus one; i_3 is the number of dimensions at level 3, minus one; $i_2 - squared$ is the square of i_2; and $i_3 - squared$ is the square of i_3. If d_1, i_2 or i_3 are negative they are set equal to zero. Details of the definitions of the variables are presented in Table 5.3.

Table 5.2 Variables and definitions of Dolan & Roberts 2002 model

variable	definition
cons	constant
difmob1	1 if the difference in mobility is 1, 0 otherwise
difsc1	1 if the difference in self-care is 1, 0 otherwise
difua1	1 if the difference in usual activities is 1, 0 otherwise
difpd1	1 if the difference in pain/discomfort is 1, 0 otherwise
difad1	1 if the difference in anxiety/depression is 1, 0 otherwise
difmob2	1 if the difference in mobility is 2, 0 otherwise
difsc2	1 if the difference in self-care is 2, 0 otherwise
difua2	1 if the difference in usual activities is 2, 0 otherwise
difpd2	1 if the difference in pain/discomfort 2, 0 otherwise
difad2	1 if the difference in anxiety/depression 2, 0 otherwise
ANY13	1 if the difference includes 0 and 2, 0 otherwise

(65)

Table 5.3 Variables and definitions of Shaw *et al.* 2005 model

variable	definition
m1	1 if mobility is at level 2, 0 otherwise
s1	1 if self-care is at level 2, 0 otherwise
u1	1 if usual activities is at level 2, 0 otherwise
p1	1 if pain/discomfort is at level 2, 0 otherwise
a1	1 if anxiety/depression is at level 2, 0 otherwise
m2	1 if mobility is at level 3, 0 otherwise
s2	1 if self-care is at level 3, 0 otherwise
u2	1 if usual activities is at level 3, 0 otherwise
p2	1 if pain/discomfort is at level 3, 0 otherwise
a2	1 if anxiety/depression is at level 3, 0 otherwise
d1	the number of dimensions moving away from level 1, minus 1
	d1=0 for state 11111
i2	the number of dimensions at level 2, minus 1
	if no level 2 in any dimension, i2=0
i22	square of i2
i3	the number of dimensions at level 3, minus 1
	if no level 3 in any dimension, i3=0
i32	square of i3

(51)

5.2.4 Predictive ability and responsiveness

After the models are estimated, the resulting models and the estimated scores are examined to choose the "best model" using the criteria stated in the beginning of the chapter. The following sections report the methods of the estimation of predictive abilities of the models, responsiveness and logical inconsistency of the estimated scores. The following are the formulas used to calculate the predictive ability of the model and the responsiveness of the models: root mean square errors (RMSE); and mean absolute difference (MAD).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^{2}}$$

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|x_i - y_i|$$

Where: n = number of health states, x_i = the score estimated from the model, i =1, 2…n, y_i = the actual score, i =1, 2…n (67)

To calculate the responsiveness of the model, the estimated scores for all 243 states are used. All possible pairs from 243 states are generated where the first state of each pair represents the baseline state and the second state is the post-treatment state. States in the pairs are arranged under the assumption that only positive transformations are generated. The responsiveness is measured using the Cohen effect size and the formula is as follows.

$$\textbf{Cohen effect size} = \frac{mean_p - mean_b}{SD_b}$$

Where: $mean_p$ = mean of post-treatment states, $mean_b$ = mean of baseline states, SD_b = Standard deviation of baseline mean (69)

5.2.5 Logical inconsistency in the estimated scores

Two methods are used to identify logically inconsistent responses. The first method is to use Stata to detect inconsistent estimated scores. The inconsistent scores are then re-examined by finding the cause of inconsistency using the coefficients in the models. By using Stata program, out of 243 states, a total of 7,625 pairs can be used to identify the logically inconsistent responses. To identify the logical inconsistency from the estimated scores the method adopted is as follows. All 243 states are ranked according to the EQ-5D numeric codes from perfect health (11111) to the worst health state (33333). Note that this will tend to rank health states roughly in terms of increasing severity. To illustrate how health state pairs are formed, see Figure 5.1. The first health states of the pairs are lined from state 1 to state 243 in a horizontal plane. The second states of the pairs from state 1 to state 243 are in a vertical plane. The illustrated diagonal loop involves the comparison of state 2 with state 1, state 3 with state 2 and so on. The second loop (not shown) represents the comparison between state 3 and state 1, state 4 and state 2 and so on. By this method, a total of 242 loops can be constructed which would cover all potential pairs. Only the pairs that can be used to identify logical inconsistencies are taken into account.

Note that only the pairs of health states up until the 45th loop had been examined. The searching for the inconsistent scores does not proceed beyond the 45th loop because, if the inconsistent scores are found from the 1st to 45th loop, it is worthless to explore further on until the 242nd loop is achieved. Moreover, the model is likely to predict the inconsistent scores for the similar health states, for example, 11311 and 11312 which are paired in the first loop. Some similar health states may be found in another loop, for example, 13112 and 13212 which are presented in the 9th loop (state

13212 is ranked at the 65th state and state 13112 is at the 56th state). It is unlikely to identify the inconsistent scores at the further loop without identifying some of the inconsistent scores prior to that. Therefore, by covering all the pairs up until 45th loop, it is likely that all the fairly similar states are covered.

5.3 Results

The results are divided into two sections. In the first section results are presented for analyses based on the preferred subgroup (Subgroup 3). The Thai data are explored using the Dolan (1997), Dolan & Roberts (2002) and Shaw *et al*. (2005) models and the preferred model is identified. This model is further explored by examining performance with additional variables, and by reviewing health states for which it predicts poorly. Finally the Thai algorithm based on the full sample (modelling and validation samples combined) is reported. In the second section the impact of selecting Subgroup 3 is fully explored. The three models are estimated for each of the four subgroups. The models are then compared in terms of: score assigned to the best and worst ill health states; logical inconsistency; number of health states with negative scores and Cohen effect size.

Figure 5.1 Identification of logical inconsistency from the estimated 243 states

State →

State ↓	1	2	3	4	5	6	7	8	9	10	11	12	13	14	……… . . .	237	238	239	240	241	242	243
1	X	21	31	41	51	61	71	81	91	101	111	121	131	141	... . . .	2371	2381	2391	2401	2411	2421	2431
2		X	32	42	52	62	72	82	92	102	112	122	132	142	... . . .	2372	2382	2392	2402	2412	2422	2432
3			X	43	53	63	73	83	93	103	113	123	133	143	... . . .	2373	2383	2393	2403	2413	2423	2433
4				X	54	64	74	84	94	104	114	124	134	144	... . . .	2374	2384	2394	2404	2414	2424	2434
5					X	65	75	85	95	105	115	125	135	145	... . . .	2375	2385	2395	2405	2415	2425	2435
6						X	76	86	96	106	116	126	136	146	... . . .	2376	2386	2396	2406	2416	2426	2436
7							X	87	97	107	117	127	137	147	... . . .	2377	2387	2397	2407	2417	2427	2437
8								X	98	108	118	128	138	148	... . . .	2378	2388	2398	2408	2418	2428	2438
9									X	109	119	129	139	149	... . . .	2379	2389	2399	2409	2419	2429	2439
10										X	1110	1210	1310	1410	... . . .	23710	23810	23910	24010	24110	24210	24310
11											X	1211	1311	1411	... . . .	23711	23811	23911	24011	24111	24211	24311
12												X	1312	1412	... . . .	23712	23812	23912	24012	24112	24212	24312
13													X	1413	... . . .	23713	23813	23913	24013	24113	24313	24313
																…..	…..	…..	…..	…..	…..	…..
																…..	…..	…..	…..	…..	…..	…..
																…..	…..	…..	…..	…..	…..	…..
																…..	…..	…..	…..	…..	…..	…..
																…..	…..	…..	…..	…..	…..	…..
																X	…..	…..	…..	…..	…..	…..
																	X	239238		…..	…..	…..
239																		X	240239	241239	242239	243239
240																			X	241240	242240	243240
241																				X	242241	243241
242																					X	243242
243																						X

There are 1,074 respondents who assigned TTO scores with fewer than eleven inconsistencies. Two-thirds or 7,137 observations are randomly assigned to be a modelling sample (internal sample). The remaining one-third (3,570 responses) are used as a validation sample (external sample).

5.3.1 Dolan (1997) model

Firstly, OLS model was used to estimate the model. This model failed the Breusch - Pagan test; the null hypothesis that the model's variances are constant was rejected at the p-value of 0.000. FE and RE models were then applied. Using FE model, the F-test of the null hypothesis that the error terms are zero was rejected at the p-value of 0.000. This was to confirm that FE model was favour compared with OLS model. RE model was also applied to estimate the model. The Hausman test is used to compare the appropriateness between FE and RE model. The null hypothesis that the difference in coefficients is not systematic cannot be rejected at the p-value of 0.573, thus RE model is more efficient than FE model. Therefore, an RE model is used to estimate the model.

From RE models, the estimated coefficients for the variables s2 and u2 were not significant (at the p-value of 0.05). The non-significant variables were dropped and the models were reanalysed. The Breusch-Pagan test is used to test the heterogeneity of variances.

5.3.2 Dolan & Roberts (2002) model

Initially an OLS model was estimated. The Breusch-Pagan test revealed that the null hypothesis that the model's variances are constant was rejected at the p-value of 0.000. FE and RE model are then applied. Using FE model, the F-test of the null hypothesis that the error terms are zero was rejected at the p-value of 0.000. This was to confirm that FE model was favour compared with OLS model. RE model was also applied to estimate the model. To compare the appropriateness between FE and RE model, the Hausman test is used. The null hypothesis that the difference in coefficients is not systematic cannot be rejected at the p-value of 0.663, thus RE model is more efficient than FE model. Therefore, an RE model is used to estimate the model.

From RE model, the coefficients of all variables are statistically significant with positive signs except the variable *ANY13* and a constant term. The Breusch-Pagan test is used to test the heterogeneity of variances.

5.3.3 Shaw *et al.* (2005) model

Unlike the models estimated earlier, the models using the variables from the Shaw *et al.* (2005) study, were estimated with no constant term. Note that the TTO scores for states worse than death in this model were transformed using the formula:

$$U' = \frac{U}{39}$$

Where *U'* = the transformed TTO scores for states worse than death, *U* = the untransformed scores from the raw data where the scores for state worse than death.

The first model was estimated using OLS. The null hypothesis of equal variance was rejected at the p-level of 0.000. Therefore, OLS model was not appropriate to fit the data. Next, Feasible generalized least square (FGLS) which is one type of RE model was used. The estimation is used where the variance components are unknown (67). The insignificant coefficients were dropped and the models were re-analysed.

The results of parameter estimates using the variables from the three models are presented in Table 5.4. Table 5.4 shows the estimated coefficients from the model specifications using the Dolan (1997), Dolan & Roberts (2002) and Shaw *et al.* (2005) models. Variables, coefficients and standard errors (SE) of the Dolan (1997) model are presented in the first three columns. The 4^{th}-6^{th} columns represent variables, coefficients and SEs of the Dolan & Roberts (2002) model and the 7^{th}-9^{th} columns represent those of the Shaw *et al.*(2005) model. Only significant variables (p-value<0.05) are presented in this table. Ten variables are significant in the Dolan (1997) model, twelve in the Dolan & Roberts (2002) model and fourteen in the Shaw *et al.* (2005) model. Variables *s2* (-0.019) and *u2* (-0.016) were dropped from the Dolan (1997) model. Mean score of state 33333 used in the Dolan & Roberts 2002 model is -0.419. This value is used to generate the dependent variable of the Dolan & Roberts (2002) model. R-squared is similar in the first two models. The Breusch-Pagan test is conducted after in the Dolan (1997) and Dolan & Roberts (2002) models. The test demonstrates that the Null hypothesis of no heterogeneity of variances is rejected (p-value is 0.000). All models suffer from heteroskedasticity.

Table 5.4 Parameter estimates and fit statistics of the three alternative model specifications

Dolan 1997 model			**Dolan&Roberts 2002 model**			**Shaw et al.2005 model**		
Variables	**Coeff.**	**SE**	**Variables**	**Coeff.**	**SE**	**Variables**	**Coeff.**	**SE**
mo	0.120	0.012	difmob1	0.310	0.012	m1	0.289	0.008
sc	0.120	0.007	difmob2	0.457	0.014	m2	0.615	0.015
ua	0.060	0.007	difsc1	0.123	0.012	s1	0.305	0.009
pd	0.074	0.012	difsc2	0.271	0.013	s2	0.525	0.017
ad	0.038	0.012	difua1	0.066	0.012	u1	0.261	0.009
m2	0.177	0.018	difua2	0.161	0.013	u2	0.423	0.015
p2	0.080	0.019	difpd1	0.169	0.011	p1	0.273	0.009
a2	0.043	0.019	difpd2	0.268	0.014	p2	0.510	0.016
N3	0.138	0.016	difad1	0.099	0.011	a1	0.236	0.009
_cons	0.200	0.015	difad2	0.167	0.013	a2	0.427	0.016
			ANY13	-0.073	0.010	i2	0.069	0.016
			_cons	-0.055	0.013	i22	-0.010	0.004
						i32	-0.027	0.001
			Mean score of state 33333	-0.419		d1	-0.256	0.013
R2(overall)	0.448			0.447			NA	
RMSE	0.102			0.106			0.257	
MAE	0.080			0.085			0.199	
Number of states with absolute difference > 0.1	28			30			59	
Number of logical inconsistencies in the estimated 243 states	0			15			37	
Cohen effect size	1.084			1.083			1.023	
scores for state								
11112	0.766			0.782			0.764	
33333	-0.452			-0.469			-0.074 (state 33232)	

Note: RMSE=Root mean squared error, MSE=Mean squared error

Out of eighty-six states, twenty-eight states have the absolute difference between the estimated and mean scores larger than 0.1 estimated from the Dolan (1997) model, thirty states from the Dolan & Roberts (2002) model and fifty-nine states from the Shaw *et al.* (2005) model. Thirty-seven logical inconsistent responses are identified in the scores estimated from the Shaw *et al.* (2005) model and fifteen inconsistencies from the Dolan & Roberts 2002 model. The Dolan (1997) model is only one model that estimates the completely consistent scores.

Cohen effect size is similar between the first two models. The effect size is lowest in the Shaw *et al.* (2005) model. The Dolan & Roberts (2002) model estimates the highest score for the best ill health state (11112) and the lowest score for the worst state (33333). Note that the Shaw *et al.* (2005) model estimates the lowest score for state 33232 rather than for state 33333.

The Dolan 1997 model is the only model estimating the scores with no logical inconsistency. This makes the Dolan 1997 model to be the best model to estimate the Thai scores. There also are other aspects that make the Dolan 1997 more favourable. Compared with the other models, this model is the simplest model, in terms of the number of variables. R-square of the model is slightly higher than that of the Dolan & Roberts (2002) model and RMSE and MAD are the lowest among the three models. The number of states with the absolute difference between the actual and estimated scores exceeding 0.1 is smallest in the scores estimated from the Dolan (1997) model.

The responsiveness of the scores estimated by the Dolan 1997 model is similar to that estimated by the Dolan & Roberts 2002 model. Although the Cohen effect sizes of the two models are similar, the Dolan & Roberts 2002 model estimates higher scores for state 11112 and lower score for state 33333. The Dolan 1997 model would have been less favourable compared with the Dolan & Roberts 2002 model if the selection was based on only this criterion. However, it is because the latter model estimates the scores with logical inconsistencies which make this model less favourable.

5.4 Adding variables to the Thai model

Because the Thai model still suffers from heteroskedasticity, more variables are incorporated to see whether the model would perform better. Two sets of variables are added in the Thai model: all two-way interaction terms; and the variable *x4* from the US Hispanic model are used in the model specifications (54). The modelling sample of subgroup 3 respondents is used in the model estimation. The models are estimated using a random effects model. The models are again assessed in terms of the logical consistency of the predicted scores, robustness and the best-worst predicted scores. Definitions of the variables are presented in Table 5.5.

Table 5.5 Definitions of the interaction terms

Interaction terms (apart from N3)	
variable	**definition**
mo_sc	The product of the interactions between *mo* and *sc*
mo_ua	The product of the interactions between *mo* and *ua*
mo_pd	The product of the interactions between *mo* and *pd*
mo_ad	The product of the interactions between *mo* and *ad*
mo_s2	The product of the interactions between *mo* and *s2*
mo_u2	The product of the interactions between *mo* and *u2*
mo_p2	The product of the interactions between *mo* and *p2*
mo_a2	The product of the interactions between *mo* and *a2*
sc_ua	The product of the interactions between *sc* and *ad*
sc_pd	The product of the interactions between *sc* and *pd*
sc_ad	The product of the interactions between *sc* and *ad*
sc_m2	The product of the interactions between *sc* and *m2*
sc_u2	The product of the interactions between *sc* and *u2*
sc_p2	The product of the interactions between *sc* and *p2*
sc_a2	The product of the interactions between *sc* and *a2*
ua_pd	The product of the interactions between *ua* and *pd*
ua_ad	The product of the interactions between *ua* and *ad*
ua_m2	The product of the interactions between *ua* and *m2*
ua_s2	The product of the interactions between *ua* and *s2*
ua_p2	The product of the interactions between *ua* and *p2*
ua_a2	The product of the interactions between *ua* and *a2*
pd_ad	The product of the interactions between *pd* and *ad*
pd_m2	The product of the interactions between *pd* and *m2*
pd_s2	The product of the interactions between *pd* and *s2*
pd_u2	The product of the interactions between *pd* and *u2*
pd_a2	The product of the interactions between *pd* and *a2*
ad_m2	The product of the interactions between *ad* and *m2*
ad_s2	The product of the interactions between *ad* and s2
ad_u2	The product of the interactions between *ad* and u2
ad_p2	The product of the interactions between *ad* and p2
m2_s2	The product of the interactions between *m2* and *s2*
m2_u2	The product of the interactions between *m2* and *u2*
m2_p2	The product of the interactions between *m2* and *p2*
m2_a2	The product of the interactions between *m2* and *a2*
s2_u2	The product of the interactions between *s2* and *u2*
s2_p2	The product of the interactions between *s2* and *p2*
s2_a2	The product of the interactions between *s2* and *a2*
u2_p2	The product of the interactions between *u2* and *p2*
u2_a2	The product of the interactions between *u2* and *a2*
p2_a2	The product of the interactions between *p2* and *a2*
x4	dummy variable, 1 if 4 or more dimensions are at level 2 or 3 0 otherwise

A Random effects model was used in the model analysis. Results of the model specifications altogether with the Thai model are shown in Table 5.6.

Table 5.6 Thai model, X4 model and interaction model compared

The Interactions model			The X4 model			The Thai model		
Variables	Coef.	Std. Err.	Variables	Coef.	Std. Err.	Variables	Coef.	Std.Err.
mo	0.129	0.014	mo	0.101	0.012	mo	0.120	0.012
sc	0.130	0.007	sc	0.109	0.008	sc	0.120	0.007
ua	0.119	0.010	ua	0.050	0.007	ua	0.060	0.007
pd	0.095	0.010	pd	0.052	0.013	pd	0.074	0.012
ad	0.091	0.006	m2	0.191	0.019	ad	0.038	0.012
m2	0.316	0.023	p2	0.103	0.019	m2	0.177	0.018
mo_ua	-0.038	0.009	a2	0.090	0.011	p2	0.080	0.019
p2_mo	0.167	0.020	N3	0.148	0.015	a2	0.043	0.019
m2_p2	-0.349	0.038	x4	0.061	0.018	N3	0.138	0.016
constant	0.159	0.015	constant	0.240	0.015	constant	0.200	0.015
No.of observations		7,133			7,133			7,133
R-squared		0.449			0.580			0.448
RMSE		0.108			0.103			0.102
MAE		0.087			0.081			0.080
No.of states with absolute diff.>0.1		34			29			28
No.of inconsistencies		0			48			0
Score for the 2nd best state(11112)		0.750			0.760			0.766
Score for the worst state		-0.440			-0.457			-0.452

The dependent variable of all models is 1 minus the model output. All models have ten significant variables with positive signs except two variables; *mo_ua* and *m2_p2,* in the interactions model. Note that all variables were estimated in the model specifications and only the variables with statistical significant at p-level < 0.05 are presented in Table 5.6. The highest R-squared is 0.580 from the X4 model. The Breuch-Pagan test is used to test the model heteroskedasticity. The null hypothesis of the variances of the model are constant is rejected at p-level = 0.000 in both the interactions and the X4 models. RMSE and MAD are similar across the three models. Thirty-four health states estimated from the interactions model, twenty-nine from the X4 model and twenty-eight from the Thai model have absolute differences between estimated and actual scores exceeding 0.1. The X4 model is the only model that predicts the scores with logical inconsistencies. Among the three models, the Thai model estimated the highest score (0.766) for the second best state (11112), the X4 model predicted the lowest score (-0.457) for the worst state. This score is similar to that predicted by the Thai model.

From Table 5.6, compared with the interactions and the X4 model, the Thai model is still the best model to estimate the preference scores because the model estimated completely consistent scores. The model is slightly more robust compared with the other two because of the smaller number of states with absolute differences exceeding 0.1 and similar RMSE and MAD although the R-squared of the Thai model is lower than the X4 model and similar to the interactions model. The other two models do not predict higher scores for the second best health state although the X4 model does estimate a slightly lower score for state 33333. In conclusion, compared with the interactions and the X4 model, the Thai model is still the best model to estimate the Thai preference scores.

5.5 Impact of choice of subgroups

Scores from the other three subgroups are used to estimate the models in this section. Coefficients of the model specifications are then compared with those estimated from the respondents with fewer than eleven inconsistencies. The model specification procedures are similar to what have been performed in the model specifications using the scores from Subgroup 3 respondents. Recall that for the Dolan (1997) and Dolan & Roberts (2002) models, OLS was initially used to estimate the model which failed the Breusch-Pagan test; the null hypothesis that the model's variances are constant was rejected at the p-value of 0.000. FE and RE models were then applied. The Hausman test is used to choose between the FE and RE model, and it indicated that a RE model is the most appropriate model.

5.5.1 Dolan (1997) model

Table 5.7 presents the parameters estimated from the Dolan 1997 model using the scores from the modelling sample of four respondent subgroups. Only the significant variables (p-value <0.05) are presented. The *s2* and *u2* variables are not significant in the models estimated from the four subgroups. The coefficients of most of the variables are gradually increased using the scores from subgroup 1 to subgroup 4. The variables that are not following this trend are *ad* where the coefficient using subgroup 3 is slightly smaller than that using subgroup 2. Coefficient of *p2* using subgroup 2 is slightly lower than that using subgroup 1. Coefficients of *N3* are gradually increased using the scores from subgroup 1 to 3 but using subgroup 4, the coefficient is lower than that using subgroup 3. Standard errors (SEs) of all variables range from 0.007 to 0.023. Note that although the scores of highly inconsistent respondents are excluded, the SEs are approximately similar across all four subgroups.

Table 5.7 Parameter estimates and the fit statistics of the Dolan 1997 model by subgroup

Variable	Subgroup 1 all respondents		Subgroup 2 with ≤ 16 inconsistencies		Subgroup 3 with ≤ 11 inconsistencies		Subgroup 4 with≤ 6 inconsisten	
	coeff	**SE**	**coeff**	**SE**	**coeff**	**SE**	**coeff**	**SE**
mo	0.100	0.011	0.109	0.011	0.120	0.012	0.137	0.014
sc	0.111	0.007	0.117	0.007	0.120	0.007	0.128	0.008
ua	0.051	0.007	0.055	0.007	0.060	0.007	0.077	0.008
pd	0.063	0.012	0.072	0.012	0.074	0.012	0.092	0.014
ad	0.034	0.012	0.039	0.012	0.038	0.012	0.061	0.014
m2	0.164	0.018	0.172	0.018	0.177	0.018	0.179	0.022
s2	-	-	-	-	-	-	-	-
u2	-	-	-	-	-	-	-	-
p2	0.066	0.018	0.065	0.019	0.080	0.019	0.088	0.023
a2	0.036	0.018	0.039	0.019	0.043	0.019	0.045	0.022
N3	0.124	0.015	0.126	0.016	0.138	0.016	0.101	0.019
_cons	0.267	0.014	0.239	0.014	0.200	0.015	0.116	0.017
N	8,746		8,091		7,133		4,235	
R2(overall)	0.351		0.396		0.448		0.538	
RMSE	0.093		0.094		0.102		0.120	
MAD	0.071		0.073		0.080		0.097	
no.of states with the absolute differences between predicted and actual scores larger than 0.1	21		22		28		21	

The highest R-squared is seen in the model using subgroup 4 respondents. However, the goodness-of-fit statistics (RMSE and MSD) are gradually increased from the models using subgroup 1 to 4. Out of eighty-six states, there are twenty-one states with the absolute difference larger than 0.1 using the scores from subgroup 1 and 4. The scores estimated from subgroup 2 and 3 have twenty-two and twenty-eight states with the absolute difference larger than 0.1, respectively.

5.5.2 Dolan & Roberts (2002) model

Table 5.8 presents the parameters estimated from the Dolan & Roberts (2002) model using the scores from the modelling sample of four respondent subgroups. All variables are statistical significant at p-level lower than 0.05. Two variables have negative signs. The coefficients of almost all of the variables are gradually increased using the scores from subgroup 1 to subgroup 4. The variables that are not following this trend are *ANY13* where the coefficient using subgroup 1 is similar to that using subgroup 2 and slightly similar to that using subgroup 3. The constant terms are gradually decreased from subgroup 1 to 3 and slightly increased using subgroup 4. Mean scores for state 33333 are also gradually decreased from subgroup 1 to 4. The SEs are approximately similar across all four subgroups.

Table 5.8 Parameter estimates and the fit statistics of the Dolan & Roberts 2002 model by subgroup

Variable	Subgroup 1 all respondents		Subgroup 2 with ≤ 16 inconsistencies		Subgroup 3 with ≤ 11 inconsistencies		Subgroup 4 with≤ 6 inconsistencies	
	coeff	SE	coeff	SE	coeff	SE	coeff	SE
difmob1	0.277	0.011	0.294	0.011	0.310	0.012	0.325	0.014
difmob2	0.401	0.013	0.427	0.014	0.457	0.014	0.480	0.016
difsc1	0.111	0.012	0.116	0.012	0.123	0.012	0.137	0.014
difsc2	0.251	0.013	0.264	0.013	0.271	0.013	0.279	0.016
difua1	0.057	0.011	0.064	0.011	0.066	0.012	0.093	0.014
difua2	0.138	0.013	0.146	0.013	0.161	0.013	0.184	0.015
difpd1	0.142	0.011	0.150	0.011	0.169	0.011	0.190	0.014
difpd2	0.229	0.013	0.245	0.013	0.268	0.014	0.303	0.016
difad1	0.086	0.011	0.094	0.011	0.099	0.011	0.119	0.013
difad2	0.148	0.013	0.161	0.013	0.167	0.013	0.201	0.015
ANY13	-0.075	0.010	-0.075	0.010	-0.073	0.010	-0.049	0.012
_cons	-0.038	0.013	-0.047	0.013	-0.055	0.013	-0.051	0.015
Mean score of state 33333	-0.346		-0.386		-0.419		-0.484	
N	8,746		8,091		7,133		4,235	
R2(overall)	0.351		0.396		0.447		0.538	
RMSE	0.095		0.097		0.106		0.112	
MAD	0.075		0.077		0.085		0.089	
no.of states with the absolute differences between predicted and actual scores larger than 0.1	22		29		30		33	

The highest R-squared is seen in the model using subgroup 4 respondents (0.538). However, the goodness-of-fit statistics are gradually increased from the models using subgroup 1 to 4. Out of eighty-six states, there are twenty-two states with the absolute difference larger than 0.1 using the scores from subgroup 1, twenty-nine in subgroup 2, thirty in subgroup 3 and thirty-three in subgroup 4.

5.5.3 Shaw *et al.* (2005) model

Table 5.9 presents the parameters estimated from the Shaw *et al.* (2005) model using the scores from the modelling sample of four respondent subgroups. The first model used to estimate the models was OLS model. The null hypothesis of equal variance was rejected and the Feasible generalized least square (FGLS) was used with no constant term. Only significant coefficients (p-level < 0.05) are used in the model analyses. Not all variables are statistical significant at p-level lower than 0.05. The *i2-squared (i22)* variable is not significant in the model using the scores from subgroup 4. The *i3* variables are not significant using the scores from subgroup 1 to 3. Two variables; *i3-squared (i32) and d1* have negative signs. The *i2-square (i22)* is negative using subgroup 1 to 3. The coefficients of almost all of the variables are gradually decreased using the scores from subgroup 1 to subgroup 4. The variable that is not following this trend is *d1* where the coefficients are gradually increased from subgroup 1 to 4. Note that SEs are approximately similar across all four subgroups.

Table 5.9 Parameter estimates and the fit statistics of the Shaw *et al.* 2005 model by subgroup

Variable	Subgroup 1 all respondents		Subgroup 2 with ≤ 16 inconsistencies		Subgroup 3 with ≤ 11 inconsistencies		Subgroup 4 with≤ 6 inconsistencies	
	coeff	**SE**	**coeff**	**SE**	**coeff**	**SE**	**coeff**	**SE**
m1	0.321	0.008	0.311	0.008	0.289	0.008	0.240	0.010
m2	0.632	0.015	0.626	0.015	0.615	0.015	0.554	0.020
s1	0.346	0.009	0.333	0.009	0.305	0.009	0.246	0.011
s2	0.548	0.016	0.537	0.016	0.525	0.017	0.459	0.021
u1	0.298	0.008	0.283	0.009	0.261	0.009	0.207	0.011
u2	0.445	0.014	0.437	0.014	0.423	0.015	0.368	0.019
p1	0.312	0.009	0.301	0.009	0.273	0.009	0.238	0.011
p2	0.528	0.015	0.519	0.016	0.510	0.016	0.467	0.020
a1	0.277	0.008	0.262	0.008	0.236	0.009	0.204	0.010
a2	0.452	0.015	0.441	0.015	0.427	0.016	0.391	0.019
i2	0.073	0.015	0.071	0.016	0.069	0.016	0.058	0.018
i22	-0.011	0.004	-0.011	0.004	-0.010	0.004	-	-
i3	-	-	-	-	-	-	0.057	0.023
i32	-0.024	0.001	-0.025	0.001	-0.027	0.001	-0.036	0.003
d1	-0.300	0.013	-0.282	0.013	-0.256	0.013	-0.211	0.018
N	8746		8091		7133		4235	
RMSE	0.232		0.247		0.257		0.277	
MAD	0.179		0.193		0.199		0.206	
no.of states with the absolute differences between predicted and actual scores larger than 0.1	58		61		59		57	

The goodness-of-fit statistics are gradually increased from the models using subgroup 1 to 4. The greatest RMSE and MAD are seen in the model using the scores estimated from subgroup 4. Out of eighty-six states, there are fifty-eight states with the absolute difference larger than 0.1 using the scores from subgroup 1, sixty-one from subgroup 2, fifty-nine and fifty-seven from subgroup 3 and 4 respectively.

5.5.4 The comparison of scores estimated from all models

The scores in the following tables are estimated from the full sample (that is, the modelling and validation samples combined). Table 5.10 compares the estimated scores estimated from the Dolan (1997) model across four subgroups. Using the scores from subgroup 4, the model estimates the highest score for state 11112 and the lowest score for state 33333. As a result, the greatest range of the best-worse scores is also identified from this subgroup. The model estimated from subgroup 3 gives the greatest number of negative scores [68]. None of the four subgroups estimates the inconsistent scores. Subgroup 4 has the highest Cohen effect size.

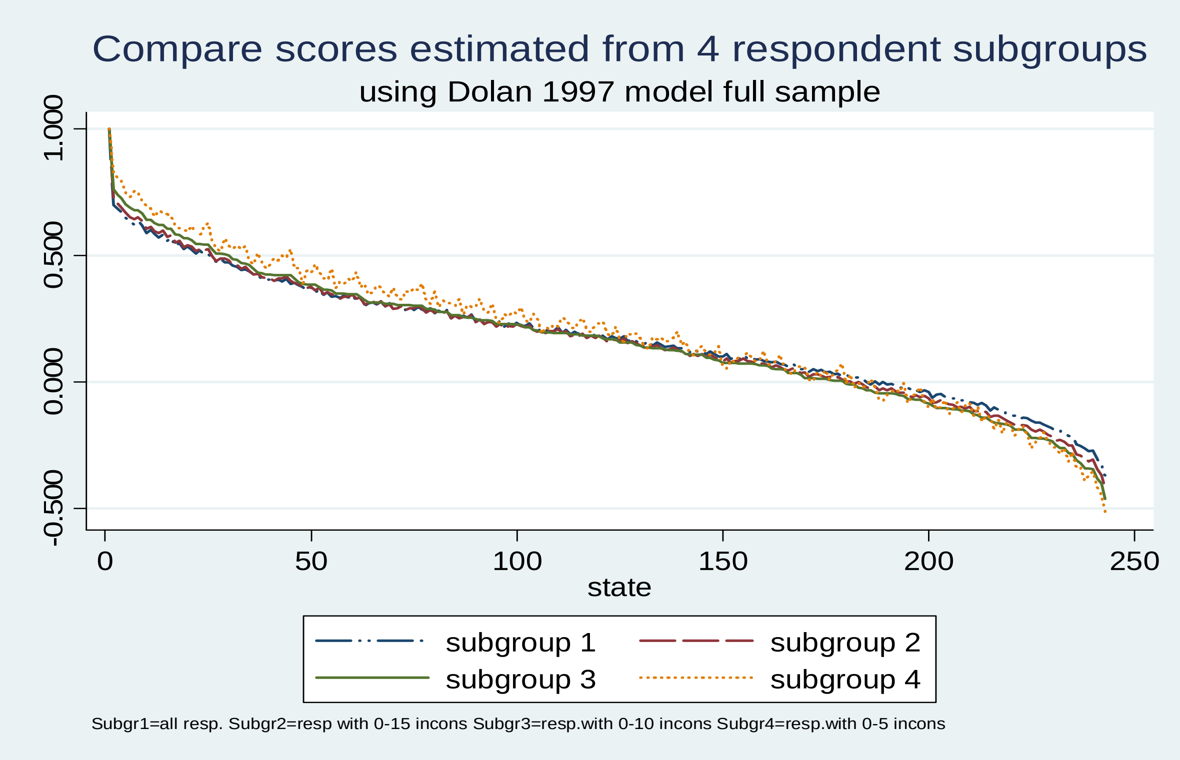
Table 5.10 Comparison of the scores estimated from the Dolan 1997 model by subgroup

	Respondents			
	Subgr 1	**Subgr 2**	**Subgr 3**	**Subgr 4**
Best ill health score (state 11112)	0.707	0.729	0.766	0.829
Worst state score (state 33333)	-0.373	-0.420	-0.452	-0.513
Range from best-worst score	1.373	1.420	1.452	1.513
Number of negative scores	54	64	68	62
Number of logical inconsistency	0	0	0	0
Cohen effect size	1.087	1.084	1.084	1.400

To see the differences of the scores estimated from the Thai model using the scores from all four subgroups, all 243 scores estimated from all four subgroups are ranked from the highest to lowest scores according to those estimated from Subgroup 3. The differences are shown in Figure 5.2.

Y-axis represents the scores ranging from -0.50 to 1. X-axis represents the health states. For most of the health states better than death, the scores estimated from Subgroup 4 are slightly higher than those estimated from other subgroups. Regarding the health states worse than death, most of the scores estimated from Subgroup 1 tend to be higher than those estimated from other subgroups. The scores estimated from the respondents in subgroup 3 and 4 are quite similar with each other.

Figure 5.2 Estimated scores comparison from 4 respondent subgroups



To see whether the model coefficients estimated from Subgroup 2 to 4 differ significantly from those estimated from Subgroup 1, 95% confidence intervals (CIs) of the model coefficients are compared and presented in Table 5.11. The confidence intervals for all coefficients estimated from Subgroup 2 overlap those estimated from Subgroup 1. The constant terms estimated from Subgroup 3 and 4 do not overlap that for Subgroup 1. The findings could be interpreted that by excluding the scores from the highly ***inconsistent*** respondents (≥ 15 inconsistencies), the resulting model is not significantly different from that estimated from the scores from all respondents. This is in contrast with the models estimated from the scores given by the higher ***consistent*** respondents (Subgroups 3, 4) where the resulting models are slightly different from that estimated from all respondents, only the differences in constant terms are significant.

Table 5.11 95% CIs of coefficients estimated from four subgroups using the Dolan 1997 model

variables	Subgr 1 (all respondents) 95% CI of coeff.		Subgr 2 (<16 incons) 95% CI of coeff.		Subgr 3 (<11 incons) 95% CI of coeff.		Subgr 4 (<6 incons) 95% CI of coeff.	
	lower limit	upper limit	lower limit	upper limit	lower limit	upper limit	lower limit	upper limit
mo	0.078	0.122	0.087	0.131	0.098	0.143	0.111	0.164
sc	0.098	0.124	0.105	0.130	0.107	0.133	0.113	0.144
ua	0.038	0.064	0.041	0.068	0.046	0.073	0.061	0.093
pd	0.040	0.087	0.048	0.095	0.050	0.098	0.064	0.121
ad	0.011	0.057	0.016	0.062	0.014	0.061	0.032	0.089
m2	0.129	0.199	0.136	0.207	0.141	0.213	0.136	0.222
p2	0.030	0.102	0.028	0.101	0.042	0.117	0.043	0.132
a2	0.000	0.072	0.002	0.075	0.005	0.080	0.001	0.089
N3	0.094	0.155	0.095	0.157	0.107	0.169	0.064	0.138
constant	0.239	0.294	0.211	0.267	0.172*	0.229*	0.082*	0.150*

* 95% CI not overlapped with Subgr 1

Table 5.12 compares the scores across the four subgroups estimated from the Dolan & Roberts (2002) model. Using the scores from subgroup 4, the model estimates the highest score for state 11112 and the lowest score for state 33333. As a result, the greatest range of the best-worse scores is also identified from this subgroup. The model estimated from subgroup 3 gives the greatest number of negative scores. The scores estimated from subgroup 1 to 3 are logically inconsistent. The greatest Cohen effect size is seen in the model estimated from subgroup 4.

Table 5.12 Comparison of the scores estimated from the Dolan & Roberts 2002 model by subgroup

	Respondents			
	Subgr 1	**Subgr 2**	**Subgr 3**	**Subgr 4**
Best ill health score (state 11112)	0.727	0.750	0.782	0.830
Worst state score (state 33333)	-0.383	-0.428	-0.469	-0.531
Range from best-worst score	1.383	1.428	1.469	1.531
Number of negative scores	59	66	67	60
Number of logical inconsistency	15	15	15	0
Cohen effect size	1.089	1.086	1.083	1.390

Table 5.13 compares the scores estimated from the Shaw *et al.* 2005 model across four subgroups. Using the scores from subgroup 4, the model estimates the highest score for state 11112 but the lowest score is estimated for state 33133 rather than state 33333. The models using the scores from subgroup 1 to 3 estimated the lowest scores for state 33232. The greatest range of the best-worse scores is identified from subgroup 4. The model estimated from subgroup 4 gives the greatest numbers of negative scores (15). None of the scores estimated from all four subgroups are completely consistent. The smallest number of logical inconsistencies is obtained from the model estimated from subgroup 3 (37). The greatest Cohen effect size is seen in the model estimated from subgroup 2.

Table 5.13 Comparison of the scores estimated from the Shaw *et al.* 2005 model by subgroup

	Respondents			
	Subgr 1	**Subgr 2**	**Subgr 3**	**Subgr 4**
Best ill health score (state 11112)	0.723	0.738	0.764	0.796
Worst state score	-0.049 (33232)	-0.059 (33232)	-0.074 (33232)	-0.085 (33133)
Range from best-worst score	1.049	1.059	1.074	1.085
Number of negative scores	10	12	13	15
Number of logical inconsistency	54	39	37	40
Cohen effect size	1.030	1.034	1.023	0.979

5.6 The Thai algorithm and the preference scores

The Thai model is estimated from the full sample of Subgroup 3 using the Dolan (1997) model. Coefficients of the Thai model are presented in Table 5.14.

Table 5.14 Coefficients of the variables in the Thai model

variable	coefficients
constant	0.202
mo	0.121
sc	0.121
ua	0.059
pd	0.072
ad	0.032
m2	0.190
p2	0.065
a2	0.046
N3	0.139

Thai preference scores are calculated from the following algorithm.

Thai score = 1-0.202-(0.121*mo)-(0.121*sc)-(0.059*ua)-(0.072*pd)-(0.032*ad)-(0.190*m2)-(0.065*p2)-(0.046*a2)-(0.139*N3)

The Thai preference scores for all 243 states are presented in presented in Appendix 4. According to the Thai model, Thai preferences on health seem to be reduced mostly by having severe problems (level 3) in mobility and self-care followed by having severe problems in pain and discomfort. Table 5.15 shows the comparisons of the model coefficients between the Thai model coefficients and the UK model by Dolan (43).

Table 5.15 **Comparison of the models**

variables	*coefficients*	
	Thai	UK
a	0.202	0.081
mo	0.121	0.069
sc	0.121	0.104
ua	0.059	0.036
pd	0.072	0.123
ad	0.032	0.071
m2	0.190	0.176
s2	-	0.006
u2	-	0.022
p2	0.065	0.140
a2	0.046	0.094
N3	0.139	0.269
MAD	0.080	0.039

Note: MAD-Mean absolute difference

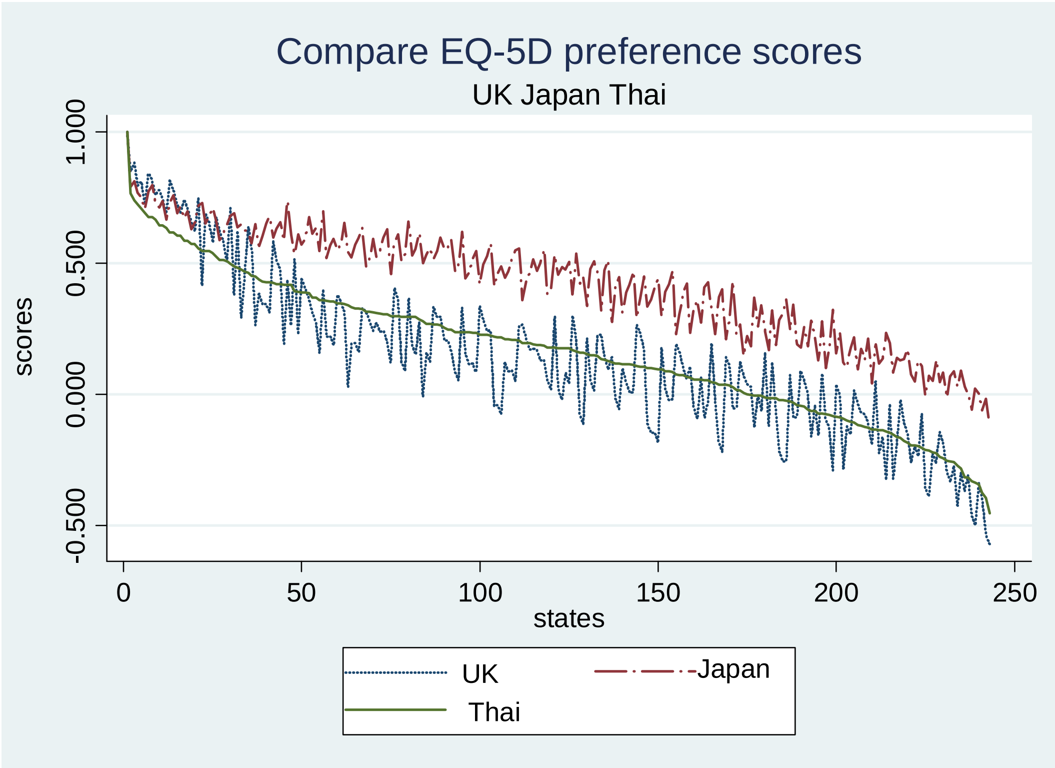
(43)

The MAD of the Thai model is twice that of the UK model. The constant term, coefficients of variables *mo, sc, ua* and *m2* are higher in the Thai model. In the Thai model the coefficients of the variables u2 and s2 are not statistically significant and are not included in the final model.

Graphical illustrations of the score comparisons between the Thai, UK and Japanese scores

The Thai scores are compared to the UK scores because the UK scores have been treated as the reference set to which a number of countries' scores have been compared. The Thai scores are also compared with those for Japanese since Japan is also located in Asia. All states between 11111 and 33333 are ranked according to the estimated Thai scores. Line graphs are plotted where the Y-axis represents the scores and the X-axis represents the ordered EQ-5D states, therefore, State 1 in the X-axis is 11111 and state 243 is 33333. The comparisons of the scores from the UK and the Japanese scores are presented in Figures 5.3.

Figure 5.3 **Compare the UK, Japan and Thai scores**



(43, 48)

In general, the Japanese scores are higher than those of the UK and Thai scores except the scores for the mild states where the Japanese and the UK scores seem to be fairly similar. In general, the Thai scores are similar to those of the UK scores. Regarding the scores for the mild states, the Thai scores are lower than both the UK and Japanese scores, for the severe states, the Thai scores seem to be higher than those of the UK but still much lower than the Japanese scores.

5.7 Discussion

This chapter reports the results of using the scores from Subgroup 3 to estimate the model using the variables from the Dolan (1997), Dolan & Roberts (2002) and Shaw *et al.* (2005) models. As reported in other studies, the Thai model was also estimated using the Random effects model. To see the impacts of the choices of subgroups, the scores from the other three subgroups were also used in the model analysis. It is shown that excluding the scores from the inconsistent respondents have impacts on the models coefficients and performances. The arbitrary classification of respondents was applied in this study, had the respondents been classified using different number of inconsistent responses, the resulted coefficients of the models would be changed.

The comparisons of the model estimated from the other three subgroups are to reassure that the Thai model could be "best" estimated using the Dolan 1997 model from Subgroup 3. The competitive model would be the Dolan & Roberts 2002 model using the respondents in subgroup 4. By using the latter model, the scores are completely consistent and the score for the best ill health state is higher and the score for the worst state is lower compared with those estimated from the Dolan 1997 model. However, the scores from Subgroup 4 are not favoured to estimate the Thai scores and if the scores from Subgroup 3 were used to estimate the scores using the Dolan & Roberts 2002 model, there are inconsistent responses in the estimated scores.

One may argue that if the scores from all respondents are favoured to be used in the final model estimation, the model estimated from subgroup 2 could be more appropriate than those estimated from subgroups 3 and 4 because none of the coefficients are significantly different from those estimated from subgroup 1. However, it is shown in the previous chapter that the scores from all respondents were not favoured because this subgroup includes the highly inconsistent respondents who may have had difficulties in assigning the scores in the TTO interview. Although the respondents with greater than fifteen inconsistencies were excluded, the exclusion may not be sufficient to make significant differences from the model estimated from all respondents. This makes the model estimated from subgroup 2 less favoured than the models estimated from subgroups 3 and 4.

Using the scores from subgroups 3 and 4, significant changes from the model estimated from all respondents can be identified although that change was seen only at the constant term. This can be used as an evidence to support an argument that excluding the highly inconsistent scores at the "appropriate" number, the models are changed. The changes in the models can be both regarded as "justifiable" or "unjustifiable". It is "justifiable" because the resulting models perform better (higher R-squared) and the estimated scores are systematically changed in a favourable fashion (minimised ceiling and floor effects). It may be "unjustifiable" because the scores of some respondents were excluded and the data were lost from the model analysis. However, the exclusion was justified here because the excluded scores were given by respondents who may have had difficulties in participating in the TTO interview. This can be used to support the selection of subgroup 3 to model the Thai preference scores.

Regarding the model performance, after excluding the inconsistent responses from the model specifications, the R-squared of the models estimated from the scores with lower numbers of inconsistent responses are higher than those with higher numbers of inconsistent responses. However, higher R-squared alone does not justify the better performance of the models. The estimated scores from the models with a lower number of inconsistent responses were much different from the actual scores. One reason for the larger differences could be that the numbers of observations were smaller in the sample with lower number of inconsistent responses. The estimated score for the second best state was higher and that of the worst state was lower after excluding the scores from highly inconsistent respondents. This evidence could be used to support that the inconsistent respondents were likely to assign scores at random with least correlations with the severity of health states. This could be the result of respondents not understanding the health state descriptions or the tasks or they were lack of concentration when participating in the interview.

The criteria used to select the best model in this study are slightly different from those reported in the Dolan (1997) study in that the responsiveness of the scores to changes in health was added to the set of criteria. It is shown that the predicted scores from all three models have high responsiveness. Note that the responsiveness in this study was estimated from the comparisons of the estimated scores with all possible positive transformations which included the transformations from the worst state to full health. This may not be the case in the real-life situations where the scores are used to measure QALYs gain from health interventions. More research should be conducted to develop more insights regarding the responsiveness of the Thai preference scores.

According to the selection criteria, the Dolan (1997) model was chosen to estimate the Thai scores. The Thai model still suffers from heteroskedasticity which is in line with the models used to estimate the preference score for EQ-5D health states in other countries. Compared with the UK model where there were twelve variables included in the algorithm, there were only ten variables in the Thai algorithm. In the UK model, variable *s2* was insignificant (at p-level <0.05) whereas variables *s2* and *u2* were insignificant in the Thai model. Unlike the UK model in which the “insignificant” variable was included, those insignificant were dropped in the Thai model because the models then performed slightly better. Moreover, both excluded coefficients had negative signs, by including these coefficients in the algorithm, the estimated scores for the states with level 3 in self-care and usual activities, other things being equal, would have been slightly higher than those estimated from the algorithm without these coefficients.

When adding the interaction terms into the Thai model, heteroskedasticity still exists. From the interactions model, there existed two patterns of interactions between dimensions: [1] mobility and usual activities and [2] mobility and pain/discomfort. Variable *N3* was not significant in this model. These findings provide more insights on the impacts on Thai preferences on health from the interactions between the attributes of health.

The Thai model tends to predict lower scores than the actual ones for health states with no problem in mobility and self-care but some or extreme problems in the last three dimensions. The reason could be that the respondents may have paid more attention to the first two dimensions on the health cards. If there is no problem in the first two dimensions, the Thai respondents may have gained the impression that "this sounds good to me" and paid less attention to the last three dimensions, the respondents then assigned high score for this state. This assumption could be applied to the states with extreme problems in the first two dimensions and no or some problems for the last three dimensions, at which the respondents may have bad impressions after reading only the first two dimensions and gave lower scores for this state than predicted by the model. If this is the case, further study should be aware of this problem and the interviewers should be certain that the respondents take all dimensions into consideration before assigning scores.

As far as I know, this is the first study using the Stata program to detect logically inconsistent responses in the estimated scores. The resulting estimates of logical inconsistency were re-examined by using the model coefficients to identify the possibilities of inconsistent responses if the coefficient of, for example, being in level 2 was higher than that of level 3. If this is the case, other things being equal, the resulting score for the better state could be lower than that of the poorer state. I am confident that the scores estimated from the Thai model are completely consistent. Unlike the scores reported in the Dolan & Roberts (2002) and Shaw *et al.* (2005) models where one of the criteria to select the best model was logical inconsistency in the estimated scores, however, by using Stata program to search for inconsistent responses, some logical inconsistent responses in the estimated scores were detected; sixty states from the Dolan & Roberts 2002 using the UK data and fifteen states from the Shaw *et al.* (2005) using the US data.

Thai preference scores differ from the UK and the Japanese scores. Possible causes of the differences in preferences may be the differences in health systems, cultures and religious beliefs of the population. Terminologies used in the translated versions of EQ-5D may also play some roles in that the terminologies used to describe the same health states in different languages may convey different meanings. If the Thai scores were unavailable, the UK scores would be, among the models presented in this chapter, the best scores to estimate the Thai QALYs. Thais and British seem to elicit fairly similar scores for the same health states. The states with highly differences between the two sets of the scores can be found at mild and severe states.

5.8 Conclusion

The Thai algorithm is based on the Dolan (1997) model using data from respondents with fewer than 11 inconsistent responses. This model is chosen because the resulting algorithm produces no logically inconsistent scores, the model is the most parsimonious, highly robust and the responsiveness is acceptably high. The effect of using the scores from other subgroups was explored to see the differences of the models across all four subgroups. The constant terms in the models were significantly changed after the scores from the highly inconsistent respondents were excluded from the model specifications. Interaction terms were added in the Thai model to eliminate the problem of heteroskedasticity but did not yield a superior algorithm. The models estimated from the fewer number of inconsistent respondents predict a higher score for the best ill health state and lower score for the worst state. The Thai model predicts approximately twenty per cent of the scores with the absolute differences from the actual scores exceeding 0.1. Compared with the UK and the Japanese scores, the Thai scores seem to closely correlate with the UK scores.

Chapter 6 Discussion and Conclusion

6.1 Introduction

The main finding reported in this document is the estimation of preferences on health from a representative sample of the Thai general population. The EQ-5D health states were used as a "proxy" to describe health and three preference elicitation methods; ranking, VAS and TTO, were used to derive preference scores from the Thai respondents. The MVH protocol was adapted to be used in a face-to-face interview in the fieldwork survey and almost one-third of all 243 EQ-5D health states were directly valued. This chapter is organised as follows. The first section focuses on the contribution of the thesis, followed by the second section which identifies the limitations of the thesis. The final section describes the priorities for future studies and the overall thesis summary.

6.2 Contributions

6.2.1 The first Thai population-based preference scores for EQ-5D health states

This is the first study to estimate preference scores for health from the Thai general population. The conventional interview procedure, i.e. the MVH protocol was redesigned and implemented and the scores were estimated using a Random Effects model which is the model most commonly used to estimate EQ-5D scores. The preference scores can be used to measure health outcomes in economic evaluation.

The criteria to select the best model stated in this study differ from those used in the previous preference studies. The criteria used in other studies were as follows: logical consistency of the modelled scores, model parsimony and robustness. As stated in Chapter 5, the criteria used in this study were the aforementioned three criteria and additionally, responsiveness of the modelled scores. The additional criterion was used to ensure the superiority of the selected model.

To be certain that the modelled scores are consistent, an innovative method to detect logical inconsistencies in the modelled scores was proposed in Chapter 5 using Stata program. Although, the estimated scores were thoroughly examined up until the certain level, it is likely that almost all of the possible inconsistently responses were identified. This method also successfully identified logical inconsistencies in the UK preference scores (using the Dolan & Roberts (2002) model) and the US preference scores (using the Shaw *et al.* (2005) model).

Unlike previous studies, the "best model" was selected not only on the basis of model performance, but also ensured that the scores came from an "appropriate respondent subgroup". Moreover, to confirm that the model was estimated from the appropriate subgroup, scores from the other subgroups were also used to estimate the models and their performances were compared with those estimated from the selected subgroup. Regarding specification of the Thai model, almost all variables and models proposed in the literature on the estimation of preference scores for EQ-5D states were utilised. Among the three models used in this study: Dolan (1997), Dolan & Roberts (2002) and Shaw *et al.* (2005) models, the Dolan (1997) model is the best model to explain Thai preference scores. Other interaction terms proposed in previous studies were used in the final model but the model performance was not improved.

The Thai preference scores estimated from this study contribute to the research community in economic evaluation in Thai settings in that QALYs of Thai population are able to be estimated using Thai preferences. This is in line with the recommendations in the economic evaluation guideline developed by the Health Impact and Technology Assessment Program (HITAP) Thailand (12). Researchers in the economic evaluation field can be reassured that the Thai preference scores were estimated using standard methodologies and the scores were derived from a large population-based survey. The Thai preference scores could also contribute by being an additional set of scores in the list of available preference scores provided by the EuroQol group.

The Thai preference scores could also be a substantial input to health outcome research communities worldwide in that the Thai preference scores could be used in cost-utility analysis conducted in the neighbouring countries, for example, Lao People's Democratic Republic, Vietnam or Malaysia, where their own preference scores are yet to be established. Given that the Thai scores were derived from the respondents likely to have similar demographic characteristics with their general population, the Thai preference scores could possibly approximate preferences on health for their general populations. The Thai preference scores could be used in the sensitivity analysis of health state values in cost-utility studies conducted in other countries.

6.2.2 Successful administration of a preference survey in Thailand

Face-to-face preference interviews were successfully conducted in a representative sample of the Thai general population. Respondents from every walk of life from all over the country were reached by the research team. Users of the preference scores can be reassured that the scores were derived, not only from the respondents who could commute to the research team, but also from those who may not be able to travel to the research office. Regarding fieldwork management, the key factors for the successful completion of the preference elicitation interviews, rest on a combination of close collaboration and good communication between the research team and the fieldwork coordinators. By providing a wider range of interview settings, a greater number of the respondents could be reached. In other studies, the respondents were interviewed either in their households or at the offices of the research organisations. It would be prohibitively expensive in Thailand if all respondents were invited to be interviewed in the researcher's offices. Moreover, if the interviews were scheduled to be conducted in the respondents' households, some may feel uncomfortable having the interviewers entering their premises. By providing interview sites in the various respondent neighbourhoods, a greater number of the respondents could be encouraged to participate including those who could not previously be identified in the fieldwork preparation phase because their addresses were not known to the field coordinators. This group of respondents were successfully contacted later in the fieldwork because some of them were known by the respondents who participated in interviews. In addition, by conducting the interview in their neighbourhoods, it was easy for them to travel to the interview sites.

Before implementing the adapted MVH protocol, the feasibility of conducting the preference elicitation interview using the original and re-designed MVH protocols was thoroughly explored and the changes were made in the protocol before implementation in this study. By conducting the pilot studies, it was learned that the original MVH protocol was unlikely to be appropriate to be conducted in the Thai general population. One reason is that the original protocol was the result of vigorous research and successful implementation in the UK, rather than in Thai settings. It is seen that some countries have administered the original protocol in their own settings. However, the original protocol may not be a "one size fits all" for other countries. Given the complexities of the tasks, efforts should be made to reduce the cognitive workload and the protocol should be redesigned to be more suitable to the competency of the population before implementing the survey.

6.2.3 Greater number of health states used in the fieldwork survey

A greater number of health states were valued directly in the Thai study. It is expected that the higher number of health states used in the direct observations of the preferences, should facilitate the accurate estimation of the model to assign scores to all possible states. However, it was shown that the Thai model still suffers from heteroskedasticity, and ceiling and floor effects of the estimated scores as in the previous studies. The results of the Thai model, as reported in Chapter 5 indicate that there were a considerable number of states with differences between the actual and predicted scores, this can be used to support an argument that simply increasing the number of health states in the preference interview may not be able to improve the model performances. This finding may shed more light on the limitations of the EQ-5D measure for describing the very mild and the very poor health outcomes. In addition, this can generate more questions on the appropriateness of using additive modelling to estimate the scores for EQ-5D health states.

The classification of health states into "mild" and "severe" categories in this study are more convincing than those classified in the UK study by Dolan in 1997. In this study, there was no level 3 in "mild" states as opposed to the "mild" states categorised in Dolan where level 3 was in up to two dimensions. The poorest health state in the "mild" category was state 21333 and the estimated score for this state is -0.110 as opposed to the score estimated for the poorest state in this study (22211) for which the score is 0.497. Regarding the "severe" state classification, there is no level 1 in the "severe" states used in the Thai study whereas the best "severe" state is state 13332 in the UK study. The estimated score for this state is 0.170. The "best" severe state in the Thai study is state 22233 and the estimated score is 0.039 which is lower than the best "severe" state in the UK study. It is unconvincing to include states with level 3 in the mild category and states with level 1 in the severe category.

One distinct methodology employed in this study is that, unlike some previous studies where health states were randomly chosen to present to respondents, all health states used in the interview were grouped as 12 sets with 11 health states in each set. This approach has three advantages; the first is that the numbers of observations per health state can be determined prior to the interview. This method was employed to ensure a somewhat similar number of observations per health state. Secondly, the complexity of the interview procedure for the interviewers could be reduced. Rather than requiring the interviewers to randomly choose health states from each severity category (as conducted, for example, in the UK study), it would be easier for the interviewers to simply randomly choose one of the eleven health states already prepared in the sets.

6.3 Limitations

All research is subject to a number of limitations and this thesis is not any different. The areas which it is appropriate to highlight are as follows: the exclusion of some of the directly observed TTO scores from the Thai model estimations; the modifications of the original MVH protocol; the cognitive burden facing respondents; issues with the time horizon when eliciting TTO scores; the representativeness of the sample; the number of interviewers and the interview sites; and number of observations per health state.

6.3.1 Exclusion of some directly observed TTO scores from the Thai model estimations

Not all directly observed TTO responses were used when estimating the Thai tariff because it is likely that the respondents giving a high number of inconsistent responses may not understand the interview task and may be randomly assigning scores to the health states. By including the scores from this group of respondents, the Thai health state values may be distorted. Only the scores from respondents with less than a given number of logically inconsistent responses were used to estimate the Thai model. To find the "appropriate" number of the inconsistent responses to lead a respondent to be excluded, respondents were arbitrarily classified into four groups. If the respondents were classified using different criteria, the estimated scores could have been differed from the results in this study. The decision as to where to set the threshold for exclusion is arbitrary and involved a judgement balancing a desire to retain data with a desire to avoid using data which might "mislead". The selection of this group of respondents does not imply that this level of inconsistency should be accepted as "normal" in the Thai respondents. The selection is based solely on the performance of the model coefficients and the estimated scores.

This study does not offer a novel model specification to estimate preferences for EQ-5D health states but utilises existing model specifications. According to the model selection criteria, the Dolan 1997 model seems to be the "best" model and performed fairly well in the score estimation but the Thai model still suffers from heteroskedasticity and the resulting model predicts the scores with some degree of errors in that approximately ten percent of the total health states have the differences between the actual and estimated scores exceeding 0.1.

The largest differences between actual and predicted scores seem to occur in health states with particular patterns. For example, the estimated scores are likely to be lower than observed scores for the states with some problems in the latter three dimensions (usual activities, pain/discomfort and anxiety/depression). One possible explanation would be that if Thais see that there is no problem in mobility and self-care, they may already "prefer" this health state, no matter what levels occur in the following dimensions. This may imply that some Thai respondents pay attention to only these two "key" dimensions. This finding could be used to support an argument that respondents may have used partial information on health states when deciding how to trade-off time to stay in full health. This assumption should be explored further to gain more understanding of the mechanisms employed by respondents when answering the preference questions.

6.3.2 Modifications to the original MVH protocol

Regarding the respondent performance in the preference elicitation interview, it seems that some Thai respondents may not understand the TTO task at the beginning and may have learned how to respond to the TTO questions later on. Although, the Ranking and VAS methods were used before the TTO method to give the respondents a "warm-up" exercise, this exercise only allows the respondents to familiarise themselves with health state descriptions, rather than with the TTO method of assigning values to health states.

To minimise the cognitive workload for the Thai respondents, two health states: "immediate death" and "unconscious" which are not EQ-5D states were excluded from the fieldwork interview protocol. One reason for the exclusion was to reduce the number of health states used in the ranking and VAS interviews, thus the cognitive workload of the respondents could be minimised. Only the health states used in the TTO interview were administered in the previous two interview methods to ensure that the respondents had chances to familiarise themselves with the health states before moving to the TTO method. But the exclusion of these two states means that the interview in the Thai study differed from the original MVH protocol. A consequence of excluding "immediate death" is that preference scores cannot be estimated from the Ranking and VAS data. However, the primary objective of the study is to estimate Thai preference scores and it was decided that the scores would be estimated using the TTO method. Therefore, the exclusion of these two health states is unlikely to jeopardise the estimation of the Thai preference scores. Another weakness of the exclusion would be that the respondents may have missed the opportunity to practice imagining health states as worse than death. However, this opportunity would have come at the cost of additional workload for the respondents in the ranking and VAS interviews.

6.3.3 Cognitive burden facing respondents

A cognitive burden could be the result of several factors namely: the descriptions of health states in Thai; the complexity of the TTO task; and the illness experience of the respondents. It might be difficult for the respondents to imagine themselves being in the hypothetical health states. The descriptions of the health states may increase the cognitive burden on respondents. The health state descriptions in Thai appear to be unclear and ambiguous to the respondents. Some of the respondents expressed their concerns regarding the ambiguous vocabulary in the self-completed questionnaire at the end of the elicitation interview. This issue could also jeopardise the "conceptual equivalence" of the Thai version of EQ-5D. The respondents could be confused by the descriptions on the health cards. For example, the card describing state 31311 with the first two dimensions read "*Mai Samart Pai Nai Dai Lae Jam Pen Tong U Bon Tiang*" for mobility and "*Mai Mee Pan Ha Dan Karn Doo Lae Ton Eng*" for self-care. Descriptions in the first two dimensions (level 3 in mobility and level 1 in self-care) begin with the term "*Mai*" but "*Mai*" in mobility is for level 3 (negative "*Mai*") and "*Mai*" in self-care is for level 1 (positive "*Mai*"). This may confuse the participants especially those who are elderly with only primary education and poor reading ability. They may have thought that either mobility and self-care are at level 1 (positive "*Mai*") in state 31311 or that both dimensions are at level 3 (negative "*Mai*").

As stated by Brazier *et al.*, values are sought from the "informed general population"(70). From the results of this study the extent to which the Thai respondents can be "informed" given the health cards used in the interview to describe the Thai EQ-5D health states is uncertain and the respondents may not be well "informed" before giving scores to health states. The comments provided by the respondents can be used to support an argument that the respondents may have had difficulties in understanding the health states described on the cards. In addition, the respondents may be "well informed" but when it comes to making a decision, the respondents may have taken only partial information (of the health states) into account or used external information in their decision making. Individual's "personal habit" with respect to making decisions may play some role at this stage in that, some respondents would prefer to carefully consider all the dimensions described in the health card before assigning the scores. On the contrary, some respondents may have rushed through the decision process and assigned scores to health states without thoroughly contemplating themselves in the given states.

6.3.4 Time horizon for the TTO questions

The users of the Thai scores should be aware of the time horizon used in the TTO elicitation methods in this study. The preference scores in this study were based on trading-off time within a ten year life expectancy. If the respondents were given a longer or shorter duration of life expectancy, the scores given by the respondents would have been different. Whether the Thai general population have a maximal endurable time (MET) should be explored, and if they do ways of correcting the TTO scores should be considered. The other TTO assumption which could be violated is that the Thai respondents may express diminishing marginal utility (DMU) in which the proportion of time traded-off may not be constant. This may cause problems when QALYs are estimated for diseases with life expectancy of longer than 10 years. The Thai scores estimated in this study may well not represent the preferences of the general population towards these particular diseases.

Biases may occur in the scores from diminishing marginal utility of additional lifetime and discounting. For extreme states, Thai respondents may have a threshold or MET for the health state and lower or negative scores could be assigned to the additional years after this threshold. The resulting scores may be biased because the scores do not take into account diminishing marginal utility. Were the scores corrected for this bias, the score would be higher. Future study should take into account the weights attached to the utility function for future life years and correct the conventional TTO scores to achieve a proper reflection of the preferences of the Thai respondents.

6.3.5 The representativeness of the sample

The sample in this study does not perfectly represent the Thai general population. Females, adults aged 20-59 years, those with only primary education and respondents living in urban areas are slightly over-represented. This may result from the fieldwork management where the specific interview schedules were sent out to the field coordinators who were responsible for the invitation of respondents. Interviews were conducted in the daytime when male respondents were more likely to be unavailable because of work commitments. This also affects the representativeness of the respondent subgroup selected to estimate the Thai scores. The respondents from subgroup 3 with fewer than eleven inconsistencies were used in the estimation of preference scores. Compared with the general population, females, adults aged 20-59 years, respondents with primary education and respondents from urban areas are still over-represented. However, compared with all respondents in subgroup 1, the proportions of the respondents with secondary level and university education were higher and the proportion of those with primary level education was lower. This implies that more of the respondents with primary education level were dropped from subgroup 1. Therefore, it is likely that the preferences from the elderly respondents are less well represented in the Thai health state values.

The respondents participating in the qualitative interviews were also not representative of the Thai general population. The interviewees were chosen based on characteristics that made them more likely to generate a greater number of inconsistencies. If the qualitative interviews were conducted in different respondent groups, more could be learned from the respondents and the findings may be different from the findings presented in this study. It is possible that other groups of elderly respondents, especially those with higher educational attainment, may assign consistent scores because they could understand the descriptions of health states and the interview method. It is also possible that the adult respondents with a lower level of education attainment may also have higher number of inconsistencies.

6.3.6 Number of interviewers and the interview sites

The numbers of interviews per interviewer were not equal. One possible cause of this limitation is that a large proportion of the interviewers were recruited from Master degree and first-degree students who were not available throughout the period of the fieldwork. Some of them were also engaged in other research projects or their dissertations. Some interviewers had full-time hospital jobs. Therefore, interviewer availability depended on their day-jobs and their academic schedule in the universities. This was also a reason why additional interviewers were recruited in the middle of fieldwork to solve the problem of shortage of interviewers. The interviewers who conducted more interviews may have gained more experience and may have been able to conduct the interview more effectively. This would also be one possible factor concerning the extent of logical inconsistency. However, greater experience may not definitely guarantee fewer inconsistent responses, if the respondents they were assigned were elderly or with less education which generates a greater number of inconsistencies independently of the level of experience of the interviewer.

Although the interview sites were not recorded, it is likely that interview sites had some effect on the performance of the respondents in the interviews. To produce reliable results, it is likely that respondents need to pay greater attention and focus throughout the whole interview procedure, therefore, a peaceful environment plays a key role in assisting the respondents in assigning scores. It is less likely that respondents could concentrate on the tasks if they were interviewed in their households or workplaces, especially if their workplaces were shops and there were customers constantly coming by. The interviews were intermittently halted to allow the respondents to serve their customers. It is likely that the results from this group of respondents were significantly affected. However, this situation can hardly be avoided given that the interview schedules were planned in advance, and the interviews had to be conducted in a given period as planned, and as many as possible of the targeted respondents had to be interviewed. Some respondents could not leave their households to attend the arranged interview sites at the designated times because they had to take care of sick family members at home or they had to take care of their children as well as their household chores.

6.3.7 Number of observations per health state

As seen in Chapter 3 the numbers of observations per health state ranged from 95 to 1,313 and some health states have less than 200 observations per health state, this is one limitation in that two-hundred observations per health state cannot be achieved in some states given the constraints on budgets and human resources in the fieldwork survey. The achieved numbers of observations per health state are still adequate for use in estimating Thai health state values because the numbers are still greater than recommended by Williams (39). Another limitation is that the numbers of observations per health set were not equal for all sets. This resulted from the poor fieldwork management in that the health sets used in the interview were not closely monitored. For future studies, the number of health sets used in the interview should be checked more regularly to ensure that the number of observations per health state would not be much different from planned.

6.4 Priorities for future studies

What we have learned from this study can be used as a springboard to conduct future studies on estimating preferences scores for health in Thailand. Policy makers can be confident that preference elicitation interviews can be conducted successfully in a nationally representative sample and lend more support to research in this field. The issues that should be prioritised in future studies are as follows: to minimise the cognitive burden on the respondents; modelling health state values for further subgroups of respondents; to recruit a sample that is more representative of the Thai general population and to improve the fieldwork management; and the possibilities of using the new version of EQ-5D. Details are as follows.

6.4.1 To minimise the cognitive burden and logical inconsistency

There are two hypotheses that might explain the possible causes of logical inconsistency in this study and offer a potential strategy to minimise the level of logical inconsistency in future studies. The first hypothesis is that the cognitive overload resulting from participation in the interview causes the inconsistent responses. It is likely that the elderly respondents with primary level education may become exhausted relatively early in the interview, but would younger respondents with less literacy and numeracy abilities be exhausted faster than the elderly with high literacy and numeric abilities?

The respondents' understanding of the health states could be enhanced by using more plausible health states, combinations of health states that are more clearly differentiated and greater use of tools. As stated in the self-completed questionnaire, some respondents cannot identify the differences between different health state descriptions. Visual presentations, for instance, cartoons or graphic illustrations that can demonstrate the differences between degrees of severity among attributes in health states might be added to questionnaires. Visual representations as described in Hadorn *et al*. may be able to assist respondent comprehension(71). However, the graphic illustrations should be tested regarding whether the same illustrations can produce equivalent understanding across all groups of respondents. As suggested in the study by Hadorn *et al*., the response variance, the test-retest reliability and the numbers of counter-intuitive results should be examined before implementing the graphic illustration. Some pictures may be emotionally offensive for some respondents. A friendly computer-based interview program is going to be used in future studies. This will also help with the illustration of health states to the respondents. The computer-based interview program will be thoroughly examined regarding the feasibility and the reliability of the program before implementing in interviews. For those who may have difficulties or may unfamiliar with using a computer, an assistant will be provided to help.

The second hypothesis is that the logical inconsistency can be minimised if the respondents can "learn" how to respond to the tasks before the actual tasks begin. To familiarise the respondents with the interview tasks, an additional "warm-up" exercise for the TTO method should be given, for example, after completing the VAS method, three health states could be used to let the respondents practice and become familiar with the TTO questions. The scores used in the "warm-up" states would not be included in the analysis because it is less likely that the scores would represent the respondent's preferences over health states. One of the limitations of using the "warm-up" states for TTO is that the overall interview duration would be increased because the respondents are also required to give scores for the additional states in the warm-up session. This limitation could be solved by using fewer health states in the Ranking and VAS stages.

6.4.2 Modelling health state values for further subgroups of respondents

To further explore the implications of inconsistent responses on the model coefficients and the estimated scores, the respondents could be classified based on different criteria which may include a group of respondents with entirely consistent respondent. In future preference elicitation studies, the respondents should be classified into, for example: all respondents and one or more subgroups with various numbers of inconsistencies. If number of logically inconsistent responses could be reduced in future studies and a larger group of respondents with completely consistent responses could be obtained, the scores from this subgroup could also be generated. The choice of the numbers of inconsistencies used in the classification is arbitrary and based on the researcher’s judgements. The model estimated from the scores assigned by all respondents could be compared with that estimated from the scores of completely consistent respondents. Additional models estimated from the scores of the respondents with various numbers of inconsistencies provide more information on how the models change along this continuum.

There would be, at least, three areas of improvement in the model specifications: alternative modelling methods; new relevant independent variables; and a new transformation of the scores for states worse than death. More effort should be paid to the development of better models to estimate Thai preference scores. Various modelling forms, for example, a multiplicative model such as used in the preference scores estimation for the Health Utility Index (HUI) measures, may be an alternative model to consider apart from the current additive model(72, 73). Further relevant independent variables, for example, the independent variables used in the estimation of the Japanese model, could be added to explain the interactions between the dimensions and the large differences between the actual and estimated models of the states with level 1 in both mobility and self-care dimensions or with level 3 in both dimensions(48).

Another interesting future study would be to correct the Thai preference scores in which the assumptions of the TTO may be violated especially with respect to maximal endurable of time and diminishing marginal utility. The methods used in the study by Attema & Brower can be adapted to address the latter issue to find a potential strategy to “correct” the Thai preference scores (74). The new TTO question format as suggested by Robinson & Spencer and Devlin *et* *al.* can be used to elicit the scores for states worse than death to place the negative scores on the same scale as the states better than death (75, 76).

6.4.3 To recruit a sample that is more representative of the Thai general population and to improve the fieldwork management

Female respondents tended to be over-represented in the sample. This may result from an inappropriate interview schedule and future studies should take this into account. One option would be to plan to interview a greater number of male and elderly respondents. To encourage respondents to participate in the interview, greater efforts should be paid to the strategy used to identify, inform and make the appointments with the respondents and a flexible interview schedule should be provided. It is likely that if respondents are provided with choices of interview times so that they can choose on the basis of their availability, then a greater number of respondents might be willing and able to participate in the interview.

Respondents should be informed regarding the research objectives and the credibility of the research team. In the present study, it was the field coordinators who contacted and informed the respondents regarding the research. It is likely that the field coordinators, given the workload of their day jobs, may not be able to clearly communicate with the respondents. An official letter should be sent directly to respondents from the research team with a range of interview times so she can choose to suit her schedule. Different strategies should be used to locate the respondents in urban and rural areas. In rural areas, an active and knowledgeable field coordinator is required to locate the respondents who do not reply to the formal letters from the research team and to remind the respondents regarding the interview schedule nearer the time. For the respondents in urban areas, including Bangkok, the respondents who can easily commute to the researchers' office may be invited to be interviewed in the office and their transportation costs covered.

Fieldwork management should be improved to increase the proportion of successful interviews. If the sample of the respondents is going to be drawn from respondents being interviewed in an NSO survey, early communication regarding the merging of data on respondent characteristics should be planned in the beginning of the research project. If the merging of databases cannot be done because of the confidentiality policy, researchers need to be informed and be prepared to collect respondent characteristics in their own survey. This has the cost of prolonging the interview but longer interviews may have to be accepted in order to secure a complete database. Also, using the same groups of the respondents may make them irritated and annoyed because participation in more than one survey may result in questions being repeated.

If a face-to-face interview conducted by an interviewer is going to be used in the next study, full-time interviewers are needed who could work with a flexible interview schedule. Interviewer training should involve more sessions for practicing the interview. Review of the interview process should be performed regularly. The number of interviews per interviewers should be similar to decrease workloads on the interviewers and to give more opportunities to gain experience for some interviewers who would otherwise have less experience because of the fewer number of interviews. Interview settings should still be arranged in the neighbourhood with, if possible, the minimum level of distraction. The best interview time would be when respondents are least likely to be distracted, for example, at the weekend or in the evening after the respondents complete the household chores.

The classification of health states into mild, moderate and severe groups can be obtained in a different way. Rather than using the 5-figure of EQ-5D to categorise health states, the estimated preference scores from this study could be used as a guide to categorise the severity of health states used in future studies. For example, severe states could be those states with negative scores (68 states) and mild states the states with the scores higher than 0.550 (21 states). The states with the scores higher than 0.556 are those states without level 3 in any dimensions. The remaining states are classified as moderate states. The chosen states should be plausible.

6.4.4 Use of the new version of EQ-5D measure

A greater number of health states can be identified using the new version of EQ-5D. It is expected that the EuroQoL group is going to launch the new version of EQ-5D (EQ-5D-5L). One of the key topics of interest for the 26th EuroQol Group Scientific Meeting in September 2009 is the research on the five level descriptive system(77). Launching the new EQ-5D version could be used as a good opportunity to conduct a new translation of the new EQ-5D health states. Translation of health states into Thai should be done taking into account the semantic and conceptual equivalence of the health state descriptions. Some activities described in the EQ-5D-5L may need to be changed for the Thai context. The utilisation of the new EQ-5D is promising and has successfully drawn support from the potential funders. The estimation of preference scores for the EQ-5D-5L health states is going to be implemented in a proposal supported by the International Health Policy Program (IHPP), Burden of Disease project (BOD) and the Health Impact and Technology Assessment Program (HITAP), Ministry of

Public Health, Thailand. The objectives include the translation of EQ-5D-5L into Thai, qualitative studies on how Thais respond to the elicitation interview following from what has been learnt from this study.

There could be more challenges awaiting in the elicitation of preference scores for the new version of the EQ-5D. Rather than 243 states as in the current version, there will be 3,125 (5^5) states described by the new system. The next interesting question would be how many states should be valued directly, given that the Thai respondents may not each be able to give scores for more than 10-11 states, how many respondents are needed in the sample. More levels of health descriptions mean greater complex of the health states. Will the Thai respondents be able to cope with the greater cognitive workloads using the new version? It is possible that the number of inconsistent responses could be greater than in the present study, more attention should be given to the treatment of the inconsistent responses.

The Thai preference scores estimated for the present version of EQ-5D can be used to guide the categorisation of health states that could be used in future preference studies. For example, it has been learned in this study that Thai respondents assigned higher weights to the mobility and self-care dimensions. The health states generated by the new version of EQ-5D with, for example, the most severe problems in these two dimensions, could be assumed to be the "severe" health states, compared with those states with mild problems in the mobility and self-care dimensions.

6.5 Conclusion

Results of economic evaluations can be used to aid decision making on the allocation of scarce resources across different health interventions. The key contribution of this research is the estimation of Thai population-based preference scores to be used in estimating QALYs as one measure of health outcome in economic evaluation in Thai settings. Logical inconsistency is also explored in this study using both quantitative analysis and qualitative interviews. It is not only the respondent demographic characteristics that influence the number of inconsistent responses, the strategies employed by the respondents when asked to state their preferences may be responsible for the inconsistent responses as well. It is assumed in this study that the highly inconsistent respondents are unable to understand the preference elicitation tasks and their stated scores may not be suitable to represent their preferences on health, thus these scores were excluded from the model specifications. This is justified because it was unlikely that the highly inconsistent respondents would be able to assign the scores given the complicated task as the TTO method. The disadvantage of the exclusion of respondents is that the scores from the elderly could be under-represented. Logical inconsistency could be minimised in future studies by adopting a number of strategies, for example, reducing the number of health states used in the interview, choosing more plausible health states, or using a computer-based preference interview.

This study demonstrates that preferences about health can be successfully estimated from the Thai general population. Using the scores elicited from other countries to represent Thai preferences in an economic evaluation could produce misleading results. Thai preference scores differ from those of other countries but are quite similar to the UK scores. Future studies can then aim to elicit preference scores for different health description systems both generic health outcome measures and condition-specific health outcome measures. Close collaborations between the several organisations involved in the research are a key factor for the successful conduct of the fieldwork. The Thai model still suffers from heteroskedasticity and some errors are identified in the estimated scores. The new version of the EQ-5D which will be launched in the near future may be used to provide the opportunity for a systematic translation of health state descriptions with more of semantic and conceptual equivalence to the understanding of the Thai population. The issues of the ceiling and floor effects of the scores as well as performance of the model could be improved. Problems may also arise because a greater number of health states being valued may lead to an increase in logical inconsistency.

References

1. McPake B, Kumaranayake L, Normand C. Health Economics: An International Perspective. London: Routkedge, 2002.
2. Gold M, Siegel J, Russell L, et al. Cost-Effectiveness in Health and Medicine. Oxford: Oxford University Press, 1996.
3. Drummond M, Sculpher M, Torrance G, et al. Methods for the Economic Evaluation of Health Care Programmes. 3rd ed. Oxford: Oxford University Press, 2005.
4. Morris S, Devlin N, Parkin D. The Use of Economic Evaluation in Decision Making. In: Morris S, Devlin N, Parkin D, eds., Economic Analysis in Health Care. Glasgow: John Wiley & Sons, Ltd, 2007.
5. Rutten F. Economic evaluation and health care decision-making. Health Policy. 1996; 36: 215-29.
6. Johannesson M. Economic evaluation of health care and policy making. Health Policy. 1995; 33: 179-90.
7. Chiawchanwattana A, Limwattananon C, Limwattananon S, et al. Cost utility of renal dialysis in Thailand Journal of The Nephrology Society of Thailand. 2003; 9: 158-69.
8. Teerawattananon Y. Cost-effectiveness and cost-utility analysis of renal replacement therapy in Thailand. In: Tangcharoensathien V, ed., Universal Access to Renal Replacement Therapy in Thailand: A policy analysis. Bangkok, Thailand: International Health Policy Program and Nephrology Society in Thailand, 2005.
9. Teerawattananon Y, Mugford M. Is it worth offering a routine laparoscopic cholecystectomy in developing countries? A Thailand case study. Cost effectiveness and Resource allocation. 2005; 3.
10. Teerawattananon Y, Russell S, Mugford M. A Sytematic Review of Economic Evaluation in Thailand: Are the Data Good Enough to be Used by Policy-Makers? Pharmacoeconomics. 2007; 25: 467-79.
11. Wibulpolprasert S. The Need for Guidelines and the Use of Economic Evidence in Decision-Making in Thailand: Lessons Learnt from the Development of the National List of Essential Drugs. Journal of Medical Association of Thailand. 2008; 91: S1-S3.
12. Teerawattananon Y, Tantivess S, Yothasamut J, et al. Historical development of health techonology assessment in Thailand. International Journal of Technology Assessment in Health Care. 2009; 25: 1-12.
13. Dolan P. Output measures and valuation in health. In: Drummond M, McGuire A, eds., Economic Evaluation in Health Care: Merging thoery into practice. Oxford: Oxford University Press, 2001.
14. Drummond M. Evaluation of Health Technology: Economic Issues for Health Policy and Policy Issues for Economic Appraisal. Social Science and Medicine. 1994; 38: 1593-600.
15. Badia X, Roset M, Heardman M, et al. A Comparison of United Kingdom and Spanish General Population Time Trade-off Values for EQ-5D Health states. Medical Decision Making. 2001; 21: 7-16.
16. NICE. Guide to the Methods of Technology Appraisal. London, UK, 2008.
17. Suksiriserekul S. The cost-utility analysis of some Thai public health programmes. Department of Economics and Related Studies. York: University of York, 1994.
18. EuroQol. EuroQol - a new facility of health-related quality of life. Health Policy. 1990; 16: 199-208.
19. Brooks R. EuroQol: the current state of play. Health Policy. 1996; 37: 53-72.
20. Konig H-H, Bernert S, Angermeyer M, et al. Comparison of Population Health Status in Six European Countries: Results of a Representative Survey Using the EQ-5D Questionnaire. Medical Care. 2009; 47: 255-61.

21. Kind P, Brooks R, Rabin R. EQ-5D concepts and methods: a development history. Netherlands: Springer, 2005.
22. Szende A, Oppe M, Devlin N. EQ-5D value sets: inventory, comparative review and user guide. London: Springer, 2007.
23. Froberg D, Kane R. Methodology for measuring health-state preferences-I: Measurement strategies. Journal of Clinical Epidemiology. 1989a; 42: 345-54.
24. Brazier J, Deverill M, Green C, et al. A Review of the Use of Health Status Measures in Economic Evaluation. Health Technology Assessment. 1999; 3.
25. Savoia E, Fantini MP, Pandolfi PP, et al. Assessing the construct validity of the Italian version of the EQ-5D: preliminary results from a cross-sectional study in North Italy. Health and Quality of Life Outcomes. 2006; 4.
26. Misajon R, Pallant J, Manderson L, et al. Measuring the impact of health problems among adults with limited mobility in Thailand: further validation of the Perceived Impact of Problem Profile. Health and Quality of Life Outcomes. 2008; 6.
27. Sakthong P, Charoenvisuthiwongs R, Shabunthom R. A Comparison of EQ-5D index scores using the UK, US and Japan Preference weights in a Thai sample with type 2 diabetes. Health and Quality of Life Outcomes. 2008; 6.
28. Rasanen P, Roine E, Sintonen H, et al. Use of quality-adjusted life years for the estimation of effectiveness of health care: A systematic literature review. International Journal for Quality in Health Care. 2006; 22: 235-41.
29. Brauer C, Rosen A, Greenberg D, et al. Trends in the Measurement of Health Utilities in Published Cost-Utility Analyses. Value in Health. 2006; 9: 213-18.
30. Richardson G, Manca A. Calculation of quality adjusted life years in the published literature: a review of methodology and transparency. Health Economics. 2004; 13: 1203-10.
31. Scuffham P, Whitty J, Mitchell A, et al. The Use of QALY Weights for QALY Calculations: A Review of Industry Submissions Requesting Listing on the Australian Pharmaceutical Benefits Scheme 2002-2004. Pharmacoeconomics. 2008; 26: 297-310.
32. Stein K, Fry A, Round A, et al. What Value Health? Applied Health Economics and Health Policy. 2005; 4: 219-28.
33. Sakthong P. Measurement of Clinical Effect: Utility. Journal of Medical Association of Thailand. 2008; 91: S43-S52.
34. Lim LL-Y, Seubsman S-a, Sleigh A. Thai SF-36 health survey: tests of data quality, scaling assumptions, reliability and validity in healthy men and women. Health and Quality of Life Outcomes. 2008; 6.
35. Leelakulthanit O. Quality of life in Thailand. Social Indicators Research. 1992; 27: 41-57.
36. Fox-Rusby J. First steps to assessing semantic equivalence of the EQ-5D. Paper presented at the 13th Plenary Meeting of the EuroQoL Group. Oslo, Norway, 1996.
37. Hunt S, Alonso J, Bucquet D, et al. Cross-cultural adaptation of health measures. Health Policy. 1991; 19: 33-44.
38. Cheung YB, Thumboo J. Developing Health-Related Quality-of-Life Instruments for use in Asia. Pharmacoeconomics. 2006; 24: 643-50.
39. Williams A. The Measurement and Valuation of Health: A Chronicle. York: Centre for Health Economics, The University of York, 1995.
40. Dolan P, Gudex C, Kind P, et al. The Time Trade-Off Method: Results From a General Population Study. Health Economics. 1996; 5: 141-54.
41. Gudex C, Dolan P, Kind P, et al. Valuing health states, Interviews with the general public. European Journal of Public Health. 1997; 7: 441-48.
42. Gudex C, Dolan P. Valuing Health States: The Effect of Duration. : Centre for Health Economics, The University of York, 1995.

43. Dolan P. Modeling Valuations for EuroQoL Health States. Medical Care. 1997; 35: 1095-108.
44. Dolan P, Roberts J. Modelling Valuations for Eq-5d Health States: An Alternative Model Using Differences in Valuations. Medical Care. 2002; 40: 442-46.
45. Murti B, Johnson J, Ohinmaa A, et al. Comparison of Finnish-and US-Based VAS Valuations of the EQ-5D. The 14th Plenary Meeting of the EuroQol group. 1997.
46. Johnson J, Coons S, Ergo A, et al. Valuation of EuroQol(EQ-5D) Health States in an Adult US Sample. Pharmacoeconomics. 1998; 13: 421-33.
47. Rupel V, Rebolj M. The Slovenian VAS Tariff Based on Valuations of EQ-5D Health States From the General Population. The 17th Plenary Meeting of the EuroQoL Group. Pamplona, Spain, 2000.
48. Tsuchiya A, Ikeda S, Ikegami N, et al. Estimating and EQ-5D population value set: the case of Japan. Health Economics. 2002; 11: 341-53.
49. Devlin N, Hansen P, Kind P, et al. Logical inconsistencies in survey respondents' health state valuations - a methodological challenge for estimating social tariffs. Health Economics. 2003; 12: 529-44.
50. Jelsma J, Hansen K, Weerdt W, et al. How do Zimbabweans value health states? Population Health Metrics. 2003; 1.
51. Shaw J, Johnson J, Coons S. US Valuation of the EQ-5D Health States, Development and Testing of the D1 Valuation Model. Medical Care. 2005; 43: 203-20.
52. Greiner W, Claes C, Busschbach J, et al. Validating the EQ-5D with time trade-off for the German population. European Journal of Health Economics. 2005; 6: 124-30.
53. Lamers L, McKonnell J, Krabbe P, et al. The Dutch tariff: results and arguments for an effective design for national EQ-5D valuation studies. Health Economics. 2006; 15: 1121-32.
54. Zarate V, Kind P, Chuang L-H. Hispanic Valuation of the EQ-5D Health States: A Social Value Set for Latin Americas. Value in Health. 2008; 11: 1170-77.
55. Jo M-W, Yun S-C, Lee S-I. Estimating Quality Weights for EQ-5D Health States with the Time Trade-Off Method in South Korea Value in Health. 2008; 11: 1186-89.
56. Kok E, Stolk E, Busschbach Jv. Influences of the number of health states on time trade-off. The 17th Plenary meeting of the EuroQol Group. Pampalona, Spain, 2000.
57. NSO. The 2007 Health and Welfare Survey. Bangkok, Thailand: National Statistical Office, Thailand, 2007.
58. Pliskin J, Shepard D, Weinstein M. Utility Functions for Life Years and Health Status. Operations Research. 1980; 28: 206-24.
59. Lamers L. The Transformation of Utilities for Health States Worse Than Death: Consequences for the Estimation of EQ-5D Value Sets. Medical Care. 2007; 45: 2328-244.
60. Sarkadi K. The consistency of the Shapiro-Francia test. Biometrika. 1975; 62: 445-50.
61. Cairns J, Pol vd, Lloyd A. Decision Making Heuristics and the Elicitation of Preferences: Being Fast and Frugal about the Future. Health Economics. 2002; 11: 655-58.
62. Dolan P, Kind P. Inconsistency and Health State Valuations. Social Science and Medicine. 1996; 42: 609-15.
63. Badia X, Roset M, Herdman M. Inconsistent responses in three preference-elicitation methods for health states. Social Science and Medicine. 1999; 49: 943-50.
64. Torrance G, Thomas W, Sackett D. A Utility Maximization Model for Evaluation of Health Care Programs. Health Services Research. 1972: 118-33.
65. Dolan P, Roberts J. Brief Report, Modelling Valuations for EQ-5D Health States, An Alternative Model Using Differences in Valuations. Medical Care. 2002; 40: 442-46.
66. Wittenberg E, Halpern E, Divi N, et al. The effect of age, race and gender on preferences scores for hypothetical health states. Quality of life research. 2006; 15: 645-53.
67. Green W. Econometric Analysis. 6th edition ed. New Jersey: Prentice Hall, 2008.
68. Hsiao C. Analysis of Panel Data. New York: Cambridge University Press, 1986.

69. Crosby R, Kolotkin R, Williams G. Defining clinically meaningful change in health-related quality of life. Journal of Clinical Epidemiology. 2003; 56: 395-407.
70. Brazier J, Akehurst R, Brennan A, et al. Should Patients Have a Greater Role in Valuing Health States? Applied Health Economics and Health Policy. 2005; 4: 201-08.
71. Hadorn D, Hays R, Uebersax J, et al. Improving task comprehension in the measurement of health state preferences. Journal of Clinical Epidemiology. 1992; 45: 233-43.
72. Torrance G, Feeny D, Furlong W, et al. Multiattribute Utility Function for a Comprehensive Health Status Classficiation System: Health Utilities Index Mark 2. Medical Care. 1996; 34: 702-22.
73. Feeny D, Furlong W, Torrance G, et al. Multiattribute and Single-Attribute Utility Functions for the Health Utility Index Mark 3 System. Medical Care. 2002; 40: 113-28.
74. Attema A, Brouwer W. The correction of TTO-scores for utility curvature using a risk-free utility elicitation method. Journal of Health Economics. 2009; 28: 234-43.
75. Robinson A, Spencer A. Exploring challenges to TTO utilities: valuing states worse than dead. Health Economics. 2006; 15: 393-402.
76. Devlin N, Tsuchiya A, Buckingham K, et al. A uniform Time Trade Off method for states better and worse than dead: feasibility study of the "lead time" approach. Department of Economics, Shool of Social Sciences, City University, London, 2009.
77. Devlin N, Busschbach J. The 26th EuroQol Group Scientific Plenary. Available from: http://wwweuroqolorg/news/news/article/EuroQol_Plenary_2009html [Accessed 29 April 2009], .

Appendix 1 Lists of the villages recruited in the fieldwork survey

North

ภาค	จังหวัด	อำเภอ	ตำบล	หมู่บ้าน
ภาคเหนือ	**ลำปาง**	เมืองลำปาง	หัวเวียง , กล้วยแพะ , บ้านเสด็จ , นิคมพัฒนา	เสด็จ, ร้อง, ปงกา, นิคมกิ่วลมเขต 6 , วังทอง
		เสริมงาม	ทุ่งงาม, เสริมซ้าย	นางอย
		เกาะคา	ลำปางหลวง	ป่าเหียง
		วังเหนือ	วังทอง	ปงทอง
		ห้างฉัตร	เวียงตาล	สันทรายเหนือ
		เถิน	เถินบุรี	สองแควตะวันออก
	พะเยา	จุน	ห้วยข้าวก่ำ, หงส์หิน	ศรีจอมแจ้ง
		เมืองพะเยา	ท่าวังทอง	ท่าจำบอน
		เชียงคำ	ทุ่งผาสุข	ไร่แสนสุข
		แม่ใจ	บ้านเหล่า	เหล่าพัฒนา
		ดอกคำใต้	สันโค้ง	ใหม่ราษฎร์บำรุง
	พิษณุโลก	เมืองพิษณุโลก	ในเมือง , ท่าโพธิ์ , บ้านกร่าง	ยาง,ยางเอน , กร่าง
		นครไทย	หนองกะท้าว	โนนนากาน,โนนนาก่าม
		พรหมพิราม	พรหมพิราม , ดงประคำ	ห้วยตั้ง , โคกสมอ
		เนินมะปราง	ชมพู	หนองหญ้าปล้อง
		ชาติตระการ	ชาติตระการ	ชาติตระการ
		วังทอง	บ้านกลาง	หินปลากลาย,หินประกาย

South

ภาค	จังหวัด	อำเภอ	ตำบล	หมู่บ้าน
ภาคใต้	**นครศรีธรรมราช**	เมืองนครศรีธรรมราช	คลัง	
		ชะอวด	ชะอวด , นางหลง	บ้านนางหลง,เนินทราย
		พรหมคีรี	พรหมโลก	ปลายอวน,นอกท่า
		ทุ่งสง	นาไม้ไผ่	วังยาง
		หัวไทร	หัวไทร	ลำคลอง
		พิปูน	เขาพระ	ทุ่งร่อน
		ทุ่งใหญ่	ปริก	ควนลำภู
	ตรัง	เมืองตรัง	ทับเที่ยง	
		กันตัง	นางเป้า	ตลาดใหม่
		สิเกา	บ่อหิน	ดุหุน
		นาโยง	โคกสะบ้า	เกาะหยี,ใสขัน
		ย่านตาขาว	หนองบ่อ	หนักแบกใต้
	ชุมพร	เมืองชุมพร	ปากน้ำ , บางหมาก	ดอนรวบ
		ท่าแซะ	หินแก้ว	ในโสม
		เมืองชุมพร	บ้านนา	ท่ามะปริง

Central

ภาค	จังหวัด	อำเภอ	ตำบล	หมู่บ้าน
ภาคกลาง	**ประจวบคีรีขันธ์**	เมือง	ประจวบคีรีขันธ์ อ่าวน้อย ห้วยทราย	หมู่ 3 ทุ่งยายเอม ลานไทร
		บางสะพาน	ร่อนทอง ทองมงคล	ห้วยเกรียบ
		หัวหิน	หนองแก	
		ทับสะแก	อ่างทอง	ทุ่งพุฒ
	ชัยนาท	เมืองชัยนาท	ในเมือง , ชัยนาท	ดอนตาไล้
		สรรคบุรี	ห้วยกรด	บางยายอ้น,ท่ากร่าง
		มโนรมย์	ท่าฉนวน	หลั่น
		หันคา	บ้านเชี่ยน	พัฒนา,ทับนา
	จันทบุรี	เมืองจันทบุรี	เกาะขวาง , วัดใหม่ , พลับพลา	เกาะตะเคียน , โป่งแรด
		ท่าใหม่	รำพัน	โป่งมะม่วงหวาน
		สอยดาว	ทับช้าง	สะพานหก
	สุพรรณบุรี	เมือง	ดอกกำยาน	ศาลาแดง
		ศรีประจันต์	วังยาง	ทุ่งชนะ
		เดิมบางนางบวช	ปากน้ำ	ท่าทอง
		อู่ทอง	หนองโอ่ง	หนองหลุม
		สามชุก	หนอกผักนาก	หนองไผ่

Northeast

ภาค	จังหวัด	อำเภอ	ตำบล	หมู่บ้าน
ภาค ตะวันออกเฉียงเหนือ	กาฬสินธุ์	เมืองกาฬสินธุ์	กาฬสินธุ์ , โพนทอง	โพนทอง
		เขาวง	คุ้มเก่า	
		ร่องคำ	สามัคคี	นาเรียง
		ห้วยเม็ก	พิมูล	พิมูล
		กิ่งอ.สามชัย	คำสร้างเที่ยง	คำสร้างเที่ยง
		กุฉินารายณ์	กุดค้าว	ทุ่งคลองไผ่
		คำม่วง	เนินยาง	สูงเนิน
		กิ่งอ.ฆ้องชัย	ลำชี	วังยาง
	ขอนแก่น	เมืองขอนแก่น	ในเมือง , บ้านทุ่ม , ศิลา , หนองตูม ,	แดงน้อย , หนองไผ่ , โคกท่า
		ชุมแพ	หนองไผ่	
		เขาสวนกวาง	เขาสวนกวาง	
		หนองเรือ	บ้านเม็ง	สว่างดอนดู่
		บ้านไผ่	หินตั้ง	โคกก่อง
		ภูผาม่าน	วังสวาบ	วังสวาบ
		สีชมพู	ศรีสุข	ลอมไผ่
		แวงน้อย	ท่าวัด	หนองโกน้อย
		น้ำพอง	กุดน้ำใส	กุดกว้าง
		ภูเวียง	ภูเวียง	หนองหญ้าปล้อง
		กระนวน	หนองโก	ศรีสมบูรณ์
		มัญจาคีรี	คำแคน	คำน้อย
	ร้อยเอ็ด	เมืองร้อยเอ็ด	ในเมือง	
		เกษตรวิสัย	เหล่าหลวง , น้ำอ้อม	หนองสังข์ , น้ำอ้อม
		เสลภูมิ	ภูเงิน	หนองกุง
		จตุรพักรพิมาน	ศรีโคตร	อีโคตร
		อาจสามารถ	อาจสามารถ	ใหม่พัฒนา
		โพนทอง	อุ่มเม่า	กระพี้
		กิ่งอ.เชียงขวัญ	บ้านเขือง	หัวแดง,หัวดง
		โพธิ์ชัย	อัคคะคำ	อัคคะ
		กิ่งอ.ทุ่งเขาหลวง	เหล่า	ช่องแมว,ซองแมว
		สุวรรณภูมิ	สระคู	หนองหว้า
		ธวัชบุรี	เมืองน้อย	มะยาง
	มหาสารคาม	เมืองมหาสารคาม	ท่าสองคอน , ตลาด	โนนแต้
		นาเชือก	หนองแดง	วังหิน
		โกสุมพิสัย	โพนงาม	โพนสว่าง
		วาปีปทุม	เสือโก้ก	บูรพาประชาราษฎร์
		เชียงยืน	นาทอง	แบก
		ยางสีสุราช	ดงเมือง	หนองงูเหลือม
		บรบือ	บ่อใหญ่	หนองทุ่ม,บ่อใหญ่

Appendix 2 Recording form

Name ☐Mr. ☐Mrs. ☐Miss ____________________

Corresponding address No. _____ Street____________

City______________ Ampur________________

Province____________________ Postcode _______

Telephon no. __________________

Age______years

Marital status ☐ Marriage ☐ Single ☐ Divorce ☐ Widow

No. of children__________

REG ☐

CWT ☐☐☐

AMP ☐☐

TMB ☐

AREA ☐

E.D. ☐☐☐

BLK/VIL ☐

PSU_NO. ☐☐☐

HH_NO. ☐☐

Date of interview __ __ / __ __ / 2007

Time start __ __ . __ __

Health set No.

O	1	O	5	O	9
O	2	O	6	O	10
O	3	O	7	O	11
O	4	O	8	O	12

Interviewer's name ____

Time end __ __ . __ __

Overall interview duration _____ _____ mins

Part 1 The Thai EQ-5D questionnaire

กรุณาทำเครื่องหมาย X ในช่องสี่เหลี่ยม (☐) ของคำถามแต่ละข้อที่ตรงกับสภาวะสุขภาพของท่านในวันนี้มากที่สุด

1. **การเคลื่อนไหว**

ข้าพเจ้าไม่มีปัญหาในการเดิน ☐

ข้าพเจ้ามีปัญหาในการเดินบ้าง ☐

ข้าพเจ้าไม่สามารถไปไหนได้ และจำเป็นต้องอยู่บนเตียง ☐

2. การดูแลตนเอง

ข้าพเจ้าไม่มีปัญหาในการดูแลตนเอง ☐

ข้าพเจ้ามีปัญหาในการอาบน้ำหรือแต่งตัวบ้าง ☐

ข้าพเจ้าไม่สามารถอาบน้ำหรือแต่งตัวด้วยตนเองได้ ☐

3. กิจกรรมที่ทำเป็นประจำ (เช่น การทำงาน การเรียนหนังสือ การทำงานบ้าน การทำกิจกรรมในครอบครัว หรือการทำกิจกรรมยามว่าง)

ข้าพเจ้าไม่มีปัญหาในการทำกิจกรรมที่ทำเป็นประจำ ☐

ข้าพเจ้ามีปัญหาในการทำกิจกรรมที่ทำเป็นประจำอยู่บ้าง ☐

ข้าพเจ้าไม่สามารถทำกิจกรรมที่ทำเป็นประจำได้ ☐

4. ความเจ็บปวด ไม่สุขสบาย

ข้าพเจ้าไม่มีอาการเจ็บปวดหรืออาการไม่สุขสบาย ☐

ข้าพเจ้ามีอาการเจ็บปวดหรืออาการไม่สุขสบายปานกลาง ☐

ข้าพเจ้ามีอาการเจ็บปวดหรืออาการไม่สุขสบายมากที่สุด ☐

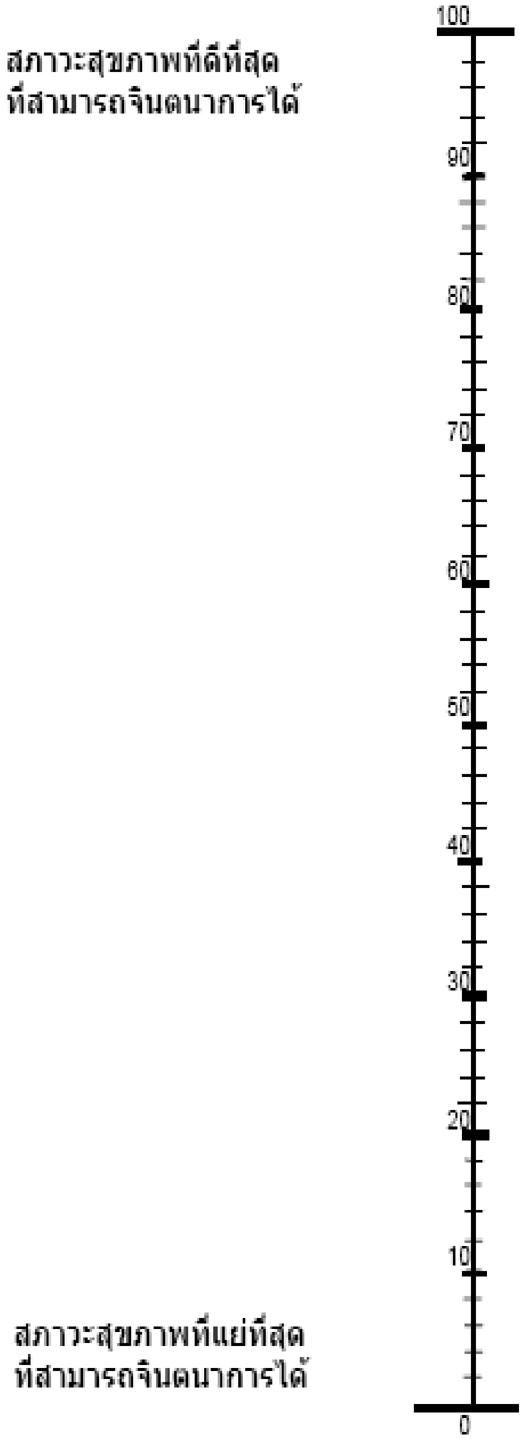
5. ความวิตกกังวล ซึมเศร้า

ข้าพเจ้าไม่รู้สึกวิตกกังวลหรือซึมเศร้า ☐

ข้าพเจ้ารู้สึกวิตกกังวลหรือซึมเศร้าปานกลาง ☐

ข้าพเจ้ารู้สึกวิตกกังวลหรือซึมเศร้ามากที่สุด ☐

Part 2 Thermometer scale “Own health state”



Part 3 Health state ranking

Time start   __ __ . __ __

Best imaginable health state

__ __   __ __

__ __   __ __

__ __   __ __

__ __   __ __

__ __   __ __

__ __   __ __

__ __   __ __

__ __   __ __

__ __   __ __

__ __   __ __

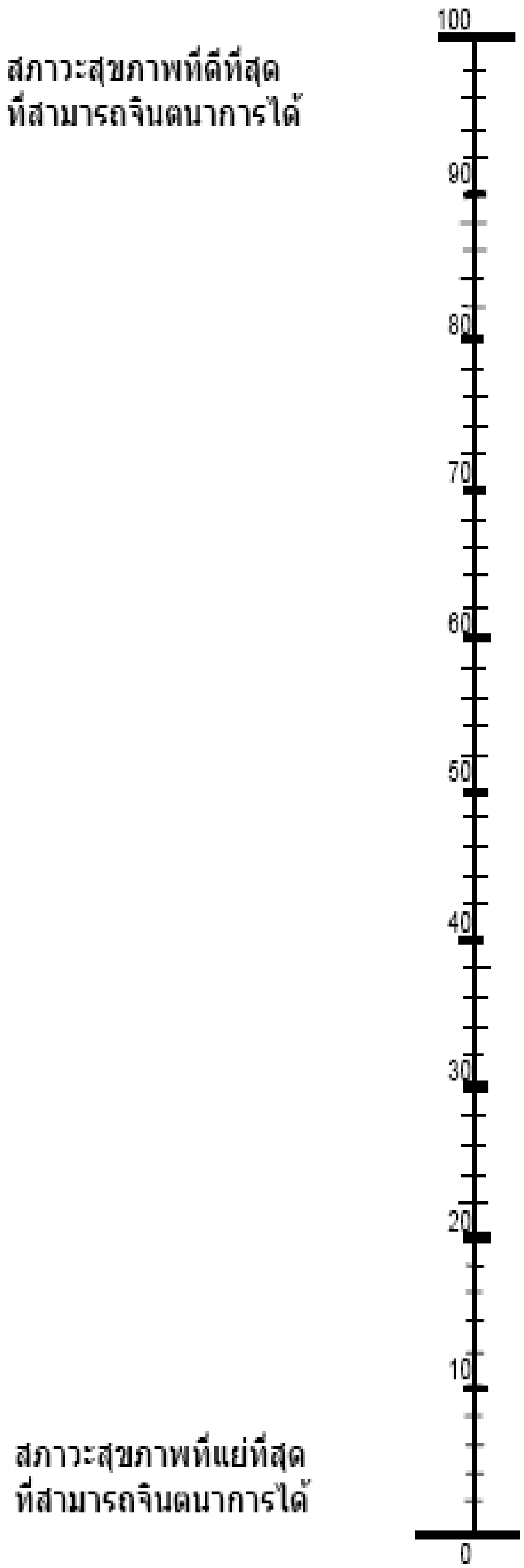
__ __   __ __

Worst imaginable health state

Time finish __ __ . __ __

Part 4 Thermometer scale for 11 health states

Time start __ __;__ __



Time finish __ __:__ __

Part 5 Time trade-off

Time start __ __ . __ __

Health state 1

X =

_ _

- o Better than death
- o Worse than death

0 1 2 3 4 5 6 7 8 9 10 จำนวนปี

Health state 2

_ _

X =

- o Better than death
- o Worse than death

0 1 2 3 4 5 6 7 8 9 10 จำนวนปี

Health state 3

_ _

X =

- o Better than death
- o Worse than death

0 1 2 3 4 5 6 7 8 9 10 จำนวนปี

Health state 4

X =

_ _

- o Better than death
- o Worse than death

0 1 2 3 4 5 6 7 8 9 10 จำนวนปี

Health state 5

_ _

X = ☐

- o Better than death
- o Worse than death

0 1 2 3 4 5 6 7 8 9 10 จำนวนปี

Health state 6

_ _

X = ☐

- o Better than death
- o Worse than death

0 1 2 3 4 5 6 7 8 9 10 จำนวนปี

Health state 7

X = ☐

_ _

- o Better than death
- o Worse than death

0 1 2 3 4 5 6 7 8 9 10 จำนวนปี

Health state 8

_ _

X = ☐

- o Better than death
- o Worse than death

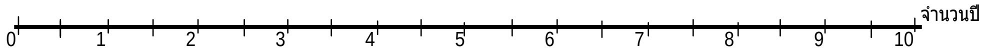
0 1 2 3 4 5 6 7 8 9 10 จำนวนปี

Health state 9

X =

– –

- o Better than death
- o Worse than death

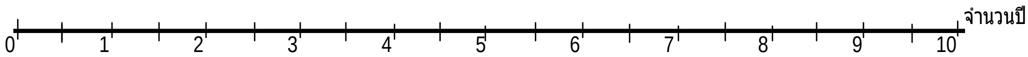


Health state 10

X =

– –

- o Better than death
- o Worse than death



Respondent's comments

1. Which part is the most difficult interview for you?

   - Ranking 11 health states
   - Giving scores from 0-100 to all health states
   - Sacrificing time of life to live in better health

2. In your opinion, what cause the difficulties in the interview?

   - I cannot imagine myself living in the states shown in the cards
   - I cannot understand the trading off time method
   - Others (please specify) ____________________

   ____

Interviewer's comments

In your opinion, how well the respondent understand the interview procedure?

- The respondent is confident participating in the interview
- The respondent became confident after the initial 2-3 questions
- The respondent is not confident throughout the interview procedure

Additional comments:

____

____

Appendix 3 **Example of health card for state 11111 (PH) and 11112 (LE)**

ไม่มีปัญหาในการเดิน

ไม่มีปัญหาในการดูแลตนเอง

ไม่มีปัญหาในการทำกิจกรรมที่ทำเป็นประจำ

ไม่มีอาการเจ็บปวดหรืออาการไม่สุขสบาย

ไม่รู้สึกวิตกกังวลหรือซึมเศร้า PH

ไม่มีปัญหาในการเดิน

ไม่มีปัญหาในการดูแลตนเอง

ไม่มีปัญหาในการทำกิจกรรมที่ทำเป็นประจำ

ไม่มีอาการเจ็บปวดหรืออาการไม่สุขสบาย

รู้สึกวิตกกังวลหรือซึมเศร้าปานกลาง LE

Appendix 4 Thai preference scores for EQ-5D health states

EQ-5D state	Score	EQ-5D state	Score	EQ-5D state	Score
1 1 1 1 1	1.000	1 2 2 2 2	0.513	1 3 3 3 3	-0.022
1 1 1 1 2	0.766	1 2 2 2 3	0.295	2 1 1 1 1	0.677
1 1 1 1 3	0.548	1 2 2 3 1	0.269	2 1 1 1 2	0.645
1 1 1 2 1	0.726	1 2 2 3 2	0.237	2 1 1 1 3	0.427
1 1 1 2 2	0.693	1 2 2 3 3	0.158	2 1 1 2 1	0.605
1 1 1 2 3	0.475	1 2 3 1 1	0.419	2 1 1 2 2	0.573
1 1 1 3 1	0.449	1 2 3 1 2	0.387	2 1 1 2 3	0.355
1 1 1 3 2	0.417	1 2 3 1 3	0.309	2 1 1 3 1	0.328
1 1 1 3 3	0.338	1 2 3 2 1	0.347	2 1 1 3 2	0.296
1 1 2 1 1	0.739	1 2 3 2 2	0.315	2 1 1 3 3	0.217
1 1 2 1 2	0.707	1 2 3 2 3	0.236	2 1 2 1 1	0.618
1 1 2 1 3	0.489	1 2 3 3 1	0.210	2 1 2 1 2	0.586
1 1 2 2 1	0.666	1 2 3 3 2	0.178	2 1 2 1 3	0.368
1 1 2 2 2	0.634	1 2 3 3 3	0.099	2 1 2 2 1	0.546
1 1 2 2 3	0.416	1 3 1 1 1	0.417	2 1 2 2 2	0.513
1 1 2 3 1	0.390	1 3 1 1 2	0.384	2 1 2 2 3	0.295
1 1 2 3 2	0.358	1 3 1 1 3	0.306	2 1 2 3 1	0.269
1 1 2 3 3	0.279	1 3 1 2 1	0.344	2 1 2 3 2	0.237
1 1 3 1 1	0.540	1 3 1 2 2	0.312	2 1 2 3 3	0.158
1 1 3 1 2	0.508	1 3 1 2 3	0.234	2 1 3 1 1	0.419
1 1 3 1 3	0.430	1 3 1 3 1	0.207	2 1 3 1 2	0.387
1 1 3 2 1	0.468	1 3 1 3 2	0.175	2 1 3 1 3	0.309
1 1 3 2 2	0.436	1 3 1 3 3	0.096	2 1 3 2 1	0.347
1 1 3 2 3	0.357	1 3 2 1 1	0.357	2 1 3 2 2	0.315
1 1 3 3 1	0.331	1 3 2 1 2	0.325	2 1 3 2 3	0.236
1 1 3 3 2	0.299	1 3 2 1 3	0.247	2 1 3 3 1	0.210
1 1 3 3 3	0.220	1 3 2 2 1	0.285	2 1 3 3 2	0.178
1 2 1 1 1	0.677	1 3 2 2 2	0.253	2 1 3 3 3	0.099
1 2 1 1 2	0.645	1 3 2 2 3	0.174	2 2 1 1 1	0.556
1 2 1 1 3	0.427	1 3 2 3 1	0.148	2 2 1 1 2	0.524
1 2 1 2 1	0.605	1 3 2 3 2	0.116	2 2 1 1 3	0.306
1 2 1 2 2	0.572	1 3 2 3 3	0.037	2 2 1 2 1	0.484
1 2 1 2 3	0.354	1 3 3 1 1	0.298	2 2 1 2 2	0.452
1 2 1 3 1	0.328	1 3 3 1 2	0.266	2 2 1 2 3	0.234
1 2 1 3 2	0.296	1 3 3 1 3	0.188	2 2 1 3 1	0.207
1 2 1 3 3	0.217	1 3 3 2 1	0.226	2 2 1 3 2	0.175
1 2 2 1 1	0.618	1 3 3 2 2	0.194	2 2 1 3 3	0.096
1 2 2 1 2	0.586	1 3 3 2 3	0.115	2 2 2 1 1	0.497
1 2 2 1 3	0.368	1 3 3 3 1	0.089	2 2 2 1 2	0.465
1 2 2 2 1	0.546	1 3 3 3 2	0.057	2 2 2 1 3	0.247

EQ-5D state	Score
2 2 2 2 1	0.425
2 2 2 2 2	0.392
2 2 2 2 3	0.175
2 2 2 3 1	0.148
2 2 2 3 2	0.116
2 2 2 3 3	0.037
2 2 3 1 1	0.299
2 2 3 1 2	0.266
2 2 3 1 3	0.188
2 2 3 2 1	0.226
2 2 3 2 2	0.194
2 2 3 2 3	0.115
2 2 3 3 1	0.089
2 2 3 3 2	0.057
2 2 3 3 3	-0.022
2 3 1 1 1	0.296
2 3 1 1 2	0.264
2 3 1 1 3	0.185
2 3 1 2 1	0.223
2 3 1 2 2	0.191
2 3 1 2 3	0.113
2 3 1 3 1	0.086
2 3 1 3 2	0.054
2 3 1 3 3	-0.025
2 3 2 1 1	0.237
2 3 2 1 2	0.204
2 3 2 1 3	0.126
2 3 2 2 1	0.164
2 3 2 2 2	0.132
2 3 2 2 3	0.054
2 3 2 3 1	0.027
2 3 2 3 2	-0.005
2 3 2 3 3	-0.084
2 3 3 1 1	0.178
2 3 3 1 2	0.145
2 3 3 1 3	0.067
2 3 3 2 1	0.105
2 3 3 2 2	0.073
2 3 3 2 3	-0.006
2 3 3 3 1	-0.032

EQ-5D state	Score
2 3 3 3 2	-0.064
2 3 3 3 3	-0.143
3 1 1 1 1	0.226
3 1 1 1 2	0.194
3 1 1 1 3	0.116
3 1 1 2 1	0.154
3 1 1 2 2	0.122
3 1 1 2 3	0.043
3 1 1 3 1	0.017
3 1 1 3 2	-0.015
3 1 1 3 3	-0.094
3 1 2 1 1	0.167
3 1 2 1 2	0.135
3 1 2 1 3	0.057
3 1 2 2 1	0.095
3 1 2 2 2	0.063
3 1 2 2 3	-0.016
3 1 2 3 1	-0.042
3 1 2 3 2	-0.074
3 1 2 3 3	-0.153
3 1 3 1 1	0.108
3 1 3 1 2	0.076
3 1 3 1 3	-0.003
3 1 3 2 1	0.036
3 1 3 2 2	0.004
3 1 3 2 3	-0.075
3 1 3 3 1	-0.101
3 1 3 3 2	-0.133
3 1 3 3 3	-0.212
3 2 1 1 1	0.105
3 2 1 1 2	0.073
3 2 1 1 3	-0.005
3 2 1 2 1	0.033
3 2 1 2 2	0.001
3 2 1 2 3	-0.078
3 2 1 3 1	-0.104
3 2 1 3 2	-0.136
3 2 1 3 3	-0.215
3 2 2 1 1	0.046
3 2 2 1 2	0.014
3 2 2 1 3	-0.064

EQ-5D state	Score
3 2 2 2 1	-0.026
3 2 2 2 2	-0.058
3 2 2 2 3	-0.137
3 2 2 3 1	-0.163
3 2 2 3 2	-0.195
3 2 2 3 3	-0.274
3 2 3 1 1	-0.013
3 2 3 1 2	-0.045
3 2 3 1 3	-0.124
3 2 3 2 1	-0.085
3 2 3 2 2	-0.117
3 2 3 2 3	-0.196
3 2 3 3 1	-0.222
3 2 3 3 2	-0.254
3 2 3 3 3	-0.333
3 3 1 1 1	-0.015
3 3 1 1 2	-0.048
3 3 1 1 3	-0.126
3 3 1 2 1	-0.088
3 3 1 2 2	-0.120
3 3 1 2 3	-0.199
3 3 1 3 1	-0.225
3 3 1 3 2	-0.257
3 3 1 3 3	-0.336
3 3 2 1 1	-0.075
3 3 2 1 2	-0.107
3 3 2 1 3	-0.185
3 3 2 2 1	-0.147
3 3 2 2 2	-0.179
3 3 2 2 3	-0.258
3 3 2 3 1	-0.284
3 3 2 3 2	-0.316
3 3 2 3 3	-0.395
3 3 3 1 1	-0.134
3 3 3 1 2	-0.166
3 3 3 1 3	-0.244
3 3 3 2 1	-0.206
3 3 3 2 2	-0.238
3 3 3 2 3	-0.317
3 3 3 3 1	-0.343
3 3 3 3 2	-0.375
3 3 3 3 3	-0.454