Valuing preferences for EQ-5D health states in the Thai general population

A thesis submitted to the University of London

for the Degree of Doctor of Philosophy

Sirinart Tongsiri

London School of Hygiene and Tropical Medicine

September 2009

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 thesis 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 Negative binomial regression model was used to analyse the determinants of the numbers of inconsistencies. Elderly respondents and those with a lower education level tend to make more inconsistent responses. 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 Thai scores and the scores derived from respondents in the other five countries were extensively compared to examine the extent of the differences. It seems that the Thai scores are more similar to those of the UK. A cost-utility analysis of the prevention and control measures for cervical cancer in Thailand was used to demonstrate the difference of cost per QALYs if the scores from other countries were used to approximate the Thai preferences.

The thesis makes a number of contributions. The modelled scores are the first original population-based preference scores on health derived from the Thai general population. The determinants of logical inconsistency were examined, as well as an exploratory qualitative interview 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

I would like to show my gratitude to my supervisor, Professor John Cairns, for making my time at the London School of Hygiene and Tropical Medicine a most precious memory of my life. Doing a PhD with John is a treasured experience. His vital encouragement, support, patience and understanding from the beginning to the final stages of my studies enabled me to develop a great deal of knowledge and accomplishments in this thesis. My sincere thanks to Prof. Paul Kind for providing the classification guideline of the EQ-5D health states to be used in the fieldwork interview and to Louis Longworth for reading the first draft of the thesis proposal.

I would like to acknowledge and extend my heartfelt gratitude to the following persons who have been supporting me through the PhD process: my family for their moral support and understanding and their help in every part of my life, Pawasoot, Piyada and Krisada for always being there to listen to all my difficulties.

I am indebted to my research assistants: Sirinya and Artidtaya; all the field coordinators, both in pre-test studies and the fieldwork survey; and the interviewers who made the extraordinary journey to do 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 their untiring help in conducting the interviews in the respondents throughout the Northeast region. I would like to express my appreciation to all respondents and their families for sacrificing their valuable time to participate in the survey. I have learned a lot from them, and without them this thesis would not have been accomplished.

I am very grateful not only for the financial support of the fieldwork survey, but also for the intellectual and moral support, 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); Dr. Yot from the Health Impacts and Technology Assessment Program (HITAP) and all the staff. I would like to offer my thanks for the helpful collaboration from the staff of the National Statistical Office (NSO) Thailand for assisting with sharing information about the sample in this study and the Health and Welfare and the Disabilities Survey 2007. Moreover, I would like to

acknowledge the collaboration of the Thai Gynecologic Oncology Collaborative Group (TGOC) for sharing the EQ-5D health states of the Thai patients with cervical cancer.

Regarding my study in the UK, the tuition fees and the monthly expenditures were covered by the Royal Thai Government scholarship. I would like to give a big thank you to all the staff, especially Khun Jitree, Khun Vorapoj and Khun Natee at the Office of Educational Affairs, Royal Thai Embassy, London. Their help and hard work enabled me to make the most of my time doing a PhD in London, without having to worry about orgainsing finances or all those essential government documents.

My sincere thanks to all my friends and colleagues in the UK, especially, Walaiporn, Phusit, Sripen, Mariana, Paco, Ayako, Nareerut, Saowaluck and Inthira for making my life in the UK feel like my home away from home and for their constant reminders and much needed motivation. I wish to extend my thanks to my friends and colleagues in Thailand- for their friendship, support and inspiration, especially, Weerachai for being the main supplier of articles to which I could not access and Lertrit for her [his] moral support when I had writing-up crises. Finally, I would also like to extend my gratitude to the PHP IT services: Mick and Caroline, all HSRU staff and all LSHTM library staff for their constant support.

Table of Contents

List of Tables

List of Figures

Not life, but good life, is to be chiefly valued

-Socrates-

Chapter 1 Economic evaluation and its role in resource allocation: the gap in economic evaluation in Thailand

1.1 Introduction

Rising costs of providing health care are evident. To maintain the health of the population, world health expenditures have increased by 35% over the past five years (1). Health care expenditure in the US has increased from 9 percent of gross domestic product (GDP) in 1980 to approximately 15 percent in 2006 (2). The US President Barack Obama called for a health care reform plan in the US because of the growing costs or the "ticking time bomb" and the concern that the health system would be bankrupted if the health care costs are out of control (3). In fact, it is not only the US government, but governments of almost all countries that encounter the common problems of increased health care costs which are partly related to advancement of health care technologies, pharmaceutical costs and the extended life expectancy of the population. Resource constraints do not allow the provision of all of the potential medical interventions. Limited resources need to be allocated across different areas of health systems and are expected to be utilized effectively and efficiently.

The global trend towards increasing health care costs is also the case in Thailand. A number of reasons for the rising health care costs are as follows. Significant changes in medical practice have occurred in Thailand with a move from traditional medicine in the late 1880s to the latest advanced medical technologies, for example: stem-cell treatments, medical material technologies and nanotechnology aiming to prolong the life of the Thai people (4-5). These high-cost medical interventions have been limitlessly and carelessly imported into the treatments of Thai patients, thus expenditure on medical supplies and equipment has soared from 2.5billion baht (£55 million) in 1991 to 15.8 billion baht (£3 billion) in 2005 (5). The 1997 Constitution of Thailand provides the rights and freedom to access to health care, and the government is obliged to provide basic health care services (5). As a result, in 2001 the Universal Coverage (UC) scheme was implemented, aiming to provide health insurance for all Thai citizens. The proportion of the Thai population who are protected by health insurance schemes increased from 94.9% in 2003 to 96.3% in 2007 (6). Health expenditures in Thailand have risen from 3.82% of GDP in 1980 (approximately 545 baht or £10 per capita) to 6.10% (approximately 7,000 baht or £130 per capita) in 2005. Of the overall health

expenditures, 66.8% was covered by the household sector whereas 33% came from the public sector (5). Pharmaceutical expenditures constituted 42.8% of the overall health expenditure in 2005. The Ministry of Public Health, after implementing the UC scheme, encountered pressure to include high cost services in the benefit package. Given the limited government budget, health care rationing is becoming a concern among stakeholders. Budgets are allocated according to the numbers of beneficiaries in the catchment areas. In 2002, the capitation "price" per patient was 1,202 baht (£22)(7). The budgets have been increased to 1,447 baht (£26) in 2004, and 2,100 baht (£38) in 2008 (8-9). It is likely that increased Thai health expenditures will become a continuing problem for the government, causing the Ministry of Health to fall into deficit if health care resources are allocated carelessly.

It is widely accepted 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 (10). There was an increasing demand for a more transparent and participatory decision making process for the allocation of resources to health (11). One of the tools used to aid efficient resource allocation across different health interventions is economic evaluation, which is defined as "a comparative study of alternative interventions in terms of their costs and benefits" (12). Economic evaluations are widely used to aid resource allocation decision making (13). The benefits of economic evaluations have been recognised in several countries. In Australia, economic evaluation is required by law for new pharmaceutical products to be listed on the Pharmaceutical Benefits Scheme in order to be subsidized by the Government(14). Economic appraisal is used by the Dutch Health Insurance Executive Board to decide the health insurance packages (14). Johannesson concluded that economic evaluation is useful in the development of treatment guidelines and reimbursement decisions for medical technologies. (15).

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 who conducted economic evaluation studies. Over time, there have been many attempts to conduct economic evaluation by Thai researchers, and indeed, a number of economic evaluations have been performed in Thailand (16-19). In 2004, Thailand introduced explicit criteria for decision making on cost and efficiency criteria in the revision of the National List of Essential Drugs (20). Chiawchanwattana *et al.* (16) and Teerawattananon (17) 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 that the Health Intervention and Technology Assessment Program (HITAP) was established to provide the health technology assessment database and the standard methodological guidelines for economic evaluation in Thailand (4).

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(12). In cost-utility analysis, a health outcome is measured in quality-adjusted life years (QALYs), both quantitatively (years of life living in a particular health state) and qualitatively (utility of being in that state), given by individuals on the 0-1 scale where 0 represents death and 1 represents full health (21). In this method, QALYs are assumed to be a cardinal measure and 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 service organisations in any health setting, 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" (22). Badia *et al.* also supported this statement (23). Ideally, the health state valuations should be relevant to the populations under study so 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 (24). The US Panel on Cost-Effectiveness in Health and Medicine recommends 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 (11).

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 (16-18). 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 (25). Preferences of health outcomes in this 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 dead. Preference scores elicited from a representative sample of the Thai general population have not yet been established.

1.2 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 following are the specific objectives to be fulfilled before the primary aim can be achieved:

- Identify an appropriate health descriptive measure and the methods to be used to elicit preference scores
- Plan and carry out a large scale survey of health preferences
- Examine the extent of logical inconsistency, its determinants and the impacts it has on the preference scores
- Conduct model specifications to estimate the scores for unobserved health states
- Compare the Thai health state values with those from other countries

1.3 Outline of the thesis

The thesis begins in Chapter 2 with a brief review of theoretical backgrounds and the methodologies of the preference elicitation techniques to estimate preference scores. Researchers in the field of economic evaluation usually use numerous terms interchangeably to refer to preferences, which may cause confusion for other researchers, especially those who are unfamiliar with research in this field. The definition of preference used in the thesis is stated in this chapter. The literature on preference elicitation methods and some of the generic health descriptive measures are reviewed here because the preference scores are expected to be used to measure health outcomes across different medical interventions. The EQ-5D is selected as the

most appropriate measure to describe health in this study because it 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 were 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 and used to estimate scores for the unobserved health states. Additional literature on the relevant specific topics, for example, logical inconsistency and the preference scores elicited from the population in other countries, are reviewed and reported in the relevant chapters.

The fieldwork study was financially supported by the Burden of Disease Project (BOD), the International Health Policy Program (IHPP) and the Health Intervention and Technology Assessment Program (HITAP). This enabled the researcher to conduct the survey in a nationally representative sample. 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 staff of the funding organizations, 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 3. 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 information sheet, consent form and recording forms are included in the appendices at the end of the thesis.

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 were very knowledgeable regarding the location of the targeted respondents. Results of the fieldwork survey, including the demographic characteristics of the Thai respondents, their health conditions in the past 24 hours, the overall interview duration and the durations of the individual interview methods, and the mean scores for the health states used in the interview are reported in Chapter 4. By conducting the interviews, more insights were gained regarding the increased cognitive overload of the respondents, 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 the next chapter. The analysis of the determinants of interview duration is reported at the end of this chapter.

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 possible causes 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. The literature on the definition, measurement and management of inconsistent responses is reviewed at the beginning of Chapter 5. Two main definitions of inconsistent responses are identified and the definition by Dolan & Kind presented in 1996 is applied here. Count data models are used to explore the determinants of logical inconsistency. A negative binomial model seems to best fit the

data. To understand more about how individuals assign scores to health states and how they cope with the complex preference interviews, a preliminary qualitative study was conducted with a group of respondents who were relatively likely to give inconsistent responses. This gives more insight into the approaches of respondents in the preference interview and more can be learned about the challenges faced by interviewers.

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 6. 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 dilemmas; 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 could be the results of the misunderstanding of the interview process. The second dilemma is that if some scores have to be excluded, how does one 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 7.

A number of different model specifications are explored in Chapter 7 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 each model will generate a different score. 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 the 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. Differences between the predicted and actual scores are also reported. 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 thesis.

Economic evaluations are often performed in countries which do not yet have preference scores derived from the general population. These authors then use the preference scores estimated from the population in other countries to estimate QALYs of health interventions in their settings. This is also a common practice in conducting economic evaluations in Thailand before the Thai preference scores become available. Given the differences in public health systems, social systems and health beliefs, care should be taken when using the scores from other countries in cost-utility analysis. The Thai population-based preference scores are now estimated as presented in Chapter 7. To explore the differences of preference on EQ-5D health states between those of the Thai population and other populations, the nature of the population-based scores of six countries, including Thailand, are extensively compared in Chapter 8. The impact of using the scores from different countries is explored using data from a recent cost-utility analysis of the prevention and control of cervical cancer in Thailand. The comparison can also be used to highlight the differences between the people from different countries when they express their preferences on health.

The thesis ends with Chapter 9, where the overall contribution of the thesis is discussed. The thesis 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. The other contribution of the thesis is that, by conducting multinational comparisons of the preference scores, variations in the preference scores of different countries can be identified. The implications of using the preference scores from other countries on cost-utility analysis are addressed in the thesis.

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 thesis 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. Additional difficulties in linking with other respondent characteristics from the HWS database of the NSO Thailand were also addressed. 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 improve the model performance. The new version of EQ-5D, which is expected to be available in the near future, may potentially offer new 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 for health of the Thai population and are applicable to decisions over resource allocation in the health sector.

Chapter 2 Literature review of health description measures and preference elicitation methods

2.1 Introduction

To elicit preferences on health, Froberg & Kane stated that the following three steps should be considered: [1] defining a set of health states of interest and selecting the preference elicitation methods; [2] identifying a judge or groups of judges to provide preferences; and [3] aggregating across the judges and determining scales (26). These three steps are used as a conceptual framework in this review to gain insight into the theoretical backgrounds and methodologies used in previous studies in other countries. The theoretical background and controversies surrounding each step were explored before decisions were made concerning the health outcome measure and preference elicitation methods to be used to estimate Thai preferences on health. Note that, Step two: identifying the judges or groups of judges to provide preferences, was not reviewed in this chapter because at the outset it was decided that preferences will be elicited from the Thai general population.

The outline is as follows: firstly, the definitions of preferences including the conceptual framework of preference elicitation used in this study are reviewed. Next, five health outcome measures are briefly described and compared before proceeding to a review of preference elicitation methods. The interview protocol which can be applied in the population survey is reported, as well as the model specifications used to estimate the preference scores. The chapter ends by presenting the selected health state measure and preference elicitation methods to be used in this study.

2.2 What is preference?

Expected utility theory and its axioms are applied to elicit preferences over health under uncertainty (12, 27). To be able to explain how an individual "ought" to make a decision under uncertainty, the Expected Utility theory (EUT) is applied. As stated in Drummond *et al.*(12), the theory states that preference exists and obeys the axioms of transitivity, independence and continuity. The EUT can be used to indicate *the cardinal utilities under uncertainty* to explain how an individual makes a decision. Individuals are

assumed to have well-constructed cardinal utilities and are rational (according to the axioms) when making a decision (28).

Several terms are interchangeably used for utility under the QALY paradigm, for example, weights, index, values, utility or preference (11). Drummond *et al.* recommended that "preference" should be used as an umbrella term to describe the overall concept, whereas "utility" is used for preference obtained by asking respondents to make a choice on health outcomes under uncertainty (12). "Value" refers to preference obtained from making a choice under certainty. The recommendations by Drummond *et al.* are used throughout this thesis.

Preference is a numerical figure informing strength of desirability to live in a health state (11-12). Asking an individual to assign numerical figures to health states could be viewed as an attempt to "quantify" an individual's judgements on health. The number is used to illustrate whether she feels "better off" living in this particular health state compared to others, and if this is the case, by "how much" (29). As opposed to a non-preference approach where the scores given to attributes of a health state have equal weights, the scores established under a preference method represent the values an individual assigns to each attribute of a particular health state. Hence, it is likely that different weights would be given to different attributes. Preference scores are measured on an "interval scale" in the sense that the distance between 0.2 and 0.4 has the same meaning as that between 0.6 and 0.8 (12). However, a "ratio scale", where the distance from zero to a health state can be identified, is also recommended to be used by Froberg and Kane (30). Preferences of individuals are revealed using elicitation methods from which preference magnitude with an interval scale are produced (31).

2.3 Health description measures

Health is multidimensional and dynamic. To define "health" one can start by considering the range of health definitions from global to specific dimensions (32). At one extreme lies the WHO definition of health as: "states of complete physical, mental and social well-being and not merely the absence of disease and infirmity" (33). The other extreme is the very narrow definitions based on a "medico-technological" definition (32). No single measure can capture all attributes of health. To describe

health in the preference elicitation exercises, Froberg and Kane suggest that the health state descriptions should be "relevant" to the outcome of interest and that fewer than nine attributes are preferred to describe health states (26). According to user's purposes, health can be defined into two categories: Condition-specific and Generic instruments (32). Condition-specific measures are used to measure health outcomes in specific health conditions or in particular patient groups, whereas Generic measures are used to measure health outcomes across different patient groups. Because the purpose of this study is to elicit preference scores to aid the decision making on resource allocation across different diseases and patient groups, only the generic measure of health is reviewed and reported in this section.

2.3.1 Generic health outcome measures

There are several generic health outcome measures. Five measures that are commonly used to measure health outcomes are as follows.

Quality of well-being (QWB)

QWB is a preference-based health outcome instrument measuring the health-related quality of life component in the General Health Policy Model. The model has three components: mortality; morbidity or health-related quality of life; and time (34). The QWB scale ranges from 0-1 where 0 represents dead and 1 is for healthy life. Health outcomes are described in two parts: Functional status; and Symptom/Problem complex. For the Functional status part, health status is categorised into three dimensions: Mobility (3 levels); Physical activity (3 levels); and Social activity (5 levels). For the Symptom/Problem complex part, 23 items are classified (34). Preference weights for all items in both parts were elicited from a representative random sample of respondents from a San Diego community using the category method, magnitude estimation and Person Trade-off technique (35).

The Short-Form 6 Dimensions (SF-6D) measure

SF-6D is a reduced form of the short-form health survey questionnaire (SF-36) which has been developed to measure subjective health status by the Medical Outcome Study (MOS) group in the US (36-37). To establish the single utility index, the measure was reduced into a smaller number of items so that respondents could process the information in a preference elicitation survey using the Standard Gamble (SG) method (38). The preference scores were estimated from a representative UK sample (36).

Health Utility Index (HUI)

HUI was known as the McMaster Health Index Questionnaire developed in Canada by a multidisciplinary group of doctors, epidemiologists and statisticians (37). Three versions the HUI: HUI1; HUI2 and HUI3 have been developed. The HUI1 was developed by Torrance *et al.* (39). This measure consists of four attributes; physical function (6 levels), role function (5 levels), social-emotional function (5 levels) and health problems (8 levels), from which 960 health states can be defined. Utility scores for each state were established using the Visual Analog scale (VAS) and Time Trade-off (TTO) methods.

The HUI-2 was developed by Torrance *et al.* (40) using seven attributes: sensation (4 levels); mobility (5 levels); emotion (5 levels); cognition (4 levels); self-care (4 levels); pain (5 levels); and fertility (3 levels), defining 24,000 states. VAS and SG methods were used to establish interval scores for the states. A transformation function was used to transform ordinal data from the VAS into interval data (SG) (41).

The HUI-3 was developed by Feeny *et al.* (42). It has eight attributes: vision (6 levels); hearing (6 levels); speech (5 levels); ambulation (6 levels); dexterity (6 levels); emotion (5 levels); cognition (6 levels); and pain (5 levels), defining 972,000 states. VAS and the SG methods were used to establish the utility scores for the health states.

The Assessment of Quality of Life measure (AQoL)

This measure was developed in the early 1970s in Australia (43). The authors argued that, although several generic preference-based health outcome measures have been developed, none of the measures have been constructed based on the normal psychometric principles. This measure has been developed using these principles with five dimensions: illness; independent living; social relationships; physical well-being and psychological well-being. Each dimension comprises three items in which four levels are constructed for each item. The authors used a Visual Analog scale and Time trade-off techniques to elicit preferences in a random sample in Australia. To increase the sensitivity of the descriptive system and eliminate bias, as well as to allow for an alternative modelling methodology, the AQoL II has subsequently been developed. Preference scores were estimated using the Time trade-off and the Person trade-off techniques (44).

The EuroQol 5-dimension (EQ-5D)

The EQ-5D measure was developed by a multidisciplinary group of experts in 1987, aiming to establish a generic health outcome measure which is easily self-completed

(45-46). The EQ-5D is widely used in the measurement of population health status, such as in a survey in six European countries, conducted to determine the health status across the populations (47). Development history of the instrument and the preference elicitation procedures are thoroughly documented (48-49). 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 (46).

Dimensions and numbers of health states described in the aforementioned six health outcome measures are presented in Table 2.1.

The mobility dimension is attributed in both QWB and EQ-5D measures. No psychological attribute is described in the QWB. The HUI-2 measure has an attribute for sensory health, whereas the HUI-3 breaks down the sensory attributes into vision and hearing. The numbers of health states identified by these measures range from a few hundred to sixteen million states. To estimate QALYs, a number indicating the population preferences needs to be attached to each health state. To estimate the preference scores for the AQoL, with sixteen million states, or HUI-2, HUI-3 and SF-6D with more than ten thousand health states each, could be more challenging than to do so for the EQ-5D with just 243 states.

Table 2.1 Summary of dimensions and numbers of health states in seven health outcome measures

Descriptive system	Dimensions	levels	Health states
Quality of well-being	Mobility,Physical activity, Social functioning	3	1,170
	27 symptoms/ problems	2	
EQ-5D	Mobility, Self-care Usual activities Pain/discomfort, Anxiety/depression	3	243
SF-6D	Physical functioning, Role limitations Social functioning, Pain Mental Health, Vitality	4-6	18,000
Health Utility Index Version 2	Sensory, Mobility, Emotion Cognitive, Self-care, Pain Fertility	3-5	24,000
Health Utility Index Version 3	Vision, Hearing, Speech Ambulation, dexterity Emotion, Cognition, Pain	5-6	972,000
Assessment of Quality of Life (AQoL)	Illness, Independent living, Social relationship, Physical senses Psychological well-being each consists of 3 sub-dimensions	4	16,800,000
Assessment of Quality of Life (AQoL) II	Social dimension, Independent living Mental health, Coping Pain each consists of 20 sub-dimensions	4-6	6,446 billion

Adapted from Dolan and Olsen (32).

2.3.2 What measure is appropriate to use in the interview with the Thai general population?

There are several reasons to argue that among the six measures described the EQ-5D is the preferred measure for use in the Thai study. Compared with the other measures in Table 2.1, the number of health states is much fewer, thus the EQ-5D seems to be the easiest for respondents to comprehend. 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 (26). 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 it can differentiate between 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.*, the EQ-5D is easy to self-complete and can be used to discriminate the health statuses of patients with chronic obstruction of pulmonary disease (COPD) and rheumatoid arthritis from general population (50). 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 (51).

The EQ-5D is ready to implement in this study because the measure was officially translated into Thai and the translation was approved by the EuroQoL group. To achieve a semantic equivalence of the original questionnaire, the official translation was conducted by the Centre Outcomes, Research and Education (CORE) at Evanston Northwestern Health care in the USA in 2002. Two forward translations were undertaken from English to Thai by native speakers and two back-translations were followed by a native English speaker fluent in Thai. The final version was tested on eight respondents. The translation process was approved by the EQ-5D translation committee and the translation certificate issued by the EuroQol Group is provided in Appendix 1.

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.* (52) and Sakthong *et al* (53). 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 (54). Brauer *et al.* reported that the number of studies using EQ-5D increased from 5.7% in 1997 to 11.5% in 2001 (55). Rasanen *et al.* reviewed the economic evaluations published during 1966-2004 and reported that the measure was the most commonly used in the QALY estimation of health outcomes (46.8% out of 81 studies) (54). 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 (56). The measure was the most commonly used measure in the Industry submissions requesting listing by the Australian Pharmaceutical Benefit schemes reviewed during 2002-2004 (57), and in economic evaluation reports to NICE (58). It is recommended in the 2008 NICE methods guide that health effects are preferably measured by the EQ-5D (24). Recently, the EQ-5D has been recommended to be used in health outcome measurement in the Thai Health Technology Assessment Guideline (59). An additional benefit of using the EQ-5D in this study is that this will offer an opportunity to compare Thai preferences on health with those of other countries.

One may question the feasibility of implementing this measure in the Thai population. Although, to the best of the researcher's 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 SF-36 has satisfactory psychometric properties (60). 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-36 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 (61).

Although the EQ-5D health state descriptions are available in Thai, there is no official report on the extent to which the Thai general population understands these descriptions or the psychometric properties of the measure. 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 (62-63). 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 (64). However, the Thai EQ-5D will be used in this study even though the issues of descriptions are yet to be solved. After all, the Thai EQ-5D has been successfully implemented in several studies and translation issues have not emerged. The issue of translation is important, but is beyond the scope of this study.

In short, the EQ-5D is selected for use in this Thai study because the measure seems to be easy to comprehend, and is already officially translated into Thai and available from the EuroQol Group. The measure is used worldwide both in economic evaluation and in the measurement of quality of life of patients with several clinical conditions. A number of organizations recommend that health outcomes should be measured using the EQ-5D. In other countries, the EQ-5D is widely accepted by respondents and the psychometric properties of the measure, such as the construct and concurrent validity, are good. 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 not yet been examined in the Thai general population, it is likely that the properties are fairly good and the measure is highly acceptable by Thai population.

2.4 Preference elicitation methods

There are several methods used to elicit preference scores for health states. Drummond *et al.* suggests that preference scores can be divided according to how the questions are framed, i.e., whether the respondent is asked about certain or uncertain outcomes. In general, the preference elicitation methods can be classified into two groups: (A) Choice-based methods, where a respondent is required to take a risk or to sacrifice her time for being in full health. To estimate preferences on health the von Neumann-Morgenstern (vNM) utility theory, involving outcomes under uncertainty, is applied to establish scores with the interval scale properties (12). (B) Choice-less methods, where a respondent is asked to express her preferences without any sacrifice. The short descriptions of the interview types using each method are as follows.

2.4.1 Choice-based methods

Three methods: Standard Gamble, Time trade-off and Discrete Choice experiments are reviewed in this section. The first two methods are selected because they are commonly used in preference elicitation interviews and the third method is chosen because it has been proposed as an alternative for the first two.

Standard Gamble (SG)

This method has a strong theoretical background because the method closely follows the third axiom of the vNM utility theory and is regarded as a "gold standard" in terms of preference elicitation (12, 27, 65). In this method, a respondent is required to choose

between two hypothetical alternatives: (1) living in health state A for *t* years with certainty or (2) a gamble between immediate dead (with the probability of 1 – P) or a return to full health (with the probability of P). The probability P is changed until the respondent is indifferent between the two alternatives. The preference score assigned to state A is P (12). The assumption underlying this method is that individuals are willing to take risks. However, one's risk attitude could range across risk averse, risk taking or risk neutral scenarios(12, 65). The drawbacks of the method are that the questions asked in the interview may be difficult for respondents to comprehend, and some respondents may find it difficult to relate to probabilities, although some researchers have developed visual aids to increase respondents' understanding (12, 27).

Time trade-off (TTO)

This method was developed by Torrance *et al.* as an alternative pragmatic method to SG to elicit preference scores, with its simple and easy-to-administer instrument (66). A respondent is required to trade-off her time being in poorer health in order to be in perfect health with certainty. Different question formats are used for states viewed as better than death and as worse than death. If a respondent regards health state A as better than death, she is asked to trade some time living in health state A (*t* years, say *t*=10 years) in order to live in perfect health. The time in perfect health (or the time in health state A that is sacrificed) is varied until the respondent is indifferent between living in state A (for *t* years) and full health (for x years). The score for health state A is $\frac{x}{10}$. If a respondent regards health state A as worse than death, she is required to choose between immediate dead or to stay in health state A for *t* years before being in perfect health for *x* years followed by immediate dead (*t* + *x*=10 years). Time *t* is changed until the respondent feels indifferent between the two choices. The preference score for health state A is $\frac{-x}{10-x}$ (67). To balance the scores for states worse than death with the score 1, assigned to perfect health, the scores are transformed using a formula recommended by Patrick *et al.* (68). The formula is $\frac{U}{1-U}$ where *U* is the score for a state worse than death. The scores were transformed differently in the preference elicitation study in the US where the scores were simply divided by 39 (69). Debates are ongoing on how to value death and how the scores for states worse than death are to be transformed (70-71). Lamers compared the methods of transformation of the scores for states worse than death and reported that different transformation methods yielded different scores (72). The possibility of assessing the worse than death

states "in exactly the same manner as better than dead states" was reported by Robinson & Spencer (73). Compared with SG, the TTO method is likely to be easier for respondents to answer (27).

The TTO method is based mainly on the following two assumptions: individuals are willing to trade life expectancy and there is a constant proportional trade-off (65, 74). The constant proportional trade-off implies that a respondent is willing to trade the same amount of time independently from her life expectancy (27). For example, if one could live for ten years, one would be willing to trade-off five years of life in health state A to live in perfect health. If she could live for twenty years, she would be willing to trade-off ten years of her life in state A. However, elicitation of preferences using the TTO method is not without controversy. In fact, there are potentially a number of violations of the TTO assumptions. Bleichrodt *et al.* reported that the TTO scores could be inconsistent as a result of loss aversion and scale compatibility (75). Scale compatibility suggests that the respondents focus more on time sacrificed to live in full health rather than on the health states of interest because the response scale in the TTO method is duration of life. Loss aversion assumes that because a person assigns different scores to health states according to one's reference point, assigning a value for a health state from a "loss" perspective has more influence on health preference than from a "gain" perspective. This is in line with Froberg *et al.* who suggested similar assumptions of loss aversion but using a different term, "framing effects", to describe this effect (76). Framing decision making from the perspective of dying gave a different set of preferences from those framed in terms of surviving. Additional explanations of the loss aversion concept were made by Buckingham *et al.* using Hicks' utility theory to divide the compensation after change into two categories: Compensating Variations (CV) and Equivalent Variations (EV) (77). Spencer explained further by proposing that the questions in the conventional TTO method are framed in the gain of full health with the sacrifice of life years (CV-gain). However, the TTO questions can be framed unconventionally by asking respondents to imagine themselves living longer in poorer health (CV-loss) (78). Dolan *et al.* also supported the view that different preferences can be anticipated when respondents are asked to imagine themselves or other persons being in ill-health(79).

Sutherland *et al.* examined the attitude towards duration of survival for different health states and argued that the attitude depends on the quantity of survival time and quality of the health state; and that respondents tend to have a "threshold" "maximal

endurable time" (MET) (80). Respondents tend to assign negative values or view as worse than death if they live additional years after the threshold duration. Attema & Brouwer claim two biases are embedded in the TTO method namely; diminishing marginal utility of additional lifetime and discounting (81). The scores given to future additional life years are often found to be non linear, therefore, simply transforming the TTO score by the conventional formula ($\frac{x}{10}$ where x is the time in full health), would over-rate the utility for additional life years. The authors recommended a correction method for the TTO score; however, the authors used only one health state (back pain) to estimate the weights to adjust the future utility. By using only one health state to estimate the weights, one could argue whether the weight would be different if different health states were used. Moreover, it is unlikely that the weights estimated from only one state can be generalized to all other health states. Issues around operational methodologies for this method need to be explored further as well as the correction of TTO scores for states worse than death. Violation of the TTO assumptions is discussed further in a study by Buckingham and Devlin in which the authors argue that the slopes of indifference curves depend on the number of years in a particular health state and the severity of the health state (82).

Discrete choice experiments (DCE)

This method has been widely used in marketing, transport and environmental research, and there is increasing interest in using this method to estimate preferences for health care (83). DCE is based on several theoretical backgrounds, namely, the theory of Demand, welfare and consumer theory, and random utility theory (65, 84). The technique is based on an assumption that the value given by an individual to a good or a service is according to the characteristics of that good or service. The relative importance of various attributes can be discovered using this technique (12). An example of DCE being used to elicit preference scores is Ratcliffe *et al.* who applied DCE and ranking data to estimate preference scores for a sexual quality of life questionnaire (85). Ryan *et al.* applied DCE to the measurement of health outcomes for social care of elderly people (86). Hakim and Pathak conducted a preference elicitation for the EQ-5D in a small US sample using the SG and DCE methods (87). To the best of the researcher's knowledge, this is the only study applying DCE to the elicitation of preference scores for the EQ-5D. There are still several methodological limitations in applying the DCE method in a field survey. As stated in the study by Ryan and Farrar, there are ongoing difficulties in the survey methodology, as follows: difficulty in defining

a number of attribute levels, the selection of scenarios to be presented to respondents, the treatment of inconsistent responders and non-traders, and the specification of the benefit function (83). The DCE may increase the cognitive workload for respondents if the number of attributes exceeds five or six, and if more than twelve scenarios are presented to respondents (84).

2.4.2 Choice-less methods

Two methods are reported in this section: Ranking and Visual Analog scale (VAS).

Ranking method

The ranking method has a theoretical background in Thurstone's law of comparative judgement (88). The method is often used as a warm-up exercise to familiarise respondents with the health states used in the interview. The outputs of the ranking exercise are ordinal data (65). Craig, Busschbach and Salomon report that the scores estimated from a ranking method show a "strong linear correlation with both TTO and VAS responses" (89). Salomon reported statistical modelling methods using the ordinal responses elicited using the Ranking method to estimate preference scores for EQ-5D health states. He compared the data obtained from the Ranking method with the TTO data (90). One advantage of the Ranking method, as stated in Salomon, is that it is relatively easy and by using this method, it may be able to simplify the preference elicitation tasks for respondents who might be less competent in literacy and numeracy (90). On this ground, Salomon argues that it could be used in preference elicitation surveys conducted in wider settings and population subgroups.

Visual Analog Scale (VAS)

In this method, a respondent is asked to place a health state on a line with variable references at the bottom and the top of the line (12). For example, the scale on a line could range between zero and 100, where zero represents the "worst possible imaginable health state" and 100 presents the "best possible imaginable health state". A respondent is asked to place the health state on a line such that the distances between each state are proportional to the differences of her preferences over the health states (12). Ryan *et al.* reported that the VAS is regarded as easier than the SG and TTO method on the basis of completion rate (65).

An advantage of a choice-less method is that it is easy to use, but a drawback is that it lacks theoretical background (27). Economists prefer to use a choice-based method to elicit preferences because the results are observable and verifiable (12). By using the TTO method, for example, an individual's well-being from time spent in state A might be seen to be valued twice as highly as time spent in state B because, the individual stated that she feels indifferent between living in state A for eight years and living in full health for ten years, and she feels indifferent between living in state B for four years and living in full health for ten years. In a choice-less approach such as VAS, it is difficult to observe whether an individual strictly observes and applies her preferences to the distance between health states. Drummond *et al.* suggested that a non-choice based method can be used as a warm-up exercise for a respondent to prepare and familiarise herself with the descriptions of the health states before proceeding to a choice-based method (12). This suggestion was also supported by Torrance *et al.* (91).

There are ongoing debates supporting the use of the VAS method to estimate preference scores. Parkin & Devlin (92) argued that the method has a role in eliciting preference scores for health on three grounds. Firstly, the method is not based on the vNM utility theory as well as other preference elicitation methods except SG, therefore if the VAS scores cannot be regarded as utilities, neither can the TTO scores. Secondly, the VAS method, in fact, has a theoretical background in psychological theories of response to sensory stimuli. If QALYs are estimated from "weights" rather than "utility" and if QALYs are viewed from the perspective of extra-welfarism theory, then the estimation of preference scores can move away from the vNM utility theory. Hence, the scores elicited by the VAS method can be used to measure "weights" on health. And finally, the authors argued that an individual, in fact, makes a choice on which point of the line to put a health state. In response to the arguments of Parkin & Devlin, Brazier & Mccabe (93) argued that the VAS method suffers from context bias and end aversion bias. A respondent is asked to give value to being dead which is not a health state, as opposed to TTO and SG in which being dead is defined as score 0. Brazier & McCabe suggested that as an alternative to SG and TTO, ordinal data estimated from the DCE method or a ranking method can be used to explain utility as "behaviour" stated by a respondent rather than as what actually exists as "utility". There is increasing research using ordinal data to estimate cardinal scores, for example, Ratcliffe *et al.* report scores using DCE and Ranking methods compared to those elicited using the TTO method (85).

2.4.3 What method is to be used to elicit preference scores from the Thai general population?

Health state ranking, the VAS and the TTO methods are the most commonly used methods in the elicitation of preference for health outcomes (94). As described by Stiggelbout and Vogel-Voogt, a number of factors, ranging from stimulus, information interpretation and integration, and judgement and responses are extensively involved in the elicitation of preferences (95). To elicit preferences for health in Thailand, the elicitation method should be based, to some extent, on the theoretical backgrounds and applicable to the competency of a representative sample selected from the Thai general population. Priority should be given to the potential cognitive burden that would be faced by a respondent during the interview. One question with respect to the burden on a respondent is whether the choice-based methods are suitable for the cognitive abilities of the Thai general population.

To elicit preferences, a great deal of reading activity is required from a respondent. On average, the Thai general population have 7.9 years of education (96). Reading competency in the general population will vary. The reading skills are maintained and increased if one continues reading in one's everyday life but, as reported in the survey of the extent of reading in the Thai general population in 2007, only 66.3% of the total population were engaged in some kind of reading (97). Note that the definitions of reading in this survey were reading activities involving any of a wide range of reading materials (from newspaper, novels, textbooks, journals, newsletters to online materials). In the population aged 7 years and older, the greatest proportion of those who read is found in Bangkok (85.8%) and the lowest in the Northeast region (58.2%). Approximately seventy percent of those aged 25-59 years read newspaper and twenty-three per cent read knowledge-based documents whereas sixty-five per cent and seventeen per cent of those aged sixty years or older read newspaper and knowledge-based documents, respectively. Examples of knowledge-based documents are textbooks, newsletters and leaflets published by an organisation. Males read slightly more frequently than females and the proportion of those who read is greater in urban areas (77.7%) compared to rural areas (61.2%). The extent of reading activities is highly correlated with the level of education; the higher the level, the more frequent is reading. Only thirty-nine per cent of the elderly (aged 60 years or more) read. From these data, it is likely that the elderly respondents, those in the Northeast region and those with less education may have difficulties in participating in the preference

elicitation interview. Some authors recommended that a formal test of literacy and numeracy tests should be conducted before starting the interview (98-100).

According to the aforementioned arguments, the TTO method is chosen for use in this study, albeit a number of biases have been described, the method has supporting theoretical background, although the theory is not based directly on the expected utility theory. Compared with SG, the TTO method seems to be easier for respondents and as reported in Ryan *et al.*, SG is the most difficult technique to understand (65). The TTO method performs better in terms of test-retest reliability (30, 74). The method is most commonly used in the economic evaluation reports to NICE (58) and also in the RCT studies reported by Richardson and Manca (56). Brazier *et al.* supported the feasibility of the TTO method in preference elicitation (50).

The TTO method seems to be feasible for Thai respondents. Some Thai researchers have successfully used the TTO method to measure quality of life (QoI) of groups of Thai patients. For example, the QoI measurement of Thai patients after cataract surgery (101) and the measurement of QoI in Thai patients with end-stage renal disease comparing hemodialysis and continuous ambulatory peritoneal dialysis (102). One study suggested that Thai respondents could understand the TTO interview fairly well (25). However, the details of how the interview was conducted, the number of health states used and the extent of the cognitive burden on the Thai respondents were not provided. But even had they been, the lessons would be limited because the interviews were conducted in small groups of patients, rather than in the general population.

Although there has been no previous report on the feasibility of conducting valuation studies in Thailand, the experience from similar types of study conducted in other Asian countries can be noted and potentially applied to the Thai setting. A group of researchers from Singapore conducted studies on the feasibility of health outcome valuation in the Singaporean general population. One study reported that based on in-depth interviews with a representative sample of Chinese and Indian Singaporeans, both TTO and SG were feasible to implement and the level of acceptability between the two methods was similar (103). The population with lower education "preferred" to be interviewed by TTO. The mean age of this group of respondents was 41 years with an average of 10 years in education. If the TTO was "acceptable" to Singaporeans, it is also, to some extent, assumed to be "acceptable" to Thais. However, given that the average

education of the Singaporean sample is higher than Thai population, Thais may encounter a higher level of difficulty in participating in the outcome valuation survey.

Immediate dead is used in the TTO interview. Talking about death is sensitive in Thailand; discussing death with Thai respondents can be regarded as offensive. One study from Singapore aimed to gain more insights on "the impact of talking about death on health state valuation, a study among Chinese and Indian Singaporeans" (104). It was reported that Chinese and Indian Singaporeans were "generally comfortable" in the discussion regarding death and the term "passed away" was less offensive compared with the terms "sudden death" or "immediate dead". The results of this study should be taken into account in the study design of the study in Thailand.

In short, the TTO method is preferred in the preference elicitation survey in Thailand because the method has a supporting pragmatic theoretical background. The method is also easier than the SG and seems to be "usable" in the Thai general population, although the elderly and those with primary education may encounter difficulties reading the health states. No study reports the results of the use of the TTO method in the Thai general population but the evidence from using TTO methods in small groups of Thai patients and a small unrepresentative group of the general public, as well as the results of the use of the measure in neighbouring countries are encouraging.

Some authors believe that the demographic characteristics of respondents may have some implications for their preferences. Others report that these effects have no impact on preference scores. Froberg *et al.* (76) showed that demographic characteristics: age; gender; race; nationality; marital status; political persuasion; and religion have no effects on preferences expressed regarding health states. But they suggest that statistically significant differences may be obscured by small numbers of subjects and inadequate power. Nonetheless, Dolan *et al.* showed that age, sex and marital status are the most important characteristics to influence health state evaluations (105). The study by Dolan & Roberts also supported this argument (106). The effect of demographic characteristics on the Thai preference scores should also be explored.

2.5 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" (107). The interview protocol was reported in several papers (108-110). Countries which have estimated preference scores from their general population are presented in Table 2.2.

Table 2.2 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	Tsuchiya *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	TTO	1,115	42	Random Effects from the US scores)	Zarate *et al*. 2008
South Korea	TTO	488	42	Random Effects	Jo *et al.* 2008

(23, 67, 69, 111-121)

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 TTO scores, however, the study by Kok, Stolk and Busschbach 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" (122). The authors also advocated the implementation of a flexible interview protocol in different settings in which the 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, preferences were elicited using the 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: OLS, Fixed effects and Random effects models. The MVH protocol will 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. Another benefit of using the MVH protocol is that the results from the Thai study can be compared with the scores from other countries to gain insights into the differences between the Thai preferences and others'.

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.

In the UK MVH protocol, all 243 EQ-5D health states were categorised into mild, moderate and severe groups (109). A computer program was used to randomly select health states from each group into 6,080 unique combinations of 11 health states. Including four of the core states, each UK respondent was assigned 15 health states in the interview. The card allocation lists were generated and distributed to the interviewer: one list per one interview. It was not indicated whether all combinations were used. Props used in the face-to-face TTO interview were extensively described in the Time Trade-off User manual (123). Each health state was described on a 100 x 30 mm card in green colour; each dimension was described in a separate line. Full health was in pink and the "Immediate dead" in blue. The Time boards were made of three layers of thick hard board with a movable sliding scale indicating duration of time

ranging from zero to ten years. The layouts of the time boards provided were different according to whether respondents valued that health state as better than or worse than death.

2.6 Model specifications

It is nearly impossible for a respondent to assign scores to all 243 EQ-5D health states although one study aimed to collect all 243 scores from a sample of medical students (124). This is not relevant to the Thai study because the interviews are going to be conducted with the general population. Froberg & Kane divide the specification of models to estimate scores for health states not observed directly in the interviews into two categories: multiattribute utility function (MAU) and statistical inference (SI) (26). Regarding the MAU method, a respondent is asked to evaluate each level of an attribute while keeping other attributes constant. Examples of health state modelling using the MAU method are HUI and AQoL (125-126). In the SI method, an algebraic model is developed on the basis of one's preferences on a set of multi-attributed health states. This model is then used to predict scores for unobserved health states. The following section presents the use of the SI method in the model specifications for EQ-5D health states.

The first model specification to estimate the preference scores for EQ-5D health states was reported in the UK MVH study where 42 states were directly observed from the UK sample, modelling was used to interpolate the rest of the 243 states of the EQ-5D (107). Exclusion criteria were applied to the respondents who assigned values for fewer than three states in VAS and TTO, ranked 11111 equal to death or were missing 11111 and/or death, who were in the top 5% in terms of logical inconsistency on VAS or TTO, or who assigned the same values for all states in TTO. The data from interviewers who had a high rate of missing values and unusable data and incomplete interviews were also excluded.

The following criteria: goodness of fit (how well the model explains the differences between the estimated and actual scores), parsimony, consistency (better states "must" have higher scores) and transparency, were used to select the "best model". The preferred model was the Dolan-N3 model where the "decrements" from full health were captured by 11 dummy variables. Two sets of dummy variables for individual

dimensions were generated. The first variable set takes the value one for level 2, two for level 3 and zero otherwise. The second set takes the value one for level 3, zero otherwise. The 11th dummy variable (N3) takes the value one if any dimension is at level 3, zero otherwise. The utility score for any health state is the result of one minus the score estimated from the model (67). Terms representing interactions between dimensions were also generated.

The robustness of the Dolan 1997 model was tested. Two-thirds of the total of the usable 35,964 observations from 2,997 respondents were randomly selected to be the "internal sample" and the model was estimated from this sample. The remaining one-third of the observations was the "external sample". A set of coefficients was estimated using the scores from the internal sample, the coefficients were then used to estimate scores for all 243 states. The modelled scores were compared with the actual scores of the corresponding states in the external sample. The modelling was performed at the individual-level data. An ordinary least squares (OLS) model and a random effects (RE) model were estimated and a Lagrange Multiplier test (LM test) indicated that the RE model performed better than OLS. A RESET test was then used to test the RE model and this revealed that the RE model suffers from heteroskedasticity. The main effects model included all of the independent variables (but not the interaction terms between the dimensions), although some of them were "insignificant" at P-level 0.05. The R -square of the RE model was 0.46 and the mean absolute difference (MAD) was 0.046. Five states had an absolute difference between the estimated and actual scores exceeding 0.1. The Dolan N3 model has been used as a reference for the model specifications in estimating the scores for EQ-5D in a number of countries.

In 2002, Dolan & Roberts employed another approach in the modelling. Rather than modelling the deviation of scores for health states from full health, the scores were modelled to deviate from the poorest state (33333). The score for any health state was the sum of the mean score for state 33333 and the score calculated from the model. Eleven dummy variables were generated. The first set took the value one if the difference between state 33333 and the state of interest (on a particular dimension) was one (dimension is at level 2). The second set took the value one if the difference was two (dimension is at level 1). The final dummy variable (ANY13) took the value one if at least one dimension was at level 1 and at least one dimension was at level 3. The criteria to select the "best model" were: sign and size of the estimated coefficients; explanatory power; Ramsey RESET test for misspecification; ability to predict the

observed mean score; and the normality of error distributions. The model was estimated from the internal model and the estimated scores were compared with the mean scores obtained from the external sample. The R-squared of the model was 0.55, MAD was 0.03 and only one state had the absolute difference exceeding 0.1 and eight states exceeding 0.05. The authors suggested that the new model performed better than the Dolan 1997 model on the basis of a slightly higher predictive ability. The valuation in the new method was based on the differences in value between state 33333 and other states, rather than the valuation of the health states themselves.

Shaw *et al.* published a study of preference scores for EQ-5D states in the US in 2005 (69). Two types of independent variables: dummy variables; and ordinal variables, were generated. Two sets of dummy variables were created for each of the five dimensions. The first set took the value one for level 2, zero otherwise. The second set took the value one for level 3, zero otherwise. Five ordinal variables were created: d_1 took the number of dimensions moving away from level 1, minus one. If there was no movement away from state 11111, d_1 was zero. The variable i_2 took the number of dimensions at level 2, minus one. If no dimension was at level 2, i_2 was zero. The variable $i_2 - squared$ was the square of i_2. The variable i_3 was the number of dimensions at level 3, minus one. If no dimension was at level 3, i was zero. The variable $i_3 - squared$ was the square of i_3. Approximately 90% of the respondents (3,600) were randomly selected into the modelling sample and the remaining 10% (400) was the validation sample. A linear transformation was used to transform the scores for states worse than death. Both OLS and Probability-weighted least squares without a constant term were used to estimate the model using individual-level data. The dependent variable was one minus the scores estimated from the model. The Random effects model was selected as the "best model" based on model robustness and logical consistency of the estimated scores. The "D1 model" was the best model with completely consistent estimated scores and the MAD was 0.024 and seven states had a difference exceeding 0.05.

Other researchers have tried different methods to improve model performance. The South Korean model to estimate the scores for EQ-5D health states was reported in 2008 (121). The "q1 log model", where ten dummy variables were generated for level 2 and level 3 in the five dimensions, was the best model. The dependent variable is the log of one minus the scores estimated from the model. This model performs better than the "q1 model" in terms of predictive ability. A Random Effects model is used,

selected on the basis of parsimony and model robustness (MAD was 0.074). The rank correlation coefficients indicate that the South Korean scores are highly correlated with the scores from the UK, US and Japan with correlation coefficients of 0.759, 0.747 and 0.721 respectively.

Other independent variables have been introduced. Tsuchiya *et al.* conducted a health outcome valuation survey in the Japanese general population using a version of the MVH protocol (115). The respondents were requested to assign values to seventeen health states. Half of the total observations were randomly assigned to the modelling sample and the remaining half formed the validation sample. Several ordinal and dummy variables were included in the model to improve performance. The variables generated were, for example, C_3, the number of dimensions in level 3 and $C_3 - square$, the square of C_3. N_1 took the value one if there was level 1 in any dimension, C_1 was the number of dimensions in level 1, $C_1 - square$ was the square of C_1 and N_3 took the value one if there was level 3 in any dimension. The best model was the "N3 model". The adjusted R-squared was 0.40 and MAD was 0.014.

Zarate *et al.* estimated preference scores for the Hispanic population in the US valuation data (120). The dependent variable was one minus the scores estimated from the model. The independent variables consisted of two sets of dummy variables representing the movement from level 1 to level 2 and level 2 to level 3 in all five dimensions. A set of ordinal variables was generated. The dummy variable X4 took the value one when four or more dimensions were on level 2 or 3. The "Latin N3+X4 model" was the best model because there was no health state with an absolute difference exceeding 0.1 and MAD was 0.031, compared with the "UK-N3" model where thirty states had the 0.1 absolute difference and MAD was 0.244. The R-squared was 0.332.

To summarise, different model specifications have been used to estimate scores for the unvalued health states. Exclusion criteria have been used to exclude some directly observed but "unusable" scores, such as responses from respondents who gave scores for fewer than three states or valued all states equally. The "best" model has been selected on the basis of logical consistency of the estimated scores, model robustness and parsimony. To test model robustness, some observations were randomly allocated to the modelling sample and the remaining scores performed as the validation sample.

All models were estimated using individual-level data. OLS models were estimated first and model performances were compared with panel data models (RE or FE models). All models suffer from heteroskedasticity and misspecifications. The dependent variable was either one minus the scores estimated from the model (that is, "deviations" from full health 11111), or alternatively measured deviations from the worst state 33333. Both dummy and ordinal variables were used as independent variables. The main variables were the dummy variables representing the movements from level 1 to level 2 and level 2 to level 3 in the five dimensions. Other variables were generated representing movements away from level 1 in more than one dimension. The estimated scores were compared with mean observed scores for the corresponding states from the validation sample. R-squared, mean absolute difference and the numbers of states with the absolute difference exceeding 0.1 or 0.05 were used to examine model robustness.

2.7 Conclusion

This chapter reviewed the theoretical background of preference and preference reversals, as well as the methodological issues of preference elicitation, including health description measures, so as to identify appropriate methods and measures to estimate Thai preferences for health. The review was conducted in the light of searching for appropriate methodology to estimate preference scores to estimate QALYs used in cost utility analysis across different health interventions in Thailand. Decisions on an appropriate health description measure and on suitable preference elicitation methods have been made based on the feasibility of implementing these methods in the Thai general population. In the absence of any previous national survey of preferences in Thailand, it is assumed that methods that have been successfully applied in other countries are applicable to the Thai general population. The wide use of EQ-5D in Thailand gives the first clue leading to the selection of the health outcome measure to be presented to the Thai respondents. Additional support for the use of the EQ-5D is that it has been officially translated into Thai and can be used free of charge by a non-profit organisation. 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 utilised in Thailand, the experiences from other countries were reviewed and adapted. Ranking and VAS methods are to be used to familiarise the respondents with the health state descriptions before introducing the

TTO method by which scores will be elicited. 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. An additional advantage is that the Thai preference scores can be compared with the scores elicited from other countries so that more insights regarding Thai preferences can be obtained. Information obtained from the review in this chapter was used to plan the study design, which is reported in the next chapter.

Chapter 3 Preparations for the fieldwork survey

3.1 Introduction

As discussed in the previous chapters, the preference scores are to be derived from a representative sample of the Thai general population. In the beginning, it was decided that the preference elicitation interview would follow the MVH protocol, which was developed in the UK but has never been implemented in Thailand. Hence, it was not known whether Thai respondents would be able to cope with the interview tasks, or what possible difficulties Thai respondents might encounter in the interview. To gain further understanding of the feasibility of conducting preference interviews in Thai settings, pre-test studies were performed and the results are reported in the first section of this chapter. Lessons from the pilot studies assisted the re-design of the interview protocol, so as to be more applicable to Thai respondents. The recommended changes developed from the pre-tests are reported after the presentation of the results of the pre-test studies.

A sample could be recruited either from one or more provinces depending on the availability of funding. To achieve a representative sample, the sample size and the recruitment of respondents should be carefully calculated and planned. The sample size calculation and the method of recruitment, as well as the development of the survey instruments, are described in the second section. A group of interviewers were recruited and trained so that a large number of respondents could be successfully interviewed. The final section describes the interview process used in the fieldwork survey.

3.2 The pre-test studies

The study was registered with the EuroQoL Group at the beginning of the fieldwork survey. The Thai EQ-5D questionnaire and the user manual were received after the registration. There were two main objectives in conducting the pre-test studies: [1] to familiarize the researcher with the MVH interview protocol, and [2] to examine the feasibility of the preference elicitation interviews in Thai respondents. Two pre-test studies were administered and the results are presented in this section.

3.2.1 The London pre-test study

The first pre-test study was conducted with five Thai PhD students, known by the researcher, in London in 2006. The MVH protocol, with the following preference elicitation interview methods: Ranking, VAS and TTO, was conducted using fifteen Thai EQ-5D health states. The respondents reported that ranking health states was a time-consuming process because it was difficult to differentiate between the health state cards. The respondents took approximately one and a half hours to complete the interview and approximately 20 minutes was allocated just for ranking the health states. The participants seemed able to cope with the VAS and TTO tasks. Given that the Thai general population has an average of 7.9 years of education (equivalent to primary level in the Thai formal education system), as reported by the National Statistic Office, Thailand (127), it is likely that Thai respondents, who generally have a lower educational level than the PhD students, would encounter similar or greater difficulties when participating in the preference interview. The decision was then made to redesign the MVH protocol to be used in the interview with the Thai general population.

3.2.2 The pre-test studies in Thailand

Prior to the beginning of the data collection phase in Thailand, ethics approval was obtained from the London School of Hygiene and Tropical Medicine and the Faculty of Medicine, Mahasarakham University, Thailand. The first task of the researcher after returning to Thailand was to seek funding for the survey. Funding was one of the key factors determining the sample size and whether the survey could be conducted at the national level. After negotiations with potential Thai funding organisations, the survey was financially supported by the Burden of Disease Project (BOD); International Health Policy Program (IHPP); and the Health Impact and Technology Assessment Program (HITAP), Ministry of Public Health, Thailand. The budget was granted and it was sufficient to conduct the survey in a nationally representative sample, recruited from randomly selected provinces.

Two pre-test studies were conducted at the beginning of the data collection phase. To demonstrate the interview protocol to the funders, and to explore further the adaptations of the interview protocol, the first pre-test study was conducted with five IHPP researchers to determine the suitable number of health states that would limit undue cognitive burden on respondents. On average, the participants used 1.30-2

hours to complete the interview with 15 health states. The interview duration was in line with the London pre-test study. Considering that very few of the respondents in the fieldwork would be as highly educated as the IHPP researchers, most of the respondents were likely to have greater difficulty assigning values to fifteen states. Therefore, the number of health states to be assessed was reduced to eleven. It was assumed that the interview should last no longer than one hour and by reducing the number of health states interviewed, the preference interview could be undertaken successfully.

To further investigate the feasibility of the interview, the second pre-test study was conducted during 6-8 January 2007 in Nakorn - Prathom province, Thailand. This province was chosen because of its proximity to Bangkok (60 km). A convenience sample was selected by the field coordinators appointed by the research team. Twenty respondents, ten from urban and ten from rural areas, were invited to participate in interviews, which were conducted in the respondents' households. There were four interviewers: the researcher and three IHPP research assistants, who were trained by the researcher to conduct the preference interview protocol. Two interviewers were assigned to interview each respondent, because the interviewers were unfamiliar with the elicitation interview methods and they could support each other during the interview process. Remuneration of 200 baht (approximately £4) was given to each respondent, and 1,500 baht (approximately £30) to each field coordinator in recognition of their help and participation in the organisation of the interviews. At the end of the study, the interviewers were invited to give feedback on issues such as the interview protocol and the arrangements of interview sites. These were used as a guide to improve the interview protocol and the preparations for the fieldwork survey.

Forty-two health states from the UK study were used in the second pre-test study in Nakorn-Prathom (67). All selected health states were categorized into four sets and each respondent was asked to value eleven health states including four core states: 11111, 33333, death and unconscious; one very mild, two mild, two moderate and two severe states as presented in Table 3.1.

Table 3.1 Forty-two health states used in the Nakorn-Prathom pre-test study

Core states	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6
11111	11112	11121	11211	12111	21111	33321
33333	12211	11133	22121	12121	22112	11112
unconscious	11122	11312	21312	21222	21133	11113
death	13212	32331	13311	22122	12222	11131
	21323	32211	12223	22331	21232	32313
	33232	23232	23321	13332	22233	22222
	22323	32223	32232	32232	33321	23313

(67)

Demographic characteristics of the respondents and the interview durations are presented in Table 3.2.

Table 3.2 Demographic characteristics and interview duration in the Nakorn-Prathom pre-test study

Characteristics		no.
Gender		
	male	10
	female	11
Age (yrs.)		
	20-44	11
	45-49	7
	60+	3
Marital status		
	single	3
	married	13
Education		
	primary	8
	secondary	3
	university	10
Average interview duration	(mins)	
	Overall	69.7
	Ranking	10.9
	VAS	14.8
	TTO	38.5

Twenty-one respondents were interviewed during this 3-day study. Half of the respondents were younger than 45 years old and approximately 50% of the respondents had university level education. The respondents seemed to understand the VAS and the TTO tasks fairly well. The overall interview duration was approximately one hour; the time taken to perform the Ranking and VAS tasks was shorter than for the TTO task. The findings from this pre-test study were in contrast with the findings from the London study in that the Ranking interview duration was shorter and the TTO interview duration

was longer. One reason for the pre-test study having a shorter Ranking duration could be that fewer health states were used. The respondents in the second pre-test study were more likely to have lower education levels, thus the level of difficulties in the TTO task would have been higher for them compared to the respondents in London. Therefore, the respondents in the second pre-test study would take more time to complete the task. The average education level of this study, again, did not represent the average education level of the general population. The key features identified from the second pre-test study that could be used to design the fieldwork survey in the Thai general population are as follows.

A peaceful environment, with as few distractions as possible, is essential to help the respondents concentrate. Both interviewer and respondent were required to focus on the tasks, and teamwork was important in performing the interview. A one-on-one interview session could be arranged in a district health office, located in the respondents' neighbourhood, which is easily accessible. Lunch and refreshments should be prepared for the interviewers at the sites to successfully cope with the number of interviews per day. The working hours could be long, starting from early morning and going until late evening. An interviewer manual should be developed. The respondents in the pre-test studies recommended that any technical terms should be omitted from the interview dialogue. The interviewers should use simple terms as often as possible. The interview instruments should be adjusted to be feasible, given the likely competency of Thai respondents. The descriptions of each health state on the health cards should be easily readable, and the TTO boards should reduce difficulties with decisions regarding trading-off time. Some respondents complained about the similarity between health cards, making it difficult for them to differentiate between health states. It was then suggested that the interviewers should emphasize the differences of the health states. The TTO boards should be made of durable materials and extra TTO boards and repair kits should be brought with the interview team in case of loss or damage.

Eleven health states per interview was regarded as appropriate, since the respondents tended to complete the interview within approximately one hour as previously determined. However, the degree of severity of health states should be rearranged to cover a range of mild, moderate and severe states.

The respondents recruited for interviews should be literate and numerate, in order to be able to read the health state descriptions by themselves. They should also be able to commute to the arranged interview sites. The field coordinator is a key person who plays a major role in successfully identifying the respondents. To assist the field coordinators in gaining more understanding of the overall interview process, it was suggested that the coordinators themselves should be interviewed. They could then be better able to explain the interview procedure to the respondents and minimize the respondents' anxieties regarding participation in the interview.

3.3 Preparation for the fieldwork survey

To successfully conduct the survey in a representative Thai sample, five pre-survey procedures were undertaken: [1] sample size calculation and the sampling method; [2] preparation of the survey instruments; [3] selection of health states; [4] preparation of the respondents; and [5] interviewer recruitment and training. Details of each procedure are as follows.

3.3.1 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 (128), 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

(109).

Table 3.3 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 3.3 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.* (109).

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 the sample needed was 3,235. This implies that 3,235 observations are needed for each health state, with every respondent given the same health states. This number of observations was unlikely to be manageable in the Thai study, given the limited time and budget. The 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 0.10 and 0.05, respectively. As a result, at least 200 observations per health state were obtained in this study and each respondent is expected to be given the same sets of health states to be valued. Alternatively, the suggestion by Williams, as proposed in the pilot study of the MVH protocol application in the population survey, could be used as a guideline (107). Williams recommended that to avoid cognitively overloading respondents, each state could be valued by at least 35 people. This latter recommendation could be applied if the first recommendation is found to not be feasible in the fieldwork survey in Thailand.

The number of respondents needed to be recruited is correlated with the number of health states used in the data collection phase. If forty-five health states are used; and each state requires approximately two hundred observations, a total of nine thousand observations are needed. From the pilot study, it is likely that one respondent could be expected to value eleven health states; therefore, to obtain nine-thousand observations, one thousand respondents would need to be recruited.

To obtain a nationally representative sample, the Thailand National Statistical Office (NSO) was invited to collaborate in the recruitment of the sample. The NSO and the IHPP have had collaborative projects on a number of national surveys, two of them being the Health and Welfare Survey and the Disabilities Survey (HWS) in 2007. It was the respondents to the surveys who were expected to be recruited for the preference elicitation interviews. The principal sampling frame was taken from the Population and Housing Consensus Survey in 2000 (6). The frame was divided into two residential areas: Municipal (urban) and Non-municipal (rural). To cover the geographical distributions of the respondents, all 76 provinces in Thailand were divided into five regions: North, Northeast, Central, South and Bangkok. A stratified four-stage sampling method was implemented: the first stage consisted of provinces; the second stage involved blocks in municipal areas, and villages in non-municipal areas, the probability of blocks or villages being selected was proportional to the size of the regions. The third stage sampling units were private households, and the fourth stage 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 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.

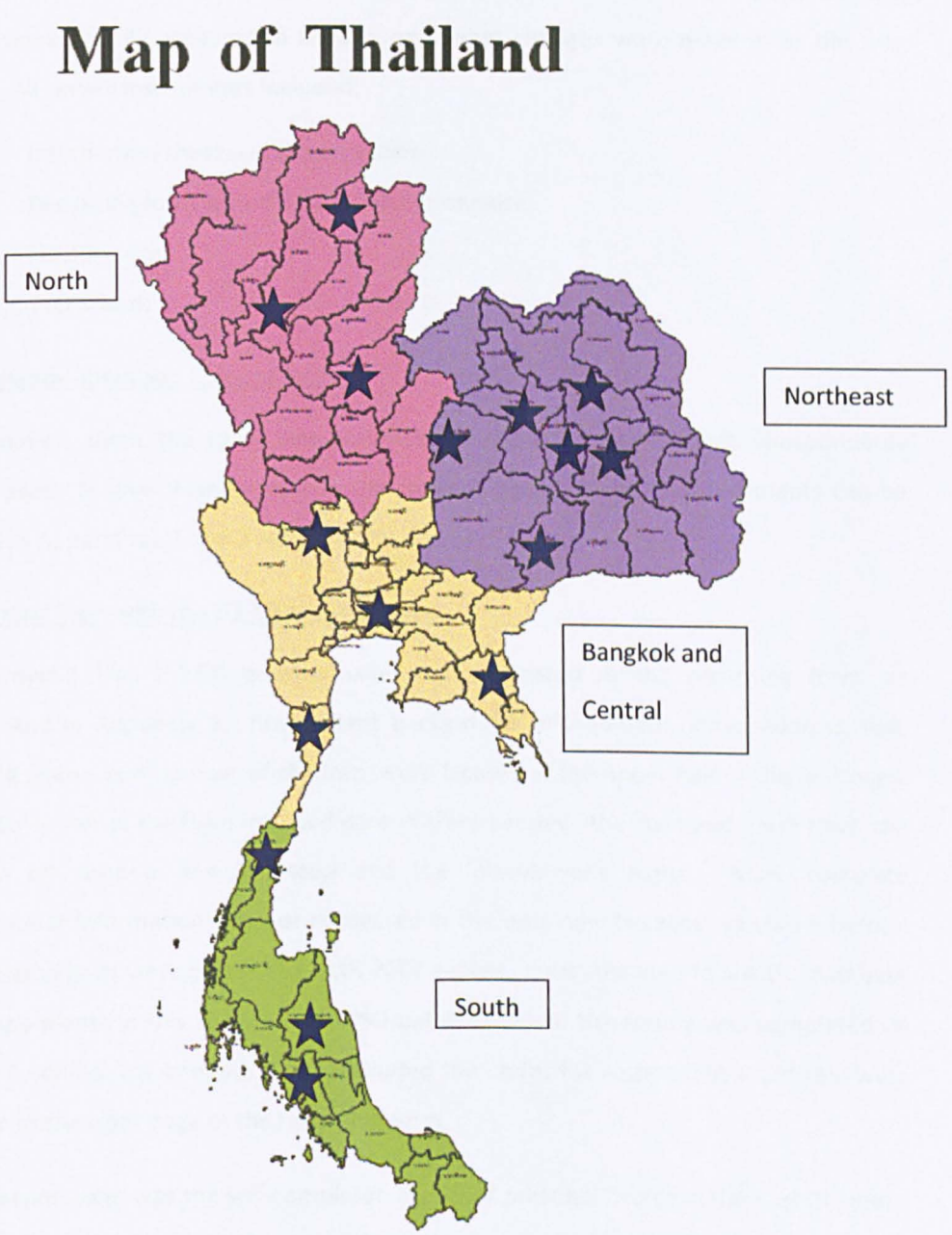
The greatest proportion of the Thai general population lives in the Northeast; therefore, the greatest number of respondents was recruited from this region. The smallest proportion of respondents was from Bangkok, followed by the South region. The three provinces in the South region were Chumporn, Nakorn Srithammarat and Trang. It should be noted that three southern provinces (Yala, Narathivas and Pattani) were not included because of political instabilities. 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. All residential areas in the Bangkok region were classified as urban.

The sample recruited for this study is presented in Table 3.4 according to regions, provinces, blocks/villages and number of respondents. The selected provinces are presented in Figure 3.1. The selected provinces are roughly indicated by the black stars.

Table 3.4 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**				

Figure 3.1 Geographical coverage of the sample



3.3.2 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:

- Information sheet and consent form
- 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. These two documents can be found in Appendices 1 and 2 respectively.

Recording form with the EQ-5D questionnaire

The original Thai EQ-5D questionnaire was integrated in the recording form, as presented in Appendix 3. 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 of the form included date of the interview, the start and finish time, the health set used in the interview and the interviewer's name. More complete 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. In order to do so, 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 presonal 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 3.

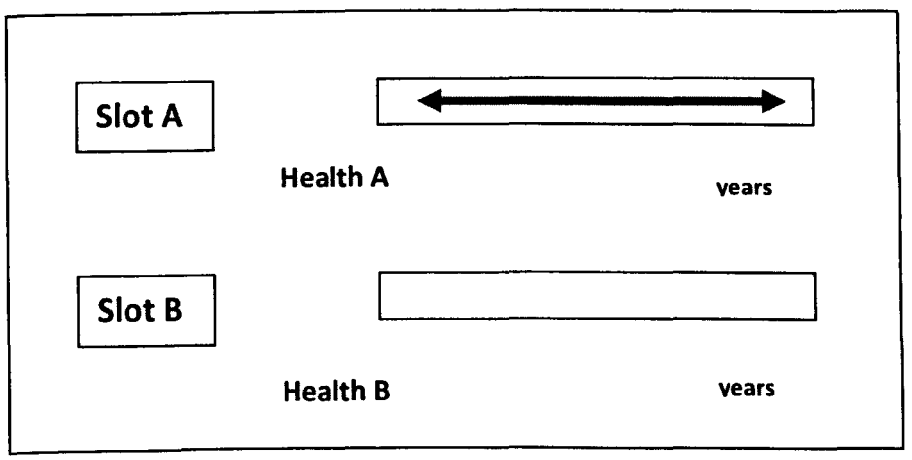
Health cards

Descriptions of health states written in Thai were printed on white paper sized 12 x 18 centimetres, with one line representing each dimension, namely: mobility, self-care, usual activities, pain/discomfort and anxiety/depression. All health cards were laminated to make them durable against possible damage (i.e. water, scratching). Eleven cards for each health set were put in an envelope with a distinctive label indicating health set number. Examples of health cards are presented in Appendix 4.

TTO boards

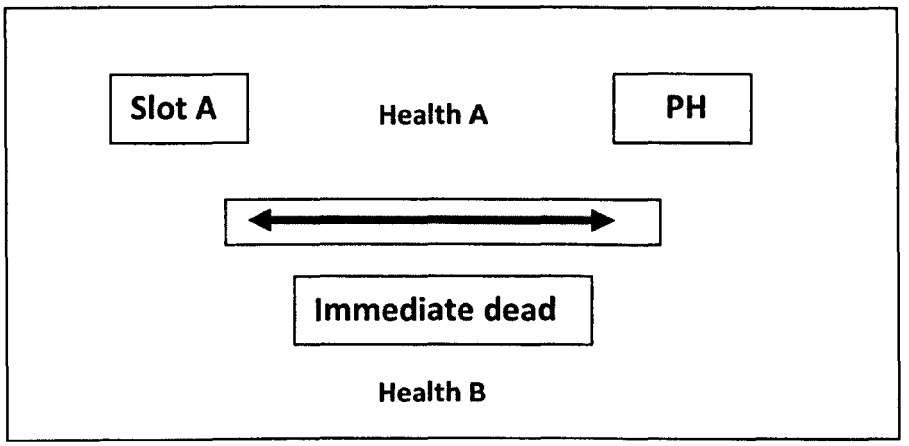
There were two TTO boards, one was used in the interview for health states which respondents regarded as better than death and the other with health states viewed as worse than death (see Figures 3.2 and 3.3). The boards were made of blue rectangular cardboard sized approximately 70 x 50 cm. The boards were intended to be larger than the original TTO boards to assist with the visualisation by the respondents.

Figure 3.2 TTO board for state better than death (TTO board 1)



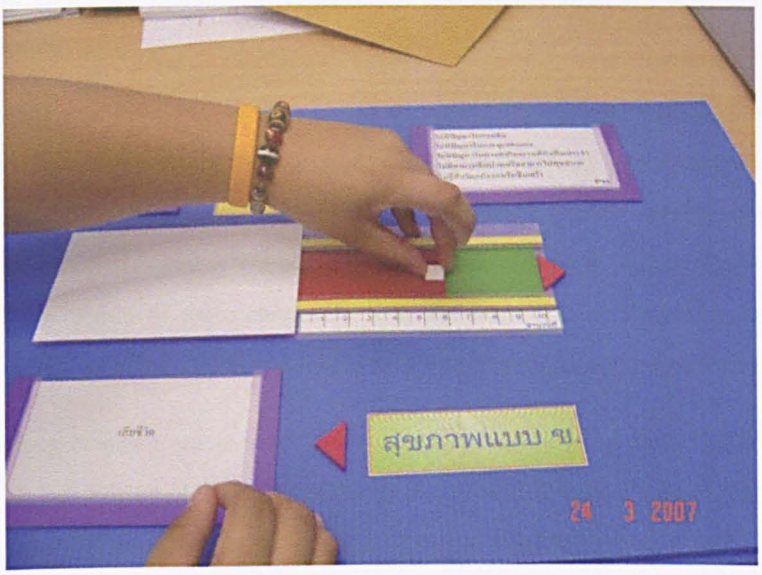
The health states used in the TTO interview were inserted in Slots A and B. Next to Slot A, a moving indicator was attached to identify years of life that the respondent would like to sacrifice. Note that the duration between 9 and a half and 10 years was divided into single months to allow respondents to choose the duration of 9 years and 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 3.3 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 dead" was a permanently fixed card described as immediate dead. 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 states worse than death is presented in Picture 3.1.

Picture 3.1 TTO board for state worse than death



3.3.3 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 would range 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 from Centre of Health Economics, University of York, 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 two at 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 Groups ranging from 1 to 9. Mild states were those in Groups 1-3 (five states in Group1, ten states in Group2 and ten states in Group 3), Group 4-6 were moderate and Group 7-9 were severe states (ten states in Group 7, ten in Group 8 and five in 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 to the mild group and twenty-five states to 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 3.5.

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 distance group 4, two states were left out from distance group 5 and one state was left out from distance group 6. Three of the health states in the mild group were randomly chosen to be combined with each of the twelve sets previously prepared. One health state from distance group 1, one from distance group 2 and one from distance group 3 were randomly selected. Since there are only twenty-five states in the mild group and only five states in distance group 1, after the states in distance group 1 were assigned (for set 1 to set 5), the states in this group were repeatedly allocated to sets 6 through 10. Thus the three mild states in sets 11 and 12 were similar to those in sets 1 and 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 the health states in all twelve sets are presented in Table 3.6.

Table 3.5 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
31311	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 3.6 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 were located at the lower right corner of the cards. The purpose of the coding was to minimise the interviewers' workload and to reduce the possibility of incorrectly recording the scores given to each health state. The coding also facilitated data entry. The first letters were A, B, C...L for

health sets 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.

Note that it was decided that a total of eighty-six health states (rather than forty-five states) would be used in the interviews. If two hundred observations were required for each health state, by using eighty-six states, a total of seventeen thousand observations would be needed. A sample of fourteen hundred respondents was calculated by the NSO and it was deemed feasible for the research team to conduct the interviews under the given budget. It was expected that this number of respondents could produce fourteen thousand observations (one respondent was asked to assign values for ten states). As a result, on average, one hundred and eighty observations were expected for each health state. This number was lower than is desirable (200 observations were expected according to Table 3.3), but this was the best number that could be achieved given the time and budget constraints.

It should be reminded that the method used to allocate health states for each interview in this study was different from other studies. In the preference studies previously conducted by other researchers, two methods have generally been used to select health states for the interviews. In the Japanese study, all respondents valued the same set of health states (115). In other cases, respondents faced different sub-sets of a larger number of health states, which was the case in the UK MVH protocol (7).

3.3.4 Preparation of the sample and the interview sites

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 letters 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 province, an outline of the interview process, arrangements for the interview site (number of tables and chairs and the lay-out of the interview sites to ensure a peaceful environment) and the props. Refreshments 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 the date and time indicated by the research team. The appointments were made via post, 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, the selection of interview sites depended on the feasibility of the areas and whether the field coordinators could find appropriate venues. The interview sites ranged from meeting rooms in hospitals or the provincial health offices to community centres, community leaders' households or the respondents' households (if the respondents could not come to the provided sites).

3.4 Recruitment and training of interviewers

Forty-eight interviewers participated in the survey. The interviewers were recruited from:

- Master's degree students from the Faculty of Pharmacy, Mahidol University.
- Bachelor's degree students from the Faculty of Pharmacy, Khon Kaen University.
- Master's degree students 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.
- Staff from the BOD, HITAP and IHPP offices
- Recently graduated bachelor's degree students from Lampang Rajabhat University. This group of interviewers were recruited in the later phase of the data collection because of reduced availability of the original interviewers.

A three-day interviewer training workshop was held 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. 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 closely matched the respondents to be interviewed in the data collection phase. This gave the interviewers a sense of what the "real" interview would be like. Regarding quality control of the interview, the performance of the interviewers was observed by the researcher and the assistants. Recording forms were checked after the end of the interview on the first day and interviewers were informed about mistakes that had been made. During this workshop, the researcher had a good opportunity to practice the preparation of interview sites and the management of 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 formed the research team: the researcher and the research assistants. 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.

3.5 The Thai interview protocol

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.

3.5.1 Respondent screening

Inclusion criteria for the respondents were that they were: [1] included in the list given by the NSO; [2] literate , with 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. Then 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 a health state card. If the respondents needed reading glasses and did not bring theirs along, 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 experienced communication problems with the interviewers.

3.5.2 The overall interview process

To ensure that each health set was used equally, 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 to be used to merge with the database of NSO, in which other personal information such as educational level and income were already available. This was to 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 the one 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 scale 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. Details of the interview protocol in the Thai study are presented in Appendix 5.

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 was 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 that occurred in the interview. In the following interview day, the

mistakes were demonstrated without blaming the interviewers, to remind the interviewers before starting the next interviews. Mistakes were mostly related to the direction of moving the time indicators in the TTO boards.

3.5.3 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 for termination, if the respondents were willing to terminate it, 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.

3.5.4 Differences between the UK MVH and the Thai interview protocol

The UK MVH protocol and the interview protocol in the Thai study share some common aspects. To familiarise the respondents with the EQ-5D health descriptions, they were asked to rate their own health states using the EQ-5D questionnaire and the thermometer scale. A face-to-face interview, consisting of Ranking, VAS and TTO tasks, was conducted with the duration of 10 years of living in the health state in question. To assist the respondents in allocating VAS scores for all health states, the bisection method was undertaken in that the best health state (from the rank given previously by the respondent) was firstly asked for the VAS score, followed by the worst health state and the state which was ranked approximately in the middle.

There are, however, several differences between the UK MVH protocol and the Thai protocol. Firstly, the number of total health states used in the interview is greater in the Thai protocol (86 states) than that of the UK MVH protocol (45 states). Since the pilot study suggested that Thai respondents may be able to cope with 10-11 health states,

only 11 health states per respondent were used in the interview. Secondly, in the UK MVH protocol, colours were used to identify differences in health cards: full health was in pink, immediate dead in blue, and health states in green cards. However, colours were not used in the Thai interview protocol because colour was not previously taken into account in the pilot studies, and it was not clear whether the same colour would provoke the same interpretations across different respondents. Therefore, it was decided that all health states would be presented on white cards. The Thai TTO boards were larger than those of the UK MVH. Thirdly, to minimise cognitive overload of the Thai respondents, the "Immediate dead" card was not used in the Ranking and VAS tasks in the Thai study. This decision was made based on the fact that the score for the "Immediate dead" state is already assigned as zero; this state is embedded into the TTO questions, rather than being a health state used to generate a score. By excluding "Immediate dead" state from the VAS task, the rescaling and transformation of the "raw" VAS scores will not be performed in this study.

The similarities and differences between the UK MVH protocol and the Thai protocol are summarised in Table 3.7

Table 3.7 Comparison of the MVH and the Thai protocols

Items	The MVH protocol*	The Thai protocol
Number of health states used in the study	45	86
Number of health states interviewed per respondent	15	10
Defined period of time	10 years	10 years
Self-report Health	yes	yes
Bisection method	yes	yes
health description cards	Colours were used to differentiate health	No color
TTO boards	smaller in size	bigger in size
Interview tasks	Face-to-face Ranking VAS TTO	Face-to-face Ranking VAS TTO
Is "Immediate death" included in Ranking and VAS ?	Yes	No

*as conducted in the UK study and reported by Dolan (67)

3.6 The qualitative study

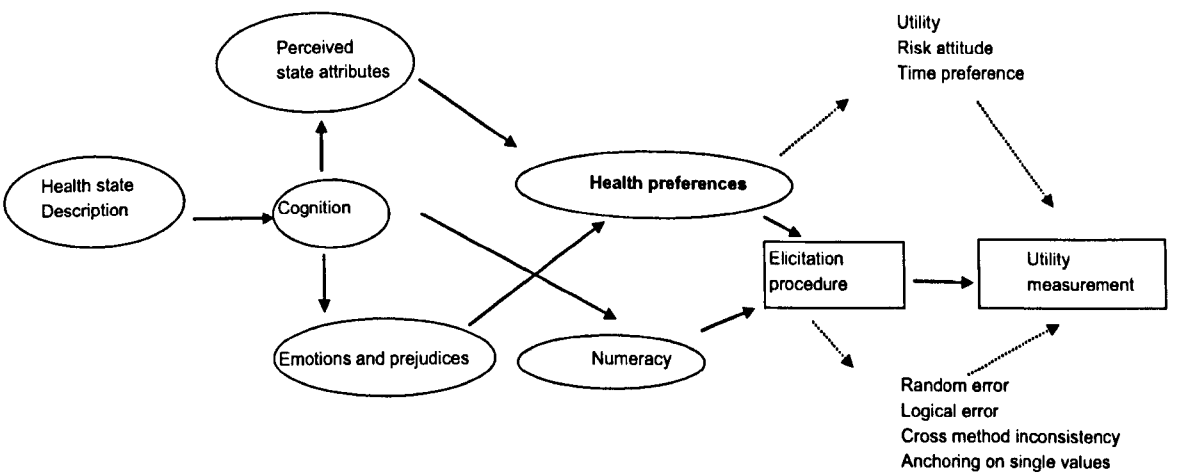
3.6.1 Background

It was apparent during data collection that elderly respondents with primary education tended to have more difficulties with the elicitation interviews and some assigned "irrational responses" or "logically inconsistent" responses with respect to the severity of the health states. This group of respondents may have had difficulties in completing the tasks due to poor literacy and numeracy as noted by Woloshin *et al.*(99). Limited numerical ability is likely to be a barrier to obtaining meaningful values from respondents using Standard Gamble (SG) and Time Trade-off (TTO) interviews. Respondents with limited reading ability are likely to encounter difficulties when

reading the health descriptions, which may lead to more difficulties when imagining themselves in the hypothetical health states and calculating number of years they are willing to sacrifice in order to live in full health. Although the investigation of the "irrational response" was not one of the main objectives of the study, an exploratory qualitative study can give initial understandings on how respondents cope with engaging in the preference elicitation interview. The findings gathered from these exploratory interviews could be used to generate the preliminary assumptions to explain the cognitive burdens of the respondents, and as a platform to develop future qualitative interviews.

To explore how the respondents cope with the preference interview, and to what extent, the conceptual framework proposed by Lenert and Kaplan could be applied to explain the thinking processes of the respondents in preference elicitation method(129). For individuals, to express their preferences, there are several factors that would influence the elicitation procedures as shown in Figure 3.4.

Figure 3.4 Model of the preference elicitation methods



How individuals respond to the preference elicitation methods

(129)

To express preferences, after reading health state descriptions, individuals are asked to imagine themselves being in the hypothetical health states. Respondents' imaginative ability depends on the adequacy of the health state descriptions, prior experiences, emotional response to the state, and how the health attributes are perceived.

Quantitative reasoning skills are needed for respondents to reveal their preferences in terms of probability or number of years they are willing to sacrifice for full health. Some other authors also argued that assigning scores to represent preferences towards health may not be an easy task. It is seemingly intractable for human cognitive functions to make a decision on the basis of multiple pieces of information (130). Gigerenzer *et al.* suggested that some groups of individuals may use a fast and frugal method or simple heuristic for decision making using only a specific part of overall information(131). Some may utilise every piece of information before making a decision, or they may employ either of the methods according to the complexity of tasks (132). Preference scores may not correspond with severity of health states, or the "transitivity" axiom of utility theory could be violated. "Irrational response" and the reasons why respondents may assign 'irrational responses' are explored by Lanscar and Louviere (132). The authors applied their research in a DCE method; they, nevertheless, also stated that these approaches can be applied in other preference elicitation tasks. One possibility could be that respondents have rational preferences but they assign 'irrational' preferences because of poor study design; they may be influenced by attributes outside those presented to them and apply those attributes in a different way that is unknown to researchers.

3.6.2 Interview procedure and sample

The preference interview method used in the qualitative study was similar to the interview in the data collection phase. The researcher explained the background of the research and the objectives of the interview. Next, the respondents were asked to complete the Thai EQ-5D questionnaire and assigned scores for their own health on the VAS scale. Then, six health states: 11111, 33333, 11121, 23231, 22221 and 23233 were used in the Ranking, VAS and TTO interviews. It should be noted that only six, rather than ten states used in the main interview, were asked in the qualitative study, because the interview in the qualitative study was used to learn about the cognitive process of the respondents engaging in the interview. By using all ten states, the respondents may have become overwhelmed by the task and not fully able to express their thinking process. The scores assigned by the respondents in the qualitative interviews were not used in the preference scores estimation. A "think aloud" technique is an appropriate tool to use to explore the coping mechanisms employed by respondents engaging in a preference elicitation interview. This technique was undertaken to examine response motivations in a DCE study by Ryan, Watson and Entwistle (133). This technique was

implemented in the qualitative research of this study. All interviews were tape-recorded. 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.

3.6.3 Data collection

The opportunity to conduct an exploratory qualitative interview emerged when the researcher reported the preliminary results of the Thai preference study to the funders in Thailand in July 2008. The researcher took this opportunity to conduct some exploratory qualitative interviews with elderly respondents. The qualitative interviews were conducted in KhonKaen province in August 2008. A convenience sample of ten respondents was recruited by the health volunteer working in the village where the researcher lives. The inclusion criteria were: [1] those aged 60 years and older, [2] those with primary educational attainment level (up to 7 years of formal education) and [3] those able to travel to the interview site by themselves. The respondents were screened by the researcher for literacy and the ability to participate in the interview. The interviews were tape-recorded and conducted in the health volunteer's household, which was in the respondents' neighbourhood.

3.6.4 Analysis

The interviews were transcribed and a content analysis was used to analyse the data. This analysis is one of the analytical methods used in qualitative studies to examine data against a pre-existing theory (134). The unit of analysis was the mechanisms employed by the respondents. After the data were collected, the mechanisms were identified and grouped according to the model described in Figure 3.4. Results of the qualitative study are reported in Chapter 5.

3.7 Conclusion

This chapter reports the preparations for the survey, including the pre-test studies, sample size calculation and sampling method, the recruitment of respondents, and the elicitation interview procedure. A stratified four-stage sampling method was implemented in this study, and potential respondents were randomly selected from seventeen provinces to ensure that the respondents were recruited from all

geographical areas. The interview protocol was designed using the MVH protocol as a prototype, 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 convenient sample using a think-aloud technique to explore the coping mechanisms of the elderly respondents in completing the interview tasks.

Chapter 4 Results of the interview and data analysis

4.1 Introduction

This chapter reports the results of the fieldwork survey which was conducted during May-August 2007. As described in Chapter 3, 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. Respondents' cognitive workloads are monitored relativelyusing interview duration, comments from the respondents and from the interviewers, and the qualitative study. The nature of the scores elicited from the TTO interviews 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 are compared. There are two parts to this chapter: the results of the survey; and the analysis of the TTO scores, the determinants of interview duration, and the cognitive workload. In the first part, a brief overview of the fieldwork management is reported, along with the numbers interviewed in each province. 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. Subsequently, 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 analysis of the determinants of interview duration.

4.2 Fieldwork managements

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

4.2.1 Locating the respondents

There were two stages to the selection process 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 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 initial mailing.

There were some respondents in the original lists who could not be located by the field coordinators. In this case, the researcher was notified by the field coordinators and new lists of the respondents from the same blocks were chosen to substitute those who could not be located, in order to maintain the agreed upon number of respondents in that block. The field coordinators were encouraged to search for the respondents until the decided number for that block was met. It was the case that some respondents were contacted later in the data collection phase by other respondents who participated in the interview. These late-found respondents were also recruited for the interview. The late-found respondents were those who worked or studied in other cities, or had moved to new addresses, without reporting this change to the registration offices. By using this strategy, the respondents who cannot be contacted by the field coordinators were invited to participate in the interview. In future studies, to be able to contact all potential respondents, more of community dwellers, rather than only hospital staffs, could be invited to help with the respondent identification.

4.2.2 The 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 schedules and the heavy traffic in Bangkok. All of the interviews in the Bangkok region were scheduled on weekends or holidays. The NSO staffs were

contacted and assigned to help in locating the expected households, as they had 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 staff 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 from hospital meeting rooms, out-patient departments (OPD) of village health offices, schools, temples, community centres, shops to the households of the respondents, field coordinators and community leaders. Examples of the interview sites and the interviews with the respondents are presented in Pictures 4.1-4.3.

Picture 4.1The respondent is given the instructions of the ranking interview

Picture 4.2 Interview at a small canteen in the respondent's neighbourhood

Picture 4.3 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 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 to 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 incorrectly understood that they were invited to an annual health check-up, rather than an interview, so they decided not to travel to the interview sites. Also in this case, the interviewers were sent to interview the respondents 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.

4.3 The Thai respondents

This section reports the numbers of respondents interviewed in the survey, demographic characteristics, numbers of interviews per interviewer and the number 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 from the respondents and the interviewers regarding the interview procedures are also included at the end of this section.

4.3.1 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 4.1. 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 44.2 (SD 12.5) years old. The highest number of respondents interviewed was from the Northeast region. The number of respondents interviewed was slightly higher than the target number, except for in the South region where, out of one hundred and seventy respondents expected to be interviewed, only one hundred and sixty-one were actually interviewed. Mean ages of the respondents from each region were lower than fifty years except in Chumporn province. The proportion of female respondents was higher than that of male respondents, except in Chaiyapoom province. Note that only respondents in rural areas were interviewed in Buriram province. To see the national representativeness of the sample population, demographic characteristics of the sample are compared with the Thai general population in Table 4.2.

Table 4.1 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 4.2. 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; the proportions of female respondents, adult age-group and those living in urban areas were greater. The proportions of elderly respondents and those with secondary and university education in the sample were lower than those of the general population.

Table 4.2 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

The 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 4.3. Interviewer No. 12 had the greatest number of interviews. Interviewers No. 20 and 21 had the least number of interviews. One interviewer was dismissed because she was not competent at the interview process; she repeatedly made mistakes when attempting to follow the interview manual, especially in the TTO method. The reason for the great differences in the numbers of interviews performed by each interviewer was because some interviewers were unavailable due to their obligations to their full-time jobs. These interviewers worked in a rehabilitation centre (i.e. as physiotherapists and occupational

therapists), and were unable to take any more leave from work in order to participate in the fieldwork. Given that most of the interviewers were master's degree students, they sometimes were unavailable to participate in the interviews, if the interviews were scheduled during their term-times. The problem was solved by recruiting new interviewers from the university located near the provinces where the interviews took place. As a result, the recently recruited interviewers were likely to have lower numbers of interviews compared to the original interviewers, and depending on their availability to participate in the fieldwork. The researcher and the research assistants also conducted some of the interviews if the interviewers were fully engaged, because some of the respondents were not available to wait for long periods of time. Note that, from Table 4.3, Interviewers No. 12-16 and No. 19-20 are those who were recruited later in the survey (the "new" group) to solve the availability problem of the interviewers who were trained in the beginning (the "old" group). On average, the "new" group interviewed approximately twenty percent of the total number of respondents. They tended to interview younger groups of respondents compared to those interviewed by the "old" group, although difference of mean ages between the respondents interviewed by the two interviewer groups was not statistical significant at p-level=0.05. The mean age of those interviewed by the new group was 43 years, whereas the mean age of those interviewed by the old group was 44 years.

Table 4.3 No. 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

4.3.2 Number of respondents per health set

The number of respondents per health set ranged from 101 to 129. The greatest proportion (9.16%) of interviews was conducted using Health Set 2, whereas the smallest proportion was done using Health Sets 10-12. Table 4.4 shows the number of respondents according to health sets.

One reason for the unequal number of respondents in the sets is the poor management of health set distribution to the interviewers in the fieldwork survey. The health sets were planned in each interview day starting from set 1 to set 12, according to the number of respondents to be interviewed on that day; for example, set 1 to set 12 were planned to be used in the interviews with twelve respondents. If the number of

respondents interviewed that day was lower than twelve (i.e. only nine respondents were interviewed), then the health sets used on that day were set 1 to set 9. On the following day, the health sets used for that day started from set 1, rather than from set 10, which was meant to have been used on the previous day, but was not. 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, previously interviewed, was checked that the researcher realized this problem. After that, the health set distribution plan was changed in order to enable the number of respondents for each health set to be equalised, especially those of health sets 10-12.

Table 4.4 Number of respondents according to health set

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

4.3.3 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 (113) assigned level 3 to at least one dimension; two of them assigned level 3 to four out of the five dimensions. Table 4.5 summarises the overall problems across 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 in respect to mobility and the second smallest proportion in respect to self-care. More than half of the respondents had some problems in pain/discomfort, while slightly fewer than half reported some problems with anxiety/depression. The highest proportion of no problems was seen in self-care, followed by usual activities.

Table 4.5 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 any one dimension 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 4.6.

Table 4.6 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 the education data are missing. Therefore, out of 645 respondents who were classified as having "good health", 581 respondents had data on 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", and of these respondents, half are male. Of those classified as having "good health", 242 respondents lived in urban areas (37.5%) and 403 (62.5%) in rural areas. Regarding the educational attainment level: 56.5% had a primary education level, while 27.0 % had a secondary education level and 16.5 % a university education level.

4.3.4 VAS scores representing health of the respondents

Two respondents considered their health to be in the "worst possible state imaginable". Over half of the respondents assigned scores of more than 80 to their own health, with the greatest proportion of the respondents (26%) giving the score of 80. Only 15% rated

their own health as the "best possible state imaginable". Table 4.7 summarizes the scores of the respondents' own health using VAS method.

Table 4.7 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	

4.3.5 Interview duration

Interview durations according to interview method are presented in Table 4.8. To examine the differences in 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 tended to have shorter interview durations in all interview types, although the differences were not statistically significant. Mean interview duration of interviews conducted by the "new" interviewer group was significantly shorter than those conducted by the "old" group (p-level=0.05). The "new" group conducted the interviews within approximately 48 mins, whereas, the "old" group required almost an hour.

Table 4.8 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

4.3.6 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 respondents expressed that Ranking and VAS methods were difficult. The reason was that they could not understand the health states descriptions, and were thus unable to imagine themselves being in these health states. Half of the respondents (51%) admitted that the TTO method was difficult, 20% of them thought so for the same

reason given to Ranking and VAS methods, while 10% expressed that they could not understand the method of trading-off time to live in full health. From the interviewer perspective, they stated that 605 respondents (45%) were confident with the tasks, 627 respondents (47%) were confident after participating in the interview, and only 8% were not confident with the interview at all.

Respondents' comments from the open-ended question

To obtain additional feedback, 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 health states, as described in the cards, because they had no previous experience of these health states. Some levels of the dimensions presented in the cards were thought to contradict each other. Some respondents admitted that they could not understand the interview tasks for the first two or three states, but eventually, after assigning values for a couple of states, their level of understanding seemed to increase. Some respondents were confused after reading several health cards, and they were unable to compare the latter states with the previous ones because they could not 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, but then the respondents gained more understanding of the tasks and, eventually, their level of confidence in assigning values for the 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 the health states. Some misinterpreted the health states and they took into account 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 comparing the two states in the TTO method. They learned that by considering two

health states as "indifferent", the new states were introduced and they could complete this task quickly.

4.4 Data management

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

4.4.1 Data entry

One research assistant was assigned to enter the results of the interview using the program Microsoft Excel 2007. Codes were generated for the respondents and the interviewers. All data for each respondent were entered in the same row. The data were then transferred using the program Stat Transfer (version 9) and were ready to be analysed using the statistical program Stata 10/SE. 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. As described in Chapter 2, 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 scores were:

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

Where: X =number of years being in perfect health (67)

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.

4.4.2 TTO scores transformations

In this study, TTO scores for states worse than death 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 (135). The equation used in the transformation is as follows:

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

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 (72).

To the best of the researcher's knowledge, the US is the only country where the linear transformation has been used in the utility score estimation model; the USD1 model (69). The equation for the linear transformation is as follows.

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

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 (69).

4.4.3 Numbers of respondents excluded from the data

The exclusion criteria followed those used in the MVH protocol (67). 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 4.9.

Table 4.9 Numbers of respondents excluded from the data and the causes of the exclusion

Causes	Ranking	VAS	TTO
Completely missing values	13	1	7
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	1,370

* 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 number of respondents who had completely missing values in Ranking is higher than that of the VAS and TTO methods because the interviews were terminated in the middle of, or just after, the Ranking method. Therefore, of fourteen respondents with completely missing values in the VAS method, thirteen respondents were excluded from the interview after the Ranking method. Of all 1,396 respondents who completed the Ranking interview and moved to undertake the VAS task: one respondent had completely missing scores in this method, two respondents gave values to fewer than three states and one respondent gave the same values for all states. There were 1,392 respondents who completed the VAS interview and moved to the TTO method. Of those, seven 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 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 dead" was not used in these two methods.

4.4.4 Mean actual TTO scores

Mean TTO scores of all 86 states observed from the overall respondents are shown in Table 4.10. Note that the monotonic transformation was used to transform the TTO scores in this table.

Table 4.10 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 4.10 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 number 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% of the total number of health states (27 states) 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.

4.4.5 Normality test

The Shapiro-Francia test was used to test the normality of the TTO scores distribution (136). The Stata program was used to calculate the z-statistics, used 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 skewness of the distribution of the scores was measured and arbitrarily classified as mild, moderate or severe. The number of health sates in each category is shown in Table 4.11.

Table 4.11 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

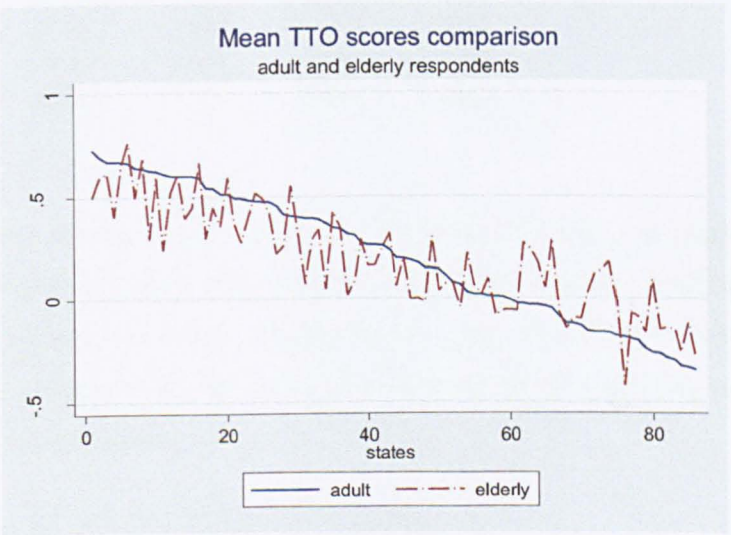
The z-statistics of less than 2.3 were classified as mild; those between 2.3 and 3.1 were classified as moderate and those greater than 3.1 were considered extreme. The distributions of more than half of the states were extremely skewed (55 out of 86 states with the z-statistics greater than 3.1). The states with one score of level 2 and level 1 in the other dimensions (mild states) and those with no scores of level 1 in any dimensions

(severe states) tended to have highly skewed distributions. The mild states tended to skew to the left (more states with positive scores) and the severe states tended to skew to the right (more states with negative scores).

4.4.6 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), a 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 4.1.

Figure 4.1 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 state (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 tended 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 4.12.

Table 4.12 TTO scores with significant difference between the elderly and adult groups

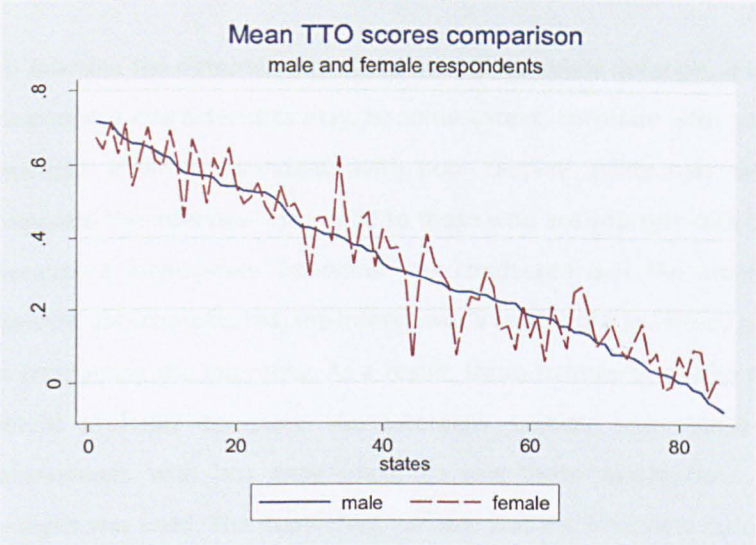
state	adult	elderly	p-value
11112	0.729	0.499	0.000
11232	0.628	0.293	0.001
21212	0.613	0.238	0.001
11221	0.673	0.405	0.004
32223	-0.263	0.100	0.013
23222	0.368	0.045	0.020
11313	0.397	0.073	0.022
12313	0.318	0.016	0.035
12221	0.543	0.296	0.037
21312	0.492	0.276	0.043
11212	0.601	0.398	0.046
23332	-0.166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

4.4.7 Mean TTO scores according to gender

A t-test was used to compare the differences between the mean TTO scores assigned by male and female respondents. Comparison of all TTO scores according to gender is illustrated in Figure 4.2.

Figure 4.2 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 state (according to the mean male scores). The 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 (approximately 0.5%) were significantly different between male and female respondents (p-value < 0.05). The states with significant differences are shown in Table 4.13.

Table 4.13 Mean TTO scores with significant difference 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 4.13, the female respondents, on average, assigned higher scores to only one state (11222), compare with those assigned by the male respondents.

4.5 Analysis of interview duration determinants

To examine the determinants of the overall interview duration, it is assumed that some respondent characteristics may, to some extent, correlate with interview duration. For example, elderly respondents with poor reading ability may take a longer time to complete the interview compared to those who are younger with better reading ability. Because a face-to-face interview was conducted and the interview procedure was considerably complicated, the interviewer's experience level may have played some role in conducting the interview. As a result, those interviewers who had more experience would probably complete the interview quicker than those conducted by the interviewers with less experience. To test these assumptions, multiple regression analysis was used. The dependent variable was the interview duration in minutes. The Independent variables were dummy variables representing the respondent demographic factors as shown in Table 4.2. A dummy variable was generated to represent whether the respondents were interviewed by the "old" or "new" interviewer groups. It is also assumed that "experienced" interviewers would complete the interview faster than those who were "inexperienced". In this analysis, the "experienced" interviewers could be either "old" or "new" interviewers, and they were assumed to gain higher level of "interview experience" after they completed the 5^{th} interview. Definitions of all independent variables are presented in Table 4.14.

Table 4.14 Independent variables for the interview duration analysis

variables	definitions
age_gr	1 for those older than 59 years old 0 otherwise
gen	1 for female 0 for male
secondary	1 for those with highest education at secondary level 0 otherwise
uni	1 for those with highest education at university level 0 otherwise
age_gr_pri	interaction term of age group and primary education level
age_gr_sec	interaction term of age group and secondary education level
age_gr_uni	interaction term of age group and university education level
hset *i*	1 if hset is *i* 0 otherwise *i=2-12*
old	1 if the interviewers were trained in the beginning of the fieldwork ("old")
exp	1 if the the interviewers interviewed their 6th respondents onward

Adults, males, those with primary education level, those given health set 1 and the respondents who were trained later in the survey ("new" group") were used as reference cases. Results of the analysis are shown in Table 4.15.

Table 4.15 Results of the regression analysis of the interview duration

Variables	Coef.	Std. Err.
age_gr	4.86**	1.55
secondary	-14.14***	1.25
uni	-17.20***	1.56
old	14.34***	1.70
old.exp	-6.14***	1.41
constant term	52.98***	1.21

*** p-level <0.001, ** p-level<0.01

The respondents who tended to have significantly longer interview duration were the elderly and those with a primary education level. The elderly tended to have interviews almost five minutes longer and those with secondary or university level education were likely to have a shorter interview duration by approximately fourteen and seventeen minutes respectively. Health sets were unlikely to have significant effects on the overall interview duration. Comparing the duration of the interviews conducted by the the "old" group of interviewers (those who were trained in the beginning of the fieldwork) with conducted by the "new" group (those who were trained later), those interviews conducted by the "new" group have significantly shorter durations by almost fifteen minutes (p-level=0.05). After the "old" interviewers gained some interview experience, i.e., the 6^{th} interview onwards, interview duration was significantly reduced by six minutes.

4.6 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. Two reasons for this success was the provision of interview sites that were easy to travel to and allowing interviews to be conducted in the respondents' household or workplace when necessary. 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 that 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 due to unexpected engagements. This situation was evaluated to prevent the same problem from happening again in the interviews in the other regions. The researcher identified a slightly higher number of expected respondents before sending names to the field coordinators in the other regions. 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, larger proportions 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 there were also larger proportions of the female respondents.

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 such severe problems, 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.

The new group of interviewers were recruited to solve the problem of unavailability of the interviewers who were recruited and trained at the beginning of the survey. Although the "old" and "new" interviewer groups were expected to perform "equally well" in conducting the interviews, and the characteristics of the respondents (for example, age) should not have been different between the two interviewer groups, it was seen that the "new" interviewers tended to interview the younger respondents, and were likely to complete the interview faster than the "old" interviewers. It was

shown in Section 4.3.5 that younger respondents required a shorter time to complete the interview. However, after controlling for age-group and taking into account whether the interviews were conducted after the interviewers gained some experience, i.e., after completing the 5th interview (Table 4.15), it is likely that the greater number of interviews conducted by the interviewers, the more experienced they gained, which lead to the shorter interview duration. This could be a reason why interview durations were significantly reduced after the interviewers had conducted sufficient interview. The interview site could have been responsible for interview duration. However, the data of interview sites were not collected; therefore, this effect cannot be analysed in this study. It is also interesting to compare the data quality in terms of the extent of logical inconsistencies obtained from the two interviewer groups. This will be analysed in the following chapters.

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 section, because of the complexities of the task. A considerable number of respondents could not understand the trading-off time task in the beginning, but gained more understanding and confidence as this section of the interview went on. This is in line with what the interviewers observed and comments made in response to the open-ended questions. The cognitive burden on respondents was partly explored using the average interview duration. The burden was still presumably high for Thai respondents although the interview protocol was redesigned from the original.

Note that the numbers of observations per health set presented in Table 4.4 are smaller than that expected from the sample size calculation in Chapter 3. 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 which suggested that Thais may not be able to cope with a preference interview using more than 11 health states (including state 11111 and 33333), in order to achieve 200 observations per *health set*, at least 1,720 respondents (or 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 4.10; twenty percent of the total health states have 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(107). 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) or 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.

Results of the statistical analysis of the interview duration suggested that the interview was significantly longer in elderly respondents and those with a lower education level. Those who assigned scores to health states in Set 1 tended to use a longer time to complete the interview tasks, although this effect is not statistically significant. These results can be used to guide further design of the interview protocols, which should be appropriate for different groups of respondents.

Most of the TTO scores for the eighty-six states were extremely skewed. The elderly tended to assign scores differently from adults. One reason could be that the elderly may have had difficulties trying to understand the descriptions on the health cards; they could have mistakenly interpreted the very poor states to be not so bad. Other reasons could be that they have had more experience of life and may be able to cope with the consequences of health states better than many of the adults. Male and female respondents appeared to assign similar scores to most of the states. Similar to the analysis in the UK study, the model specifications in this study were decided to be estimated as raw data, i.e. without transformations.

From Table 4.10, it was shown some of the actual mean scores were not consistent with severity of health states, in that mean scores of some poorer health states were higher than those of better health states. Some respondents assigned a score of 1 to the very severe health states. For example, mean score of state 22332 was -0.017 and that of state 22333 was 0.056. One reason could be because the respondents may have had difficulty in understanding either the health states or the elicitation questions; as a result, it may not be appropriate for these scores be taken to represent their

preferences. Alternatively, they may have expressed genuine preferences regarding these extreme states. This issue is explored further in the next chapter.

4.7 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 approximately one hour, with the longest time being spent on the TTO interview. Elderly respondents and those with primary level education tended to have longer interview durations. Interviewer characteristics have significant effects on interview duration, in that the interviewers who were trained at the beginning of the fieldwork tended to take more time to complete the interview. The "new" interviewers tended to complete the interview quicker than the "old" interviewers. After the interviewers gained experience in the interviews, they tended to complete the interviews faster. The distributions of the actual TTO scores of almost all health states interviewed were skewed. The elderly tended to assign lower scores for mild states and higher scores to poorer states. Gender has no significant 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 5 Logical inconsistency: Number of logical inconsistencies in the Thai study and the determinant factors

The mean TTO scores for each health state were calculated in the previous chapter. It was shown that some scores were not consistent with the severity of the health state and that mean scores of some better states are lower than some poorer states. The issues of inconsistent responses are examined in this chapter. The outline of this chapter is as follows. The treatment of logical inconsistency in previous studies and the literature on factors associated with logical inconsistency are reviewed in the first section. Secondly, the logical inconsistency apparent in the Thai study is rigorously investigated using both quantitative and qualitative methods. The association between the number of inconsistencies and a range of factors in the Thai study setting is analysed statistically. Results of the exploratory qualitative interviews are reported before the discussion and conclusion at the end of the chapter.

5.1 Literature review

Logical inconsistency and the effects of including the logically inconsistent responses in the estimation of preference scores are reported in a number of studies (116, 137-142). Age and education level have significant impacts on logical inconsistency, in that elderly respondents and respondents with less education have higher inconsistency (113-114, 116, 141). Lamers *et al.* reported that respondents who considered themselves to be religious tend to have higher levels of logical inconsistency (119). Respondents with poor health may have difficulties participating in the elicitation interviews and in giving logically consistent values for health states (113, 143). Retired people and smokers tend to make more inconsistent responses (116). Lower income respondents reported more inconsistent responses than did those with higher income (113).

One possible cause of logical inconsistency in TTO values was explained by Stalmeier, Wakker and Bezembinder using the concept of preference reversal phenomenon (144). Preference reversal occurs when individuals change their stated preferences orderings when different procedures are used to elicit their preferences (145). A large number of researchers explore the preference reversal phenomenon, for example, see the review of preference reversal by Seidl (146). Stalmeier, Wakker and Bezembinder demonstrated inconsistent TTO values using an example of living with migraine. In

their study, subjects preferred living 10 years with having migraine 5 times a week (10, 5 Mig) to living 20 years with having migraine 5 times a week (20, 5 Mig) but assigned higher value to (20, 5 Mig) than (10, 5 Mig). The authors explained this finding as a result of the violation of unilateral procedural invariance because, unlike the traditional preference reversals where choices are compared for more than one attribute, only one attribute (time) is changed in the TTO questions.

Two definitions have been used to measure the extent of logical inconsistency. 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 (140). Badia *et al.* and Devlin *et al.* also identified inconsistency in this way (141). The mean inconsistency rates reported in the Spanish study are 24.4, 25.9 and 59.2 in Ranking, VAS and TTO elicitation methods (116, 141). In the New Zealand (NZ) study, the majority of the respondents (80%) have fewer than 6 inconsistencies. Note that a face-to-face interview was conducted in the Spanish study using Ranking, VAS and TTO methods and the NZ study used a postal VAS survey.

According to Ohinmaa and Sintonen's definition, a health state is inconsistently valued if at least one better state has a lower score (138). The inconsistency rate is then the number of inconsistently valued health states as a proportion of the number of potentially inconsistently valued health states. This approach is also followed by Lamers *et al.* in the Netherlands (NL) study (139).

Causes of logical inconsistencies have also been explored. Inconsistent responses may result from 'irrational preferences'. Miguel *et al.* have conducted a qualitative analysis to see why there are 'irrational' stated preferences using a thematic approach (147). The authors suggested that the themes that emerged from the 'irrational' responses are as follows. The respondents may have used additional information or made their own assumptions about the health states by using their own experiences or using information learned from another choice. The authors also mentioned that the complexity of choices affects respondent's consistency. A respondent may become fatigued and bored with the complexity of demanding tasks.

5.2 Methods

Logical inconsistencies in the Thai study are systematically explored in this chapter. It is interesting to examine, firstly, the extent of the logical inconsistency in the scores

directly observed from the Thai respondents. The numbers of inconsistencies in the three elicitation interviews are reported. The relationships between the number of inconsistencies, respondent characteristics and other potentially relevant factors are examined using statistical analysis. Numbers of logical inconsistencies identified in the respondents interviewed by the "new" group of interviewers are also compared in this chapter. Recall that the "new" interviewers were those who were recruited in the later stage of the fieldwork to replace the "old" interviewers who were unavailable to conduct the interviews.

5.2.1 Measurement of logical inconsistency

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 consists of two health states with at least one dimension which is lower, or better, than a corresponding dimension in the other state, given other dimensions being equal. For example, considering two scenarios:

Scenario 1: the pair-wise comparison between state **12121** and **11221**

In these two states, mobility, pain/discomfort and anxiety/depression dimensions are at the same levels. Self-care in the first is worse (level 2) than in 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 a 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.

Scenario 2: the pair-wise comparison between state **11221** and **11222**

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

In this thesis, 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 for two reasons. First, it appears better able to capture the extent of inconsistency displayed by a particular respondent. The Ohinmaa and Sintonen's approach appears unable to distinguish between respondents who report only one inconsistent pair and those who report many inconsistent pairs of health states. In contrast, with Dolan and Kind the respondent who reports fewer inconsistent pairs of health states is recognised as having greater consistency. As a consequence, the Ohinmaa and Sintonen definition will tend to produce higher estimates of the rate of inconsistency. Suppose that a respondent scores the following health states from high to low: 11211 12112, 22112, 23223, 13123, 21332, 23323, 33233, 33333, 31131. The rate of inconsistency according to Ohinmaa and Sintonen would be 3/7 (43%), whereas it would be 3/27 (11%) according to Dolan and Kind.

Second, asking if at least one better state has a lower score can produce a different inconsistency rate from asking if at least one poorer state has a higher score. This asymmetry appears unsatisfactory. Let *i* represent a better health state and *j* a poorer health state in the sense that the level of at least one dimension is lower in *i* than in *j* and no level is higher. Ohinmaa and Sintonen ask for all (*i, j*) pairs, 'is there at least one state in *i* which has a lower score than *j*?' In the example in Table 5.1 there are three states (11211, 12112, and 31131) for which this can never be true. If instead the question is, 'for all (*i, j*) pairs is there at least one state *j* which has a higher score than *i* ?', then there is only one health state (33333) for which this can never be true. The number of potentially inconsistently valued pairs of health states will be independent of the order in which the health states are compared. But the distribution of these pairs is changed when the order is changed (see Table 5.1). Going down the list there is only one state with which no other health state can be compared; whereas going up the list there are three health states with no comparable health states. By using the Dolan and Kind approach, the number of logical inconsistencies is the same regardless of the direction in which the pairs of health states are counted. The number of logical inconsistencies will be examined in the scores elicited by the three interview methods. Distributions of the numbers of inconsistencies in the three interview methods are illustrated using histograms and the normality of the distributions is tested.

Table 5.1 Potential number of logical inconsistencies and the order of the states compared

Health states	Direction of counting	Potential number of logical inconsistencies	Direction of counting	Potential number of logical inconsistencies
11211	↓	5	↑	0
12112		6		0
22112		4		1
31131		2		0
13123		4		1
21332		1		1
23223		3		4
23323		1		5
33233		1		6
33333		0		9
Total number of logical inconsistency		27		27

5.2.2 Determinants of the number of inconsistent responses

The relationship between the demographic characteristics of the respondent and the number of inconsistencies is estimated using models for count data: Poisson regression (PRM); Negative binomial regression (NBRM); Zero-inflated Poisson regression (ZIP); and Zero-inflated negative binomial regression (ZINB) models. Three models, one for the scores elicited from Ranking, one for VAS and one for TTO data are estimated. Dependent variable (y_i) is the number of inconsistent pairs for respondent *i*. The number of inconsistent pairs is assumed to be associated with respondent characteristics, such as age, gender, number of children, interview duration, residential area, education level, health sets and interviewers. The general formula is:

$$y_i = \beta_0 + \beta_1 age + \beta_2 sec + \beta_3 uni + \sum_{j=4}^{14}\sum_{k=2}^{12} \beta_j hset_k + \sum_{j=15}^{61}\sum_{g=2}^{48} \beta_j interviewer_g + \beta_{59} gen + \beta_{60} no.of\ children + \beta_{61} duration + \beta_{62} ownVAS + \beta_{63} urban$$

PRM specifies that y_i given $\boldsymbol{x}_i$ has the Poisson distribution with a log link function:

$$f(y_i|\boldsymbol{x}_i) = \frac{e^{-\mu_i}\mu_i{}^{y}}{y!}, y_i = 0,1,2,\ldots \quad (1)$$

with expected number of independent pairs $E[y_i|\boldsymbol{x}_i] = \mu_i = \exp(\mathbf{x}_i'\beta)$ where $\mathbf{x}_i'$ is a vector of independent variables and β is a vector of coefficients (148). Overdispersion may result from unobserved heterogeneity that is not covered adequately by the

Poisson function. To account for overdispersion, NBRM with gamma distribution is used. Density of NBRM is:

$$f(y|\mu,\alpha) = \frac{\Gamma(y+\alpha^{-1})}{y!\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}}\left(\frac{\mu}{\alpha^{-1}+\mu}\right)^{y}, \alpha \geq 0, y = 0,1,2,\ldots \qquad (2)$$

Where α represents variance and if α =0, this density function is Poisson distribution. Mean inconsistency rate is $\mu_i^* = \exp(\mathbf{x}_i'\beta)v_i$, and v_i denotes unobserved heterogeneity for observation i (148).

If there are a considerable proportion of completely consistent respondents (number of inconsistent pairs = 0), a zero-inflated model: ZIP and ZINB are used.

Zero-inflated model

There are two processes in the zero-inflated model. One process is for the completely consistent respondents (number of inconsistent pairs = 0). Let i denote an indicator variable with value 1 if $y_i = 0$, and 0 otherwise. The outcome is binary which can be modelled using the logit or probit models. The other process for the inconsistent respondents ($y_i > 0$) is to use PRM for ZIP and NBRM for ZINB. For the zero-inflated model, two probabilities (from the two processes) are mixed.

For ZIP model: Probability of completely consistent respondent ($y_i = 0$) is:

$$\Pr(y_i = 0|x_i) = \varphi_i + \{(1-\varphi_i)\Pr(y_i = 0|x_i)$$

$$= \varphi_i + \{(1-\varphi_i)\frac{e^{-\mu_i}\mu_i^{y}}{y!}$$

If $y_i = 0$, then $\mu_i^{y} = 0$, therefore, $\Pr(y_i = 0|x_i) = \varphi_i + \{(1-\varphi_i)e^{-\mu_i}$

The probability of being an inconsistent respondent ($y_i > 0$) is:

$$\Pr(y_i > 0|x_i) = (1-\varphi_i)\Pr(y_i > 0|x_i)$$

$$=\{(1-\varphi_i)\frac{e^{-\mu_i}\mu_i^{y}}{y!}\}$$

For ZINB model: the probability that a respondent is completely consistent ($y_i = 0$) is:

$$\Pr(y_i = 0|x_i) = \varphi_i + \{(1-\varphi_i)\Pr(y_i = 0|x_i)$$

$$=\varphi_i + \{\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}}\}$$

The probability of being an inconsistent respondent (y_i>0) is:

$$\Pr(y_i > 0|x_i) = (1 - \varphi_i)\ \Pr(y_i > 0|x_i)$$

$$= \{(1 - \varphi_i)\frac{\Gamma(y+\alpha^{-1})}{y!\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}}\left(\frac{\mu}{\alpha^{-1}+\mu}\right)^{y}\}$$

Where: φ = probability of being a completely consistent respondent

$$= \frac{\exp(\gamma_i z_i)}{1+\exp(\gamma_i z_i)}$$

where γ_i is a vector of coefficients and z_i is a vector of inflation variables (149).

Overdispersion test

Two tests are used to examine overdispersion in the resulting models; Likelihood-ratio test (LR test) and Pearson chi-square test are used to test the null hypothesis $H_0: \alpha = 0$ (PRM).

LR test = 2 (LR Poisson – LR negative binomial)

And the Pearson chi-square test:

$$\sum_{j=1}^{J}\frac{(n\overline{p_j} - n\widehat{p_j})^2}{n\widehat{p_j}}$$

Where: $\overline{p_j}$= observed frequencies with *y=j*, $\widehat{p_j}$ = predicted frequencies with *y=j* (148)

The Pearson chi-square statistic divided by the number of the model degrees of freedom is used to test for overdispersion. A Pearson statistic close to one indicates that the model is not overdispersed (150).

Goodness of fit

Three methods are used to examine the models' goodness of fit; Deviance, Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC).

Deviance: $D(y, \hat{\mu}) = 2\ \{L(y) - L(\hat{\mu})\}$

Where: $L(y)$ is the maximum log-likelihood of the full model, $L(\hat{\mu})$ is the maximum log-likelihood of the fitted model (148).

AIC:

$$AIC = \frac{\{-2ln\hat{L}(M_k)+2P_k\}}{N}$$

Where: $\hat{L}(M_k)$ = the likelihood of the model, P_k = the number of parameters in the model, N = the number of observations (149).

BIC:

$$BIC = D(M_k) - df_k lnN$$

Where: $D(M_k)$ = deviance of the model, df_k = degree of freedom associated with the deviance. The more negative the BIC and the smaller the AIC, the better the fit of a model (149).

To select non-nested models – the Vuong test

An LR test is used to compare two competing models which are non-nested. The model selected is the model that is "closest to the true conditional distribution" (151). The null hypothesis is that model F_θ is equivalent to model G_γ.

The ZIP model is competing with PM and ZINB is competing with NBM. The Vuong test is used to select between the two competing models. The selected model is the "closest model to the true conditional distribution". ZIP is the null model and PM is the alternative. ZINB is the null model and NBM is the alternative. A crucial element of Vuong's analysis is that it need not be the case that either competing model is "true"; they may both be incorrect but the analysis attempts to identify the model that is "closet" to the truth (151).

Vuong statistic is as follows: $V = \frac{\sqrt{n}\,(\frac{1}{n}\sum_{i=1}^{n} m_i)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(m_i-\bar{m})^2}}$

Where: $m_i = lnL_{i,0} - lnL_{i,1}$, n = number of observations, $L_{i,0}$ = the i contribution to the likelihood function under the null hypothesis, $L_{i,0}$ = the i contribution to the likelihood function of the alternative model. The Null hypothesis is rejected if the Vuong statistic has a large positive value, and the alternative model is favoured (152).

5.2.3 The qualitative interview

As stated in Chapter 3 the qualitative interviews were conducted in Khon Kaen province to gain initial understandings on the coping mechanisms employed by the respondents engaging in the preference elicitation interviews. Recall that a convenient sample of ten respondents were interviewed using the same procedure as in the fieldwork interview except that only six states, rather than ten, were used. The reduction of number of health states was applied to minimise respondent workloads on giving their preferences on health states and to urge them to reveal their coping mechanisms on participating in the interview. The respondents were encouraged to "think aloud" on how they understand the questions and assign preferences on health. A content analysis was used to analyse the data.

5.3 Results

5.3.1 Logical inconsistencies in the Thai study

The Stata software was used to detect logically inconsistent responses. To identify the logically inconsistent responses, the appropriate pairs of health states for this task were first obtained. Then the scores or the ranks of those pairs were investigated to see whether the respondents ranked or assigned scores corresponding to the relative severity of the health states. The maximum pairs of health states that could be used to detect the logical inconsistent responses differed according to the health set and the elicitation method as shown in Table 5.2. The greatest number of potential inconsistencies is seen in Set 2, followed by Set 4 and Set 8. Note that the number of possible inconsistencies is fewer in the TTO method because only ten states were used whereas eleven states were used in the Ranking and VAS methods.

Table 5.2 Maximum numbers of logical inconsistencies according to health sets

Health set	Number of logical inconsistencies		
	Ranking	VAS	TTO
1	37	37	27
2	43	43	33
3	38	38	28
4	40	40	30
5	39	39	29
6	38	38	28
7	37	37	27
8	39	39	29
9	35	35	25
10	38	38	28
11	36	36	26
12	38	38	28

Mean inconsistency rate in Ranking method is 13.8% (SD 15.2), VAS 16.1% (SD 15.4) and TTO 25% (SD 18.3). Mean logical inconsistency rate according to health sets and the interview methods are presented in Table 5.3.

Table 5.3 Mean logical inconsistency rates by health set and the interview methods

Health set	Ranking		VAS		TTO	
	Mean	SD	Mean	SD	Mean	SD
1	16.4	16.5	19.3	15.6	27.3	17.3
2	13.2	14.2	16.4	15.1	23.7	17.0
3	14.1	16.7	16.1	15.4	24.7	16.8
4	14.1	14.5	15.1	13.6	27.2	19.5
5	14.2	16.0	15.6	16.1	26.6	21.0
6	17.6	18.6	17.2	17.5	26.9	21.6
7	13.5	12.6	15.3	14.4	25.7	16.2
8	11.8	13.4	17.5	16.6	22.9	16.8
9	11.8	13.8	13.5	11.8	20.5	16.5
10	11.7	12.1	15.4	16.0	24.7	16.2
11	15.6	16.2	16.6	16.1	25.7	17.7
12	12.8	14.9	16.7	14.7	24.1	16.9

The highest rates of inconsistency tended to be with Set 1 except in Set 6 where the inconsistency rate identified in the Ranking method is higher than that of Set 1. Among the three interview methods, the TTO method had the highest inconsistency rates. The SDs of the inconsistency rates of all three methods were slightly lower or higher than the means of the inconsistency rates. The number of logically inconsistent values

identified in the respondents interviewed by the "new" and "old" groups of interviewers, according to the three interview types, are compared in Table 5.4. A t-test was used to compare the differences in the mean number of logical inconsistencies between the "old" and "new" interviewer groups.

Table 5.4 Inconsistent values by interviewer group

Interview type	Mean number of logical inconsistency		
	new	old	p-level
Ranking	6.6	5.0	0.0001
VAS	7.2	6.0	0.0017
TTO	7.8	6.9	0.0120

Table 5.4 shows that the number of logical inconsistencies was significantly lower (at p-level =0.05) in the respondents interviewed by the "old" group of interviewers for all three elicitation methods.

The distribution of numbers of inconsistencies in Ranking, VAS and TTO methods are shown in Figures 5.1, 5.2 and 5.3 respectively.

Figure 5.1 Distribution of numbers of inconsistencies in Ranking

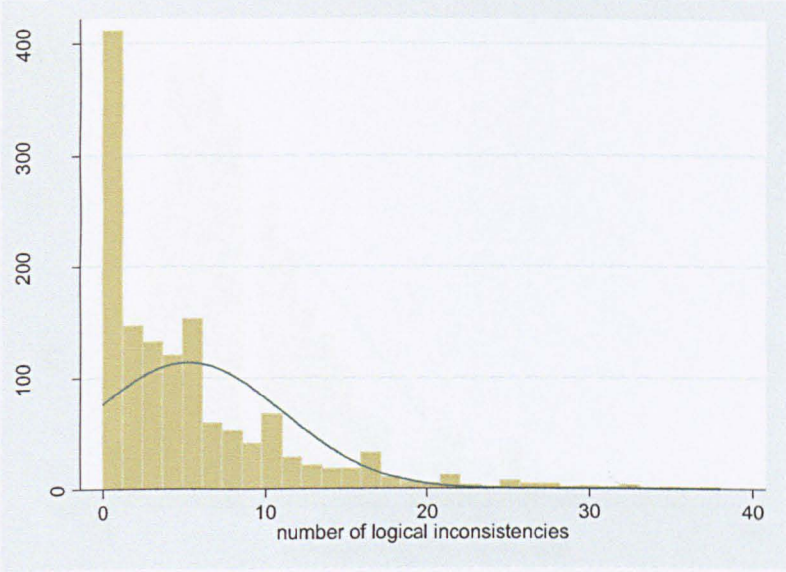


Figure 5.2 Distribution of numbers of inconsistencies in the VAS method

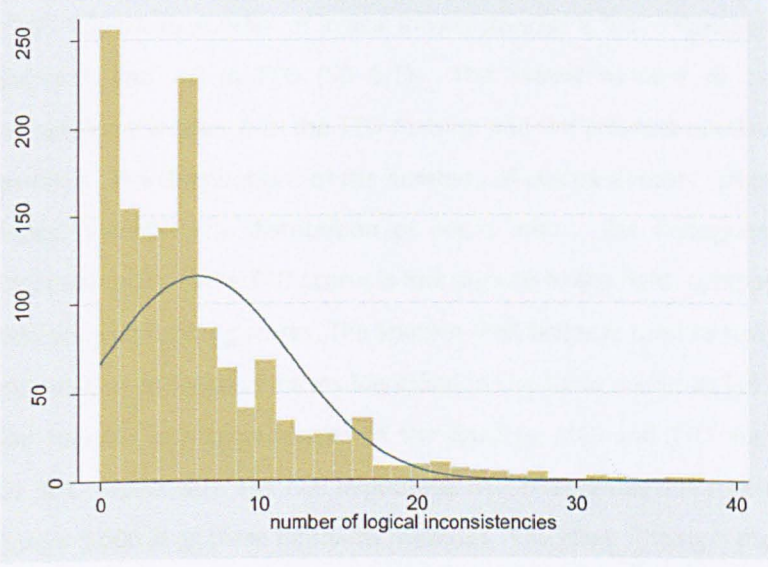
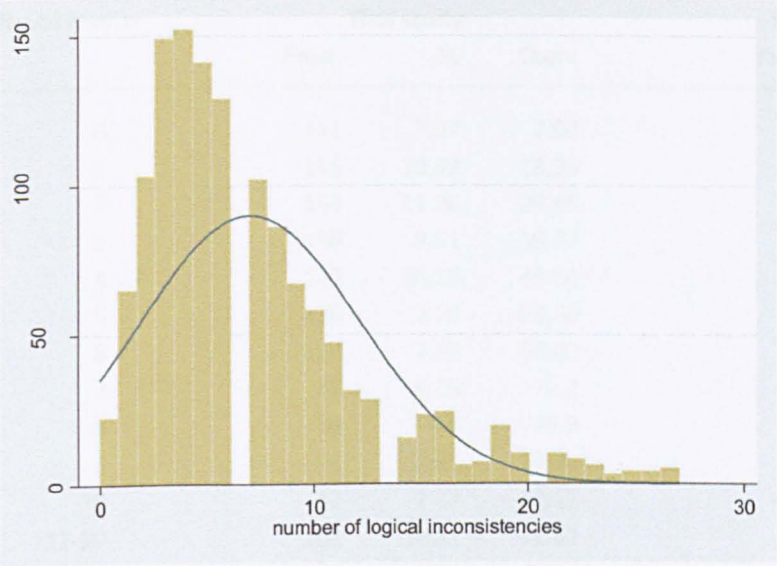


Figure 5.3 Distribution of numbers of inconsistencies in the TTO method



There were 204 completely consistent respondents with Ranking, 99 in VAS and 22 in TTO. The mean number of logical inconsistencies is 5.3 in Ranking (SD=5.8), 6.2 in VAS (SD=5.9) and 7.0 in TTO (SD 5.1). The lowest number of completely consistent respondents is seen with the TTO method and the greatest number is with the Ranking method. The distributions of the numbers of inconsistencies from the three interview types resemble the distribution of count data. The histogram of the number of inconsistencies in the TTO scores is less skewed to the right, compared with those of the VAS and the Ranking scores. The Shapiro-Wilk test was used to test the normality of the distribution of inconsistencies identified in the three methods (153). The Z-statistics of the number of inconsistencies in the Ranking, VAS and TTO are 12.103, 12.183 and 11.304 respectively. The null hypothesis that the distribution is normal is rejected at p-level = 0.000 in all three interview methods. The other finding is that when the numbers of inconsistencies increase, the probability of zero count of inconsistency decreases. Inconsistency rates in this study were compared with those obtained from the NZ and NL studies and presented in Table 5.5 and Table 5.6 respectively.

Table 5.5 Comparison of numbers of inconsistencies between Thai and New Zealand studies

No.of incons	Thai study			New Zealand study		
	Freq.	%	Cum.	Freq.	%	Cum.
0	111	7.97	7.97	189	20.57	20.57
1	145	10.42	18.39	207	22.52	43.09
2	154	11.06	29.45	137	14.91	58.00
3	138	9.91	39.37	98	10.66	68.66
4	143	10.27	49.64	56	6.09	74.76
5	126	9.05	58.69	48	5.22	79.98
6	102	7.33	66.02	30	3.26	83.24
7	86	6.18	72.2	18	1.96	85.20
8	64	4.6	76.8	17	1.85	87.05
9	41	2.95	79.74	14	1.52	88.57
10	33	2.37	82.11	12	1.31	89.88
11-20	202	14.51	96.62	50	5.44	95.32
21-30	37	2.66	99.28	22	2.39	97.71
31-40	10	0.71	99.99	13	1.41	99.13
Total	**1,392**	**100.00**		**911**	**100.00**	

(116)

No. of incons= number of logical inconsistencies

The numbers of inconsistencies obtained from the VAS method in the two studies are shown in this table. Only those from the VAS method was shown here because the postal VAS survey was used in the NZ study. The proportion of the respondents with completely consistent values was higher in the NZ study. Eighty percent of the NZ respondents had fewer than 6 inconsistent scores, compared with only 60% in the Thai study. The number of respondents with fewer than two logical inconsistencies is lower in the Thai study than in the NZ study. Note that the comparison in Table 5.5 was performed in order to examine the possibilities of the respondents being inconsistent, rather than to identify the definite differences of the number of inconsistencies between the two studies, because the health states used in the interviews in the NZ study were different from those used in the Thai study. There would be a higher chance of assigning inconsistent values to a pair of health states if the differences between the two states are small, for example, state 33323 and state 33223. The number of inconsistencies in the Thai study could be higher simply because the health states used to interview the Thai respondents are very similar, thus difficult for respondents to differentiate between two health states. The findings in Table 5.5 could be interpreted

as suggesting that the Thai respondents have a higher chance of being inconsistent than the respondents in the NZ study.

Table 5.6 Comparison of the inconsistency rates between the Thai and the Netherlands (NL) studies

Inconsistency rate	Cumulative proprotion of respondents	
	Thai	NL
0	1.6	10.7
0-8.3	14	31.5
0-16.7	38.3	51.6
0-25	63.9	70.7
0-33	76.5	82.1
0-42	85.8	91.2
0-50 and more	90.7	99.9

(139)

Note that the proportions of respondents shown in Table 5.6 are presented as cumulative proportions because the numbers of respondents in each category of inconsistency rate cannot be obtained from the published paper of the NL study. The definition of inconsistency in Table 5.6 follows Ohimaa and Sintonen. The inconsistency rates were obtained from the TTO elicitation interview. Compared with the respondents in the NL study, a lower proportion of the Thai respondents were completely consistent. Half of the respondents in the NL study had the inconsistency rate lower than 20% whereas only 40% of the respondents in the Thai study had the inconsistency rate lower than 20%. The inconsistency rate is higher in the Thai respondents than in the Dutch respondents. Similar to the comparison in Table 5.6, the comparison in Table 5.6 was conducted to examine only the possibilities of the respondents assigning logically inconsistent values to health states in the two studies.

5.3.2 Factors associated with inconsistent responses

Definitions of the independent variables and the data of each variable are presented in Table 5.7. Interviewer was also used as an independent variable I_i where i represents interviewer no. 1 – 48. Interviewer no.12 is used as a reference because the number of interview conducted by this interviewer is greatest.

Using the PRM to fit the data from the three methods, the models suffer from over-dispersion (1/df Pearson is 5.80 and 1/df Deviance is 4.95 for the ranking data, 4.37 and

4.00 for the VAS data and 4.38 and 4.00 for the TTO data). Thus NBRM is used to account for the over-dispersion. The values of 1/df Pearson and 1/df Deviance from three groups are lowered to approximately 1. LR tests support that NBRM is favoured over PRM for the three methods. LR test is 3023.31 (p-value=0.000) for the ranking data, 2056.07 (p-value=0.000) for the VAS data and 1337.32 (p-value=0.000) for the TTO data. To account for the number of completely consistent respondents, ZIP and ZINB are also used to fit the model. The Vuong test cannot reject the null hypothesis that the NBRM is favoured over ZINB 0.88 (p-value=0.190) for the TTO data. AIC and BIC statistics are smallest using NBRM. The NBRM model is, therefore, the best model to predict numbers of inconsistent pairs for the three methods.

Table 5.7 Summary statistics for the independent variables

Variable	Definition	Elicitation methods		
		Ranking mean(SD)	VAS mean(SD)	TTO mean(SD)
number of observations		1,392	1,393	1,324
age	respondent age in years age is treated as a continuous dat	44.2(12.5)	44.2(12.5)	44.2(12.5)
secondary	a dummy variable 1 if the respondent's highest education level is secondary 0 otherwise	0.2(0.4)	0.2(0.4)	0.2(0.4)
university	a dummy variable 1 if the respondent's highest education level is university 0 otherwise	0.1(0.3)	0.1(0.3)	0.1(0.3)
gen	a dummy variable for gender 1=female 0=male	0.55(0.5)	0.55(0.5)	0.55(0.5)
no. of children	a continuous data number of children	2.3(1.4)	2.3(1.4)	2.3(1.4)
duration	a continuous data duration of the overall interview in minutes	56.1 (19.7)	56.1(19.7)	56.1(19.7)
ownVAS	a continuous data self-rated VAS score for own health	79.4(15.7)	79.4(15.7)	79.4(15.7)
urban	a dummy variable for residential area 1=urban area 0=rural area	0.3(0.5)	0.3(0.5)	0.3(0.5)
dummy variables for health sets				
hset2	a dummy variable for health set 2	0.09(0.29)	0.09(0.29)	0.09(0.29)
hset3	a dummy variable for health set 3	0.09(0.28)	0.09(0.28)	0.09(0.28)
hset4	a dummy variable for health set 4	0.09(0.28)	0.09(0.28)	0.09(0.28)
hset5	a dummy variable for health set 5	0.09(0.28)	0.09(0.28)	0.09(0.28)
hset6	a dummy variable for health set 6	0.08(0.27)	0.08(0.27)	0.08(0.27)
hset7	a dummy variable for health set 7	0.08(0.27)	0.08(0.27)	0.08(0.27)
hset8	a dummy variable for health set 8	0.09(0.28)	0.09(0.28)	0.09(0.28)
hset9	a dummy variable for health set 9	0.08(0.27)	0.08(0.27)	0.08(0.27)
hset10	a dummy variable for health set 1	0.07(0.26)	0.07(0.26)	0.07(0.26)
hset11	a dummy variable for health set 1	0.08(0.27)	0.08(0.27)	0.08(0.27)
hset12	a dummy variable for health set 1	0.07(0.26)	0.07(0.26)	0.07(0.26)

* health set 1 is used as a reference

Table 5.8 Results of model analysis using data from 3 elicitation methods

Variable	Ranking			VAS			TTO		
	coeffi	SE	p-value	Coef.	SE	p-value	Coef.	SE	p-value
age	0.016	0.003	0.000	0.018	0.002	0.000	0.010	0.002	0.000
secondary	-0.179	0.081	0.027	-0.248	0.063	0.000	-0.161	0.051	0.002
uni	-0.696	0.101	0.000	-0.698	0.081	0.000	-0.234	0.062	0.000
hset9				-0.236	0.088	0.008	-0.351	0.073	0.000
I6	-0.467	0.202	0.021						
I7									
I10				-0.300	0.108	0.006			
I17							-0.206	0.090	0.022
I23				-0.462	0.232	0.047			
I24				-0.368	0.189	0.051			
I25				-0.543	0.207	0.009	-0.208	0.091	0.029
I31	-1.003	0.244	0.000	-0.349	0.177	0.049			
I33							-0.363	0.166	0.036
I34				-0.438	0.167				
I35	-0.806	0.391	0.039	-1.071	0.330				
I36	-0.887	0.254	0.000						
I38	-0.788	0.290	0.007						
I39				-0.726	0.228	0.001			
I41				-0.456	0.214	0.033			
I47	-1.996	0.985	0.043	-2.830	1.129	0.012			
_cons	1.058	0.130	0.000	1.186	0.100	0.000	1.621	0.080	0.000

Coefficients of the NBRM for the three elicitation methods are presented in Table 5.8. Only statistically significant coefficients are presented.

For the three elicitation methods, the number of inconsistent pairs increases by the factor of 1.01 in respondents an additional year older. Compared with the respondents who valued Set 1, those who valued Set 9 by VAS, the numbers of inconsistencies are decreased by the factor of 0.78, and 0.70 by TTO. Health set has no significant effect on the numbers of inconsistencies in the Ranking method.

Compared with those who have primary level education (less than seven years in school), the respondents who have highest education at university level (more than twelve years in school) have lower number of inconsistent pairs by the factor of 0.50 in Ranking, 0.50 in VAS and 0.79 in TTO. Those who completed secondary level (7-12 years in school), have lower number of inconsistent pairs by 0.84 in Ranking, 0.78 in VAS and 0.85 in TTO.

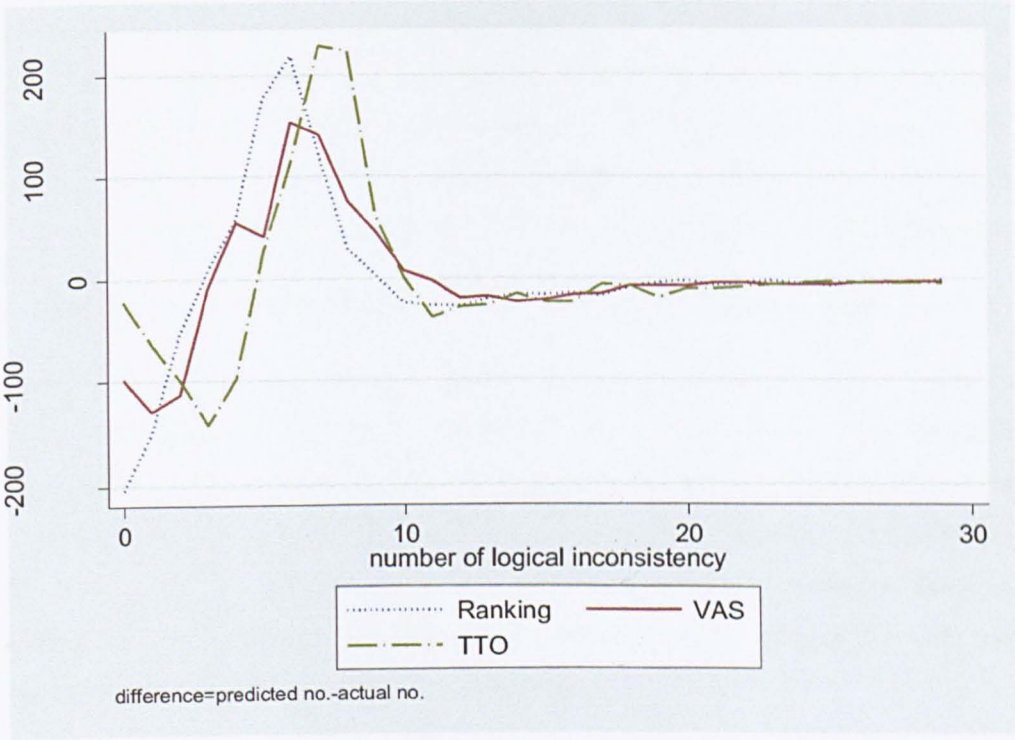
Compared with Interviewer no. 12, in Ranking, those who were interviewed by Interviewer no. 6, 31, 35, 36, 38 and 47 have decreased number of inconsistent pairs by the factor of 0.47, 1.00, 0.81, 0.89, 0.79 and 2.00 respectively. In VAS, the respondents who were interviewed by Interviewer no. 10, 23-25, 31, 34, 35, 39, 41 and 47 have a lower number of inconsistent pairs by the factor of 0.30, 0.46, 0.37, 0.54, 0.35, 0.44, 1.10, 0.73, 0.46, and 2.83 respectively. In TTO, the respondents who were interviewed by Interviewer no. 17, 25 and 33 have lower number of inconsistent pairs by the factor of 0.21, 0.21 and 0.36 respectively. Note that the old interviewers are Interviewer no. 1-11, 17-18 and 21-48.

The three resulting models are used to predict the number of inconsistent pairs for the respondents. A line graph comparing the differences between predicted and actual number of health state pairs with inconsistent responses is presented in Figure 5.4. The predicted, actual numbers of pairs and their differences are presented in Appendix 6.

From Ranking method (dotted line), the model underpredicts (predicted<observed frequency) the logical inconsistencies ranging from 1-2 inconsistencies, and overpredicts (predicted>observed frequency) the inconsistencies from 4-8. From 9-11 inconsistencies, the model tends to perform better with the differences between the number of respondents with actual and predicted pairs of corresponding inconsistent

responses is less than 31 respondents. The model does not predict cases with more than 11 inconsistencies.

Figure 5.4 Comparison of the differences between the predicted and actual number of logical inconsistencies



From VAS data (dashed line), the model underpredicts the number of logical inconsistencies with the range from 0-3 inconsistencies, and overpredicts with the range from 5-11 inconsistencies. The difference between the frequency of predicted and actual inconsistencies is small (<30 respondents) at the 10-14 inconsistencies. The model cannot predict the respondents with 15-17 and higher than 18 inconsistencies.

From TTO method (solid line), the model underpredicts the frequency of predicted logical inconsistencies between 0-4 inconsistencies and 11-12 inconsistencies. The model performs better when the number of predicted inconsistencies is larger than 9 (differences are smaller than 35). However, the model cannot predict the respondents with zero inconsistencies and more than 12 inconsistencies.

5.3.3 Results of the qualitative study

Ten respondents were interviewed; five of them were male. The average age of all respondents was 55.5 years. All of them had a primary education attainment level. As stated in Chapter 3, to assign preference scores to health, a considerable level of literacy and numeracy would be needed. Respondents would take into account extra-information outside health attributes to trade-off time living in full health and some may use simple heuristic approaches to express their preferences on health. The analysis revealed that three major themes regarding the understanding of the health descriptions and their coping mechanisms with the interview tasks emerged as follows.

1. Difficulties in perceiving health states and following the interview tasks

Almost half of the respondents had difficulties in imagining themselves living in hypothetical health states. Some respondents were unable to understand the tasks and what was asked of them. Some refused to believe that, after living in a state worse than death for some period, they would recover completely and stay in perfect health for some years. As seen in the following comments, some respondents failed to differentiate between two different health states in the beginning of the task, but developed an understanding later in the interview.

> *"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?"*

2. Incorporating extra-information beyond the health attributes or taking only parts of health dimensions into consideration

Some respondents used non-health related information to contribute to their decision making regarding trading-off time. After reading the cards, the respondents may not have only considered their life in the particular health state, but also included their family members in the scenarios. For example, the respondents were concerned about the family members who were going to have to take care of them if they had to be confined to bed. They were worried that nobody would help them if they were ill. So if there was a question asking to trade-off time of the health state with 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 for my medicines?"*
>
> *"I don't want someone to take care of me when I stay in bed. By causing burdens on my children, it is sinful.I"*
>
> *"If I have a good family, when I fell sick, my grand children would come to take care of me."*

Some respondents wanted to live as long as possible because they want to see their children (and grand-children) grown up and living their lives. In contrast, some participants may have used only partial information from health cards or applied a simple heuristic approach in assigning preference scores. They may have considered only the "key" dimension (154). 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 at level 1, no matter what the levels are for other dimensions, they may assign high values for those states without taking other dimensions into account. 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 in which they do not want to suffer, especially 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"*

3. Difficulties in understanding the TTO question for states worse than death

These difficulties could result from the reading difficulties identified in the first theme. In the TTO question for health states worse than death, where the respondents were asked to choose between Health state A: living in poorer health states for some period of time before living in full health and Health state B: Immediate dead, it was not uncommon for respondents to immediately choose to live in perfect health without taking into account that she has to live in the poorer 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 experience.

5.4 Discussion

This chapter reports logical inconsistencies. Dolan & Kind's definition was applied to identify numbers of inconsistencies in the scores obtained from the three elicitation interviews. A greater number of inconsistent responses were seen in the scores obtained from the TTO interview. The proportions of completely consistent respondents are greatest in the scores elicited from Ranking following by those elicited from VAS methods, because the required tasks in Ranking and VAS are simpler compared with the TTO task. Respondents also had opportunities to review and change their ranks or scores of all health states at the end of the tasks if the ranks or scores did not correspond to their preferences. In TTO, however, the respondents were asked to complete the task without an opportunity to review their scores upon completion. Therefore, they would not know or remember the scores given to the previous states. Moreover, the TTO interview followed after the Ranking and VAS interviews had been completed. The respondents could have been exhausted from the first two tasks. In

addition to the complexity of the TTO task this may explain the higher number of logical inconsistencies in the TTO scores.

The inconsistency rates in the Thai study are lower than that of the Spanish study. As reported in Section 5.1, the mean inconsistency rates reported in the Spanish study are 24.4, 25.9 and 59.2, compared to 13.8, 16.1 and 25 in Ranking, VAS and TTO elicitation methods, respectively, in the Thai study. The reason could be because the numbers of health states used in the Thai and Spanish studies differed. In the Thai study, up to eleven states were used in Ranking and VAS and ten in TTO interview methods. In the Spanish study, thirteen health states were valued. The Spanish respondents may have faced a greater cognitive burden because they had to "work" with more states than respondents in the Thai study, thus leading to more logically inconsistent scores.

The Thai respondents were likely to have higher numbers of inconsistencies compared to the respondents in the Netherlands and New Zealand studies. Note that the scores from the New Zealand studies were obtained from a postal VAS survey which was different from the survey conducted in the Thai study. The causes of higher inconsistencies in the Thai data could be that the respondents were, on average, older and had lower education levels than those in the Netherlands study. The Dutch respondents had previous experience in participating in preference elicitation interviews whereas the Thai respondents had no experience in preference elicitation interviews. A computer-based program was used in the Netherlands study which then could avoid the interviewer effects present in the Thai study. These may explain the lower numbers of inconsistencies in the Netherlands study.

The relationships between number of inconsistent responses and respondents' demographic characteristics were systematically explored in the statistical analysis. This appears to be the first study using count data models to analyse the factors influencing numbers of logical inconsistencies. NBRM is the best fitting model for the data from the three elicitation methods. The models estimated for each of the three methods tend to underpredict the frequency of respondents with less than five inconsistent pairs and overpredict in the range of 5-10 inconsistencies. The models also tend to under-predict the number of participants whose responses imply more than ten inconsistencies. A particular weakness is that they do not predict that any individuals will display very high numbers of inconsistency, namely more than twelve inconsistencies (TTO), more than eighteen inconsistencies (VAS) and more than eleven inconsistencies (Ranking). Other

relevant explanatory variables (if available) might be added to account for the highly inconsistent responses, for example, whether the interview was free from interruptions or distractions.

Age, education level, and interviewer effects were observed, confirming findings of other researchers. Badia *et al.* and Devlin *et al.* reported that age and education have significant effects on logical inconsistency in that older respondents and respondents with lower education tend to exhibit greater inconsistency. Badia *et al.* and Jelsma *et al.* also reported interviewer effects. In the Thai study, the interviewers could be categorised into the "new" and "old" groups of interviewers. As stated in Chapter 3 and 4, because of the unavailability of the interviewers previously trained in the beginning of the survey (old group), a new group of interviewers were recruited, trained and conducted the interviews at approximately the second half of the survey. It was also shown in Chapter 4 that the "new" interviewers tended to interview the respondents who were slightly younger than those interviewed by the "old" interviewers. The analysis in this chapter showed that numbers of logical inconsistencies identified in the respondents interviewed by the "new" interviewers were significantly higher in the three interview tasks. Interview durations were also shorter (shown in Chapter 4). One reason could possibly be that respondents tended to want to complete the interview as quickly as possible and the "new" interviewers, being less experienced, might have had difficulties trying to "slow down" the respondents and encouraging them to take all attributes of health into consideration. This is only one hypothesis attempting to explain this finding. More evidence is required before drawing definite conclusions.

This study also reports a new finding that different combinations of health states are associated with different levels of inconsistency. This was observed despite attempts to make the twelve health sets comparable in terms of the mix of severity of health states. Set 9 seems to generate fewer inconsistent pairs. Set 9 may be comprised of health states that respondents find easier to differentiate between, thus, fewer inconsistent pairs are seen. More consideration should be paid to the selection of health states in future elicitation studies.

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. It may have been

difficult for the respondents, especially the elderly with primary education, to imagine themselves in hypothetical health states and understand such a complicated question as used in the TTO interviews. These findings identified, according to the conceptual framework proposed by Lenert and Kaplan (129) as shown in Figure 3.4 , that health state descriptions and elicitation procedures have some degree of influence on the coping mechanisms employed by the respondents engaging in the interview. However, other factors, for example: emotion and prejudices; and utility and risks were not identified in this study. Future studies should be designed to explore these issues further.

The findings are also in line with the study by Miguel *et al.* (147) in that the elderly respondents with primary education tended to have difficulty imagining living in the health states given in the interview. This may have resulted from the lower level of reading competency in this group of respondents. If they had 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. What we have learnt from the qualitative study could lead to increased understanding of why logical inconsistency or "irrational responses" were common in the Thai study. It is possible that the respondents assigned scores according to their understanding of health states. If they had poor understanding concerning both the health states and the interview tasks, especially in the beginning of the interview, they may have had assigned scores that did not accord to their preferences. Later on, after they were able to "get it right" or acquire more understanding of the tasks, then they would assign scores more consistently with their preferences for these health states. Note that the qualitative analysis in this study was treated as an exploratory study from which the results can be used to generate hypothesis for future qualitative studies.

Results from this study suggest a number of ways in which it might be possible to reduce logical inconsistency 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, computer-assisted interviews with a prompt when logical inconsistency is identified may reduce the number of logical inconsistencies. 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 to avoiding combinations of states that might be particularly hard for respondents to differentiate between. Plausibility of health states should be taken into account.

While it is important to understand what factors may be associated with higher or lower levels of inconsistency with a view to collecting more consistent data, such insights might also assist decision making over data exclusion. Including data from inconsistent respondents is likely to affect the mean utility scores for health states. The next chapters will explore the impact of exclusion of respondents with differing 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.

5.5 Conclusion

As in the studies in other countries, logical inconsistency is identified in the responses given by the Thai general population. Age, education level, combinations of health states, interviewers and types of elicitation method have significant effects on the extent of logical inconsistencies. Inconsistencies are more likely to occur with the more complex interview methods. The negative binomial regression model best fit the data collected by the three different methods. Older respondents and those having completed only primary education exhibited higher numbers of inconsistent values. Interviewers and combinations of health states are associated with the extent of logical inconsistencies. Those interviewed by the "new" interviewer group were likely to have higher inconsistencies in the three interview tasks. From the qualitative study, it was shown that respondents tended to use both additional information apart from the health state card and partial information or a simple heuristic approach to assist the decision over the trading-off of time. Older respondents had difficulties in understanding the health descriptions and the complex interview questions. Cognitive overload and a learning process may play some role in responding to the interview. Results from this study can be used as a guide to develop interview methods to minimize the extent of logical inconsistency in future preference studies.

Chapter 6 Effects of logical inconsistency on preference scores

6.1 Introduction

The previous chapter explored the possible causes of, and the extent to which, logical inconsistencies were displayed by the Thai respondents. 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 respondents, due to the fact that they may not have been able to understand the preference elicitation tasks and therefore randomly assigned scores to the health states. The extent to which inconsistent respondents should be excluded from the model estimations depends partly on the impact of these inconsistencies on the mean scores. This chapter sheds light onto these effects. The outline of this chapter is as follows: first, literature on the treatment of inconsistency is briefly reviewed. Second, the methods used to examine the effect of excluding data from respondents exhibiting differing levels of inconsistency are described. Third, the effect of the exclusion of logical inconsistencies is thoroughly explored. Possible causes of inconsistencies are also further investigated on the basis of the findings from the previous chapter. The most appropriate group of respondents to use when estimating the Thai tariff is chosen at the end of the chapter.

6.2 Literature review

There is no agreement regarding the recommendations on the treatment of inconsistent responses. On one hand, a so-called "inconsistency threshold" is introduced and respondents exhibiting a level of inconsistency beyond the threshold level are excluded from the estimation of preference scores. Ohinmaa and Sintonen recommend that the data from respondents with more than three inconsistencies are excluded (138). In the Korean study by Jo *et al.*, 12 respondents (2.4% of total respondents) with more than three inconsistencies were excluded without explicit discussion as to why these respondents were excluded (121). Presumably, the authors

of the Korean study were following the recommendations of Ohinmaa & Sintonen. It is not clear which definition of inconsistency was adopted in the Korean study, and no analysis of the implications of the inconsistent responses on the mean health state scores was presented.

On the other hand, as recommended by Lamers *et al.*, all inconsistent values should be included in the model estimations (139). Lamers *et al.* used the Ohinmaa & Sintonen definition to identify inconsistency and argued that by excluding the scores from the inconsistent respondents, the representativeness of the sample was likely to be affected. The implications of inconsistent scores on mean scores are investigated using the interview-based VAS and TTO tasks. Differences between utility scores of the pairs of health states with inconsistent values exceeding 0.1 were identified, whereas differences smaller than 0.1 were argued to result from measurement errors. Sixty-five per cent of the respondents were inconsistent in response to the VAS task and eighty-nine per cent with respect to the TTO task. To see the effects of inconsistent values on mean VAS and TTO scores for health states, the respondents were divided into groups based on the number of inconsistencies. Three groups were classified for VAS values. The first group was the completely consistent respondents, the second was the respondents with one to three inconsistently values and the last group was those with four or more inconsistent values. Four groups were formed for TTO values: the first group was the completely consistent respondents and those with only one inconsistency. Those with two or three, four or five and with six or more inconsistent values were classified into the second, third and forth groups respectively. Mean VAS and TTO values of each health state were compared between the groups using a t-test. For VAS values, the differences in mean scores of nine out of seventeen states (53%) were statistically significant between the groups. Six states out of seventeen (35%) had statistically significant differences in all mean scores across the various groups of inconsistencies in TTO values. To see the effects of removing the scores from highly inconsistent respondents, the authors estimated the models using the scores from all respondents and again with only the respondents with 0-5 inconsistencies (8% of respondents excluded). The coefficients from the two groups were not significantly different; R-squared was slightly increased after removing the inconsistent responses, and the MAD was slightly smaller. The authors concluded that the exclusion of scores from inconsistent respondents made no difference and that the model is robust to be able to estimate preference scores although the inconsistent responses are included.

An alternative treatment was offered by Devlin *et al.* (116) in the New Zealand study where a postal VAS survey was conducted using thirteen health states. The definition of inconsistency followed Dolan & Kind (140). Two hypotheses regarding the interpretation of logical inconsistency were stated: the inconsistent scores could be regarded as the values genuinely representing respondent's values on health states; or they could be the scores incorrectly representing health state values because the respondents "crucially failed to understand the task". Two groups of respondents were distinguished in the analysis: all respondents and highly consistent respondents (with 0-1 inconsistencies). The health state rankings of the two groups were highly correlated. The authors concluded that there is no generalisable rule to guide the extent to which the inconsistencies should be excluded when modelling preferences. Researchers should report the effects on the results of excluding the inconsistent responses and the treatment of inconsistency relies on the researcher's judgment. Badia *et al.* also support this recommendation (141). The recommendations by Devlin *et al.* will be followed in this analysis; the effects of logically inconsistent responses on the scores will be thoroughly explored before using this analysis as evidence to support a decision on the number of inconsistencies that will be used as grounds for exclusion.

6.3 Methods

6.3.1 Examination of the validity of the scores

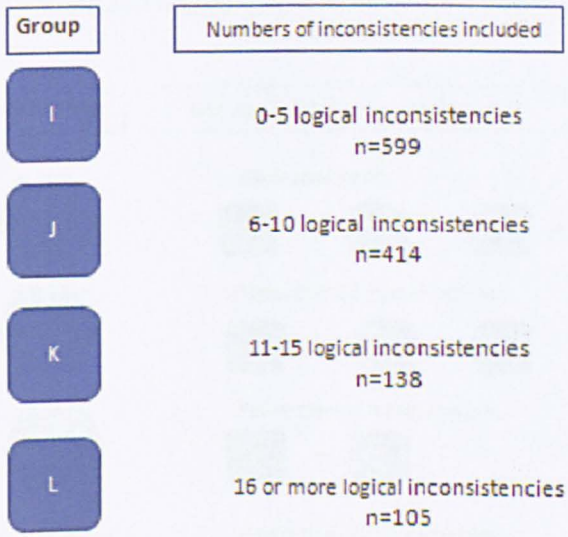
The TTO scores are used in the analysis of the effects of exclusion because the Thai preference scores are to be estimated from the TTO data. The validity of the scores is examined based on the assumption that the respondents with fewer inconsistencies are those who can assign TTO scores according to the severity of health states, i.e., with higher scores assigned to better states and lower scores to poorer states. It was not an objective in this study to identify logically inconsistent values at the health state level, therefore the method used in the study was not designed for this purpose.

Recall that the inconsistent values in this analysis were identified in the individual respondents or at the health set level, rather than at the health state level. As stated in Chapter 3, the total eighty-six health states classified into twelve health sets were administered in the fieldwork interview, one set was interviewed per respondent. Almost one-fourth of the health states were allocated to more than one health set, for

example, state 11211 was allocated to Sets 1, 6 and 11 and state 21332 to Sets 1 and 11. To be able to identify logical inconsistency in this study, two characteristics are required. Firstly, the pairs of health states which are possible to identify inconsistency and secondly, these pairs must be part of the same health set, with the scores assigned by the same respondents. Therefore, to identify inconsistent values using states 11211 and 21332, only values from the respondents assigning values to Set 1 could be used. Inconsistency in this study was not examined at the health state level, for example, it was not possible to compare mean scores of states 11211 and 21332 to identify logical inconsistency because the values were assigned using different health sets, thus, some of the scores were observed from different groups of respondents.

The extent to which logical inconsistencies occurred in individual respondents was used as a guideline to classify the respondents into several subgroups. To examine the validity of the scores, the respondents were divided into four groups according to the number of logically inconsistent responses, as presented in Figure 6.1. 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 6.1 Four respondent groups with various numbers of inconsistencies



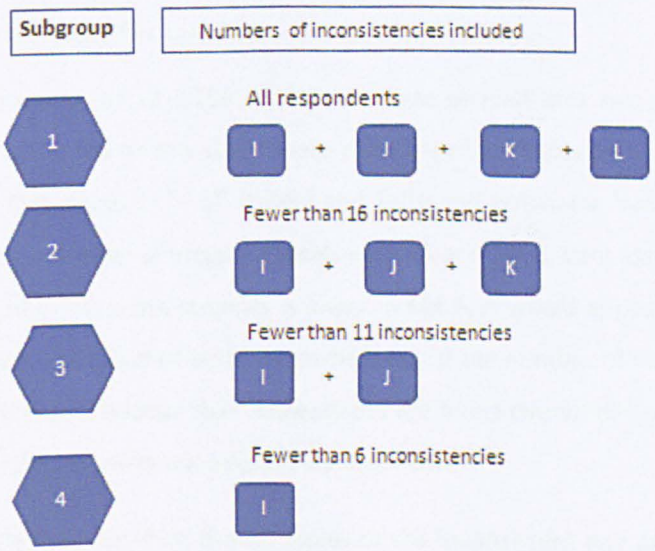
n=number of respondents

Out of 1,256 respondents who have complete demographic data and assigned scores using the TTO interview, almost fifty percent had fewer than six inconsistent values (Group I) and fewer than ten percent had more than fifteen inconsistent values (Group L). 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 respondents in this group assigned the scores with the least inconsistency. These scores are assumed to be a potentially robust basis upon which to base the EQ-5D tariff for Thailand.

6.3.2 The examination of the impact of excluding data from inconsistent respondents

To explore the effect of excluding the inconsistent scores 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, are 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 form subgroup 4. The four subgroups of respondents thus generated are shown in Figure 6.2.

Figure 6.2 Four respondent subgroups and numbers of inconsistencies included



Subgroup 1 is regarded as a reference group because it contains the scores from all respondents who participated in the study. Mean scores for Subgroups 2-4 were compared with that of Subgroup 1 and the differences in mean scores from each subgroup were explored. Spearman rank correlation coefficients were 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 exhibited more logical inconsistencies. Demographic characteristics and average interview durations are reported for the different subgroups. These factors could be utilized when trying to justify the exclusion of scores from inconsistent respondents.

6.3.3 Possible causes of logical inconsistency

As noted in the exploratory qualitative study in the previous chapter, respondents may learn how to respond to the TTO task during the initial questions and they may become increasingly tired as a result of tackling the complex tasks involved in the preference elicitation interviews. The analysis in this section aims to explore the hypothesis that respondents with fewer inconsistencies exhibit a "learning effect", where the skills have been developed in the beginning of the task, in this case within the first five questions. The respondents then apply these skills when responding to subsequent questions. Whereas, the respondents exhibiting more inconsistency are becoming tired, are, therefore, less able to concentrate, less able to learn, and in short are being "overwhelmed" by the task. In this case, the numbers of inconsistencies should be higher in the latter five questions.

The TTO scores of all 1,256 respondents are divided into two sets according to the order in which the health states were ranked by the respondents. Set A comprises the first five TTO scores (1^{st} - 5^{th} states) and Set B the remaining five TTO scores (6^{th} - 10^{th} states). The number of inconsistencies in Set A and Set B were identified and compared. If the number of inconsistencies is lower in Set B, it would appear that learning effects outweighed the effect of being overwhelmed. If the number of inconsistencies is higher in Set B, it would appear that respondents are being overwhelmed by the task, and any learning effect is being outweighed by tiredness.

To identify inconsistency, if both states of the inconsistent pair are either in Set A or in Set B, the logical inconsistency can be identified as having occurred in Set A or B. A problem arises when one state of the pair is in Set A and the other is in Set B. In this

case, the inconsistency cannot be identified as definitively belonging to either Set A or Set B. These are classified here as Set C inconsistencies.

Finally, at the end of the chapter, the findings are used to make a decision as to which respondents (if any) to exclude due to their demonstrated inconsistency. The chosen respondent subgroup is used in the model analysis in Chapter 7.

6.4 Results

6.4.1 Demographic characteristics of the respondents in all four subgroups

Demographic characteristics of the respondents in the four subgroups (subgroups 1-4) are compared with those of the Thai general population in Table 6.1. Compared with the general population, all subgroups have higher proportions of females and respondents living in urban areas, and a lower proportion of elderly people. Note that subgroup 4 has a lower proportion of respondents with a primary education level compared with that of the general population.

6.4.2 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 from the respondents in Groups I-K are reported in Table 6.2. The mean scores were ranked from highest to lowest relative to the scores in Group I. The mean scores of health states in Group J, Group K and Group L were 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 were made: Group J-Group I; Group K-Group I and Group L-Group I. A t-test was 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 percent (21 states) of the mean scores of respondents in Group J are significantly different from the mean scores of the corresponding health states from respondents in Group I. Almost 40 percent (34 states) of the scores from the respondents in Group K, and 45 percent (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 6.1 Demographic characteristics of the respondents in all four subgroups

Participant characteristics	Thai general population** (x 1,000,00	%	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	606	45.77	600	46.01	498	46.37	299	47.31
Female	31.82	50.67	718	54.23	704	53.99	576	53.63	333	52.69
Mean age yrs. (SD)	32.8		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
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

** 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 6.2 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

* statistically 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.

Table 6.2 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*

* statistically significant difference from Group I (p-level <0.05 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.

6.4.3 Mean scores of the four respondent subgroups

Numbers of observations and mean actual scores of all 86 health states in Subgroups 1-4 are estimated and shown in Table 6.3. Mean scores were ranked from highest to lowest according to 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 for state 13123 in Subgroups K or L.

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, the 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 6.3 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 6.3 Mean scores of health states after excluding scores from the inconsistent respondents (continued)

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)

6.4.4 Spearman rank correlation coefficients

As stated in Section 6.3.2, Subgroup 1 is regarded as the reference subgroup. To see whether the rank of mean scores of the health states changed after excluding the scores from the respondents with various numbers of inconsistent values, Spearman rank correlations were used. The scores from all subgroups are ranked from the highest to 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 6.4.

Table 6.4 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.

6.4.5 Identification of the possible causes of logical inconsistencies

The number of inconsistencies in Sets A, B and C are shown in Table 6.5. Recall that Set A represents the numbers of inconsistencies that arose in the first half of the TTO interview, Set B represents those in the second half and Set C represents those which cannot be categorised into Set A or Set B.

Table 6.5 Numbers of inconsistencies in Sets A, B and C

No. of inconsistencies	No. of respondents	Total no.of Inconsistencies Set A	Set B	Set C	No.of inconsistencies per respondent Set A	Set B	Set C
0	22	0	0	0	0	0	0
1-5	610	457	450	1124	0.723	0.712	1.778
6-10	442	707	697	1955	1.600	1.577	4.423
11-15	144	394	373	1041	2.736	2.590	7.229
16-20	66	238	254	681	3.606	3.848	10.318
21+	40	174	196	565	4.350	4.900	14.125
Total		**1970**	**1970**	**5366**			

Set A=the first five states
Set B=the second five states
Set C= cannot be identified when the inconsistencies occurred

Twenty-two respondents were completely consistent and 40 respondents had more than 20 inconsistencies. The number of inconsistencies in Sets A, B and C are shown in the 3rd -5th columns. The average number of inconsistencies in Sets A, B and C are reported in the 6th -8th columns. In total, the number of inconsistencies in Set A is 1,970, Set B is also 1,970 and Set C is 5,366. On average, respondents had 1.49 inconsistencies (1970/1324) in Set A and Set B, and 4.04 inconsistencies (5366/1324) in Set C. Of all 632 respondents who had fewer than six inconsistencies, 457 inconsistencies occurred in Set A (0.72 per respondent), 450 inconsistencies occurred in Set B (0.71 per respondent) and 1,124 inconsistencies in Set C (1.78 per respondent). Forty respondents who had more than twenty inconsistencies had 174 inconsistencies in Set A (4.35 per respondent) and 196 inconsistencies in Set B (4.90 per respondent) and 565 inconsistencies in Set C (14.13 per respondent). Note that for the respondents with fewer than 16 inconsistencies, the total numbers of inconsistencies identified in Set A were slightly greater than those identified in Set B.

6.5 Discussion

This chapter explores 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). In this study, logical inconsistency has been identified using pairs of health states from the same health set. Also, logically inconsistent values were identified by higher scores being given to poorer states regardless of the size of the differences between the two scores. For example, Respondent *i* could have given 0.50 to state 12111 and 0.60 to state 12112 (a difference of 0.10). Respondent *j* could have scored 0.60 to state 12111 and 0.62 to state 12112 (difference is 0.02). In this study, logical inconsistency in both respondents was equally counted as one. This treatment is different from the study reported by Lamers *et al.* in that to consider the scores to be logical inconsistency, the difference between two scores had to be 0.1 or greater. If the Lamers *et al.* definition had been followed in this study, logical inconsistencies among the Thai respondents would have been reduced.

From the analysis of the possible causes of logical inconsistency in respondents with fewer than sixteen inconsistencies, the inconsistency proportions are slightly lower in the second half of the TTO task. The inconsistencies are slightly greater in the second half of the task in the respondents with more than fifteen inconsistencies. The differences between the first and the second half of the task may have resulted from either the learning or overwhelming effects that would have developed in the respondents. It seems that respondents with fewer inconsistencies could have "learned" how to cope with the TTO questions, and although they may have been fatigued by the level of difficulty in the interview task, learning effect may have been stronger than fatigue effect. Whereas the highly inconsistent respondents may have been fatigued in the second half of the task, thus, fatigue effect could have been stronger than learning effect. However, the differences between the two sets are too subtle to assume that both effects definitely developed. More should be explored on this issue.

There are four respondent subgroups, and one of these subgroups is expected to be used in the model specifications. Since excluding the scores from the inconsistent respondents has significant effect on the mean actual scores, the question arises as to which subgroup should be used. If the scores from all respondents are chosen to be used to estimate the 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 preferred. By excluding the scores from the respondents with more than five inconsistencies (subgroup 4), almost one quarter of the health states are significantly different from the scores estimated from all respondents. Based on the score validity, the scores from subgroup 4 respondents should be selected because these scores given were assumed to be the most "valid", due to the fact that the extent of inconsistent values was lower than in any other subgroup. The respondents in this subgroup tended to understand the task and were likely to assign 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 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 the four subgroups generated in this study. By this ground, the

respondents with fewer than eleven inconsistencies (subgroup 3) were chosen to estimate the Thai scores.

Additional supporting evidence that the scores from subgroup 2 are not preferable is that the scores from the respondents in this subgroup tend to be lower in "quality" because many inconsistencies were identified in their responses and high scores were assigned to very poor health states. It is also shown in Table 6.1 that the proportions of elderly respondents and those with primary education were slightly higher in this subgroup. The average numbers of inconsistencies were also high in this group of respondents as presented in Table 6.5. The respondents with 11-15 inconsistencies have 2.73 inconsistencies per respondent in Set A and 2.59 inconsistencies per respondent in Set B, compared with 1.6 in Set A and 1.58 in Set B identified in the respondents with 6-10 inconsistencies. By excluding the inconsistent respondents, as presented in Table 6.4, the rank of the scores ordered from highest to lowest is still highly correlated with the scores assigned by all respondents. This evidence can offer reassurance that it is the mean scores of health states, rather than the rank, that is likely to change after the exclusion of the highly inconsistent respondents.

6.6 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 were compared. The scores from the respondents with fewer than eleven inconsistencies were 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. Results of the analysis regarding causes of logical

inconsistency could be used as a platform to further explore the roles of learning or fatigue effects in the development of logically inconsistent values.

Chapter 7 Estimation of the Thai health states preference model

The previous chapter has shown that inclusion of responses from inconsistent respondents systematically changes the mean scores given to different health states. The scores from the respondents with fewer than eleven inconsistencies were chosen to be used in the estimation of a model of Thai health state preferences for the reasons given in the previous chapter. This chapter reports the results of estimating a model for health state scores using the chosen subgroup. The scores from the other three subgroups are also used to estimate the models in order to show what difference it would have made had these subgroups been preferred.

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. The models are compared and the "best" model is chosen. The impact of choice of subgroups on the "best model" is 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.

7.1 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 *et al.* (2005) models are explored in this study (67, 69, 155). 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 *et 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. The simple main effects model with no variables representing the interactions between dimensions is estimated first. Next the models including the interactive terms are estimated, then the

performance of the models (with and without interactive terms) is compared. To identify whether the different combinations of health states in the twelve health sets have a significant effect on the estimated scores, health sets are also incorporated in the model estimations. Note that this is undertaken in order to see whether the different combinations of health states have a significant effect on the estimated scores, rather than being intended to include the health sets variable in the final model to estimate Thai preference scores.

7.1.1 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 comply 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)" (66). 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 in economic evaluation. Therefore, the first priority is that the model estimates higher scores for better health states, that is, it produces logically consistent scores.

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 assess model robustness is the number of states with the absolute difference between the predicted and the actual scores larger than 0.1 (67, 155). 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, all other things equal, the model estimating the highest score for the best ill health state and lowest score for the worst state is favoured.

7.1.2 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 problems), or at level 3 (severe problems).

A general model is:

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

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, x'_i = the explanatory variables, β = a vector of coefficients, ϵ_{ij} = an error term. Scores for individual health states are uncorrelated as shown by the following formulae.

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

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

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

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 (156).

A model for panel data is:

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

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, x'_{ij} = the explanatory variables, z'_i = the individual effects on the scores, ϵ_{ij} = an error term (152)

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:

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

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 beliefs and the personal beliefs on health. Although some of these factors are observed (age, gender, residential area, interviewer effect), some are not.

$$E[c_i|X_i] = h(X_i)$$

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:

$$\bar{y}_i = \frac{1}{N}\sum_{j=1}^{N} y_{ij} \quad , \bar{x}_i = \frac{1}{N}\sum_{j=1}^{N} x_{ij} \quad (152, 157)$$

The RE model is estimated by the following equation:

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

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

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

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

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

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

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

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

7.1.3 The variables

The variables specified in Dolan (1997), Dolan and Roberts (2002) and Shaw *et al.* (2005) used in this analysis are as follows (67, 69, 155). 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 (*N3*) 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 7.1.

Table 7.1 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

(67)

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 7.2.

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 7.3.

Table 7.2 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

(155)

Table 7.3 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

(69)

7.1.4 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 are the formulae 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).

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

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|x_j - y_j|$$

Where: n = number of health states, x_j = the score estimated from the model for health state j, j =1, 2...n, y_j = the actual score of health state j, j =1, 2...n (152)

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.

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

Where: $mean_p$ = mean of post-treatment states, $mean_b$ = mean of baseline states, SD_b = Standard deviation of baseline mean (158)

7.1.5 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. The second method is to find the cause of inconsistency using the coefficients in the resulted models. By using the 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 7.1. 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,

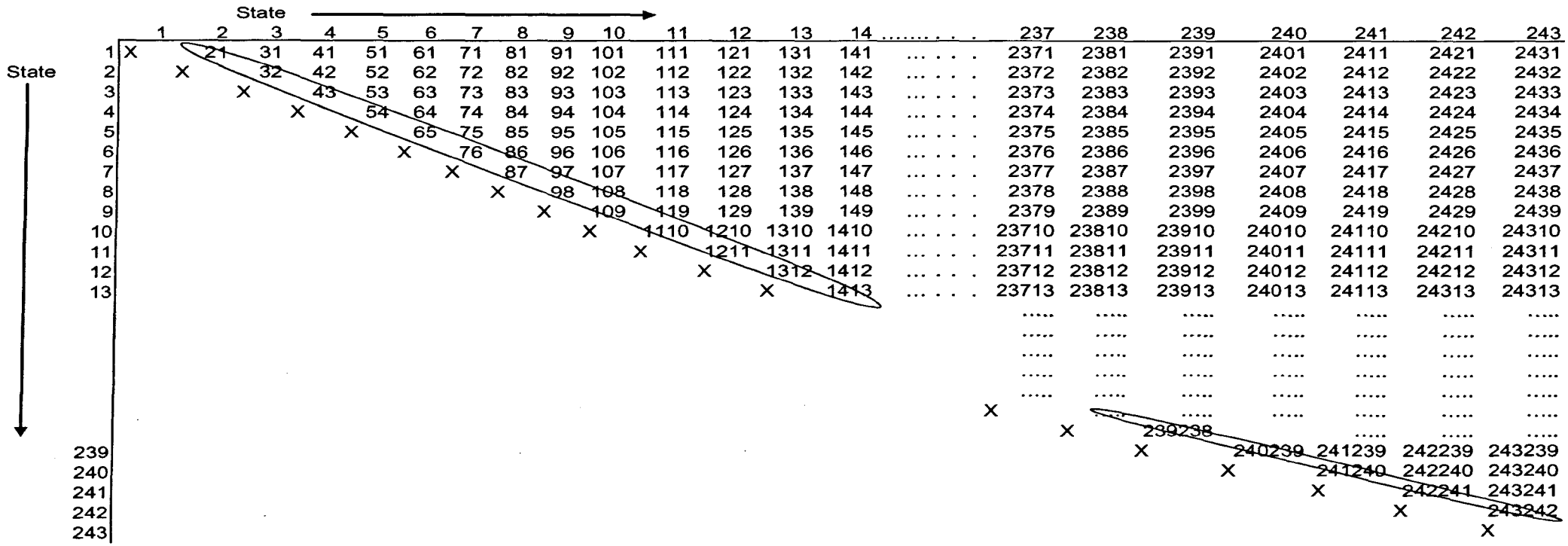
a total of 242 loops can be constructed which would cover all potential pairs. Only the pairs that can be used to identify logical inconsistencies are taken into account.

Note that only the pairs of health states up until the 45th loop had been examined. 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 inconsistent scores in later loops 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. An example of the do-file used in the identification of logically inconsistent TTO values in Health set 1 is presented in Appendix 4. The same strategy is applied to identify inconsistent values in the modelled scores for all 243 states.

7.2 Results

The results are divided into two sections. In the first section, results are presented for analyses based on the preferred subgroup of respondents (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 7.1 Identification of logical inconsistency from the estimated 243 states



7.2.1 Analyses based on Subgroup 3

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).

Dolan (1997) model

The main effects model (without an interaction term) was firstly specified, followed by the model with an interaction term (N3). At first, OLS 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 the 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 confirmed that the FE model was preferred to the OLS model. The Hausman test was used to compare the FE and RE models. The null hypothesis that differences in the coefficients are not systematic cannot be rejected at the p-value of 0.573, thus RE model is more efficient than FE model. Therefore, RE was used to estimate the model.

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

Dolan & Roberts (2002) model

Similar to the Dolan (1997) model, where the main effects model was specified first and the model with interaction terms were specified later. 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 models were then estimated. Using the 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 confirmed that the FE model was preferred to the OLS model. The FE and RE models were compared using the Hausman test. 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, RE was 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.

Shaw et al. (2005) model

Unlike the models estimated earlier, the models using the variables from Shaw *et al.* (2005), 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 for states worse than death.

The first model with no interaction term was estimated using OLS. The null hypothesis of equal variance was rejected at the p-level of 0.000. Therefore, the OLS model was not appropriate to fit the data. Next, Feasible generalized least squares (FGLS), which is one type of RE model, was used. This was used where the variance components were unknown (152). The insignificant coefficients were dropped and the models were re-analysed. The models with interaction terms were specified afterward.

The results of parameter estimates using the main effects models from the Dolan 1997, Dolan & Roberts 2002 and *Shaw et al.* 2005 models are presented in Table 7.4 and the models with the interaction variables using the RE model are shown in Table 7.5. Results of the FE model, including the interaction terms, are presented in Appendix 5.

Table 7.4 Parameter estimates and fit statistics of the three alternative model specifications using the main effects model

Dolan 1997			Dolan & Roberts 2002			Shaw *et al.* 2005		
Variables	Coeff.	SE.	Variables	Coeff.	SE.	Variables	Coeff.	SE.
mo	0.142	0.011	difmob1	0.303	0.012	m1	0.182	0.006
sc	0.130	0.007	difmob2	0.446	0.013	m2	0.157	0.008
ua	0.080	0.006	difsc1	0.127	0.014	s1	0.334	0.008
pd	0.101	0.012	difsc2	0.259	0.011	s2	0.155	0.007
ad	0.065	0.012	difua1	0.074	0.012	u1	0.167	0.007
m2	0.161	0.018	difua2	0.161	0.014	u2	0.225	0.008
p2	0.071	0.019	difpd1	0.172	0.013	p1	0.210	0.008
a2	0.040	0.019	difpd2	0.272	0.011	p2	0.126	0.007
cons.	0.189	0.014	difad1	0.106	0.012	a1	0.142	0.007
			difad2	0.172	0.013	a2	0.162	0.008
			cons.	-0.075	0.013			
			Mean score of state 33333	-0.419				
R2 (overall)	0.445			0.445			NA	
RMSE	0.111			0.111			0.298	
MAD	0.087			0.087			0.230	
Number of states with absolute difference >0.1	29			29			61	
Numbe of logical inconsistencies in the estimated 243 states	0			0			143	
Cohen effect size	1.047			1.055			0.541	
scores for state								
11112	0.746			0.750			0.858	
33333	-0.497			-0.494			-0.113 (state 22323)	

Note: RMSE= Root mean squared errors, MAD=Mean absolute difference

Table 7.4 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. Nine variables are significant in the Dolan (1997) model, eleven in the Dolan & Roberts (2002) model and ten in the Shaw *et al.* model. All coefficients have positive signs. The 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 presented for 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. The Breusch-Pagan test suggested that after applying the robust estimator, the variances are still heterogeneous (159). The robust estimates of the SEs are slightly lower compared with those estimated using the non-robust SEs. Twenty-nine states have absolute differences between the actual and estimated scores exceeding 0.1 in the Dolan 1997 and Dolan & Roberts 2007 models, with sixty-one states in the Shaw *et al.* 2005 model. No logical inconsistency is identified in the first two models, and one hundred and forty-three inconsistently valued health states are identified in the third model. The smallest Cohen effect size is seen in the third model. Among the three models, the Shaw *et al.* 2005 model predicts the highest score for state 11112 and the lowest score is not predicted for the worst state (state 33333).

In Table 7.5, 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.

Table 7.5 Parameter estimates and fit statistics of the three alternative model specifications including interaction terms

Dolan 1997 model			Dolan&Roberts 2002 model			Shaw et al.2005 model		
Variables	Coeff.	SE	Variables	Coeff.	SE	Variabl	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			d1	-0.256	0.013
			of state 33333	-0.419				
R2(overall)	0.448			0.447			NA	
RMSE	0.102			0.106			0.257	
MAD	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, MAD=Mean absolute difference

Out of eighty-six states, twenty-eight states have an absolute difference between the estimated and mean scores larger than 0.1 using the Dolan (1997) model, thirty states using the Dolan & Roberts (2002) model and fifty-nine states using the Shaw *et al.* (2005) model. Thirty-seven logically 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 the only model that estimates 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. Performance of the FE and RE models using the interactive terms are quite similar. The number of health states with absolute differences between the actual and estimated scores exceeding 0.1 are slightly greater in the RE model. To see whether the different combinations of health states (twelve health sets) have significant effects on the coefficients, twelve dummy variables are generated, one variable for each health set, and incorporated in the model estimation. By comparing the models with and without the total twelve dummy variables for all health sets (not illustrated), it is suggested that health sets have no significant effects on the estimated scores at p-level=0.05.

Comparing the coefficients and model performances between the Dolan 1997 main effects model and the model with interaction term (N3), coefficients of variables *mo, sc, ua, pd* and *ad* were slightly greater but *m2, p2* and *a2* were smaller in the main effects model. R-squared was also slightly smaller. RMSE and MAD were slightly greater. Twenty-nine states had the differences between the actual and estimated scores greater than absolute 0.1, compared with twenty-eight states in the N3 model. There was no logical inconsistency in the estimated scores. The Cohen effect size was slightly smaller and the estimated score for state 11112 was slightly higher, and for state 33333 it was slightly lower.

In the comparison between the Dolan & Roberts 2002 main effects model and the model with interaction terms, three coefficients were slightly lower, whereas seven were slightly higher and one was similar in the main effects model. R-squared was slightly lower and RMSE and MAD were higher than in the models with interaction terms. Only twenty-nine states had differences between the actual and estimated scores exceeding 0.1. There was no logical inconsistency predicted from the main effects model. The Cohen effect size was smaller and the estimated scores for both of states 11112 and 33333 were lower than those estimated using the interaction model.

In the Shaw *et al.* 2005 main effects model, only the coefficient of variable m1 was smaller than that of the model with interaction terms. Other coefficients were higher. RMSE and MAD were slightly higher. Sixty-one states had differences between the actual and estimated scores exceeding 0.1, compared to fifty-nine states in the interaction model. There were a considerably higher number of health states with logically inconsistent values: 143 states versus only 37 states from the model with interaction terms. The Cohen effect size was half of that of the interaction model. The score for state 11112 was higher. The lowest score (-0.113) was predicted for state 22323, rather than for state 33333.

The Dolan 1997 model is the only model estimating the scores with no logical inconsistency in both the main effects and the N3 models. Thus the Dolan 1997 model is the preferred model with which to estimate the Thai scores. The N3 model is favoured because although R-squared is slightly smaller in the main effects model, RMSE and MAD of the N3 model are slightly smaller. The N3 model predicts one fewer number of states with the absolute differences between the estimated and actual scores exceeding 0.1, and the score estimated for the best ill health state is slightly higher (0.766).

There are also 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. The R-squared 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 an 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, because the latter model estimates scores with logical inconsistencies this model is less favourable.

7.2.2 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. The following variables are added in the Thai model: all two-way interaction terms; and the variable *x4* from the US Hispanic model because these variables have been recently used in the model specifications along with other variables reported in previous studies (120). The variable x4 is a dummy variable taking value 1 if four or more levels are at level 2 or 3. By taking this variable into account, there could be some possibility to improve performance of the Thai model. The modelling sample from subgroup 3 respondents is again used in the model estimation. The models are estimated using an RE model. Again, the models are 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 7.6.

Table 7.6 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

Results of the X4 model and the interactions model specifications are shown in Table 7.7, along with the Thai model coefficients.

Table 7.7 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 significance at p-level < 0.05 are presented in Table 7.7. The highest R-squared is 0.580 from the X4 model. The Breusch-Pagan test is used to test the model heteroskedasticity. The null hypothesis that 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 7.7, 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.

7.2.3 Health states with large differences between the actual and estimated scores

Using the Thai algorithm to estimate the scores, there are twenty-eight states with the absolute differences between the estimated and the actual scores exceeding 0.1. Users of the Thai scores could be reassured that the Thai algorithm is able to predict scores with relatively high accuracy. Almost seventy percent of the estimated scores (out of eighty-six states used in the interview) are relatively close to the actual scores. Health states with poorer score estimations are presented in Table 7.8.

Table 7.8 Health states with the differences between the actual and estimated scores exceeding 0.1

Health states with the differences between the actual and estimated scores > 0.1

states	scores		
	actual	estimated	differences
1 1 2 2 3	0.511	0.409	0.102
1 1 2 3 2	0.676	0.336	0.340
1 1 3 3 2	0.415	0.276	0.139
1 2 1 2 3	0.506	0.349	0.158
1 2 3 3 1	0.331	0.194	0.138
2 1 1 1 2	0.762	0.642	0.120
2 1 2 1 1	0.729	0.620	0.109
2 1 2 3 1	0.388	0.253	0.134
2 1 3 1 2	0.497	0.384	0.113
2 2 1 1 3	0.547	0.303	0.244
3 2 2 3 2	-0.086	-0.202	0.116
3 3 2 2 2	-0.050	-0.168	0.118
3 3 2 2 3	-0.120	-0.248	0.128

Health states with the differences between the actual and estimated scores <-0.1

states	scores		
	actual	estimated	differences
1 2 1 2 1	0.423	0.605	-0.183
1 2 3 1 3	0.059	0.303	-0.245
1 3 1 2 3	0.118	0.229	-0.111
2 1 2 2 1	0.390	0.545	-0.155
2 1 3 1 3	0.201	0.303	-0.102
2 3 1 1 3	0.068	0.183	-0.115
2 3 1 3 1	-0.038	0.073	-0.111
2 3 3 2 3	-0.138	-0.011	-0.126
2 3 3 3 2	-0.202	-0.085	-0.117
3 1 1 3 1	-0.122	0.016	-0.138
3 1 2 1 3	-0.160	0.066	-0.226
3 2 2 2 3	-0.281	-0.128	-0.153
3 2 3 2 2	-0.249	-0.108	-0.141
3 3 1 2 1	-0.319	-0.070	-0.248
3 3 2 3 2	-0.426	-0.322	-0.104

The greatest difference is seen with the state 11232 where the actual score for this state was 0.340 higher than estimated. Large differences also arise with states 33121, 12313 and 22113. States 33223 and 21211 are underestimated, whereas the fairly similar states 33232 and 21221 are overestimated.

7.3 The Thai algorithm

The Thai model is estimated from the full sample of Subgroup 3 using the Dolan (1997) model. Coefficients of the Thai model, as well as the 95% confident intervals of the coefficients, are presented in Table 7.9.

Table 7.9 Coefficients of the variables in the Thai model

Variables	Coef.	95% CI	
constant	0.202	0.178	0.226
mo	0.121	0.103	0.139
sc	0.121	0.111	0.131
ua	0.059	0.048	0.070
pd	0.072	0.053	0.092
ad	0.032	0.013	0.051
m2	0.190	0.162	0.219
p2	0.065	0.035	0.095
a2	0.046	0.017	0.076
N3	0.139	0.114	0.164

Thai Utility 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 Appendix 6. To take prediction errors of the algorithm into account, the upper and lower levels of the estimated scores are also provided in the appendix.

7.4 Impact of choice of subgroups

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 was used to choose between the FE and RE models, and it indicated that an RE model was the most appropriate model.

7.4.1 Dolan (1997) model

Table 7.10 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 do not follow this trend are *ad*, where the coefficient using subgroup 3 is slightly smaller than that using subgroup 2. The coefficient of *p2* using subgroup 2 is slightly lower than that using subgroup 1. The coefficients of *N3* are gradually increased using the scores from subgroup 1 to 3, but using subgroup 4 the coefficient is lower than that for 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 7.10 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 subgroups 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 subgroups 1 and 4. The scores

estimated from subgroups 2 and 3 have twenty-two and twenty-eight states with the absolute difference larger than 0.1, respectively.

7.4.2 Dolan & Roberts (2002) model

Table 7.11 presents the parameters estimated from the Dolan & Roberts (2002) model using the scores from the modelling sample of the four respondent subgroups. All variables are statistically significant at a p-level less than 0.05. Two variables: constant and ANY13, have negative signs. The coefficients of almost all 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 7.11 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 across the models from 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.

7.4.3 Shaw *et al.* (2005) model

Table 7.12 presents the parameters estimated from the Shaw *et al.* (2005) model using the scores from the modelling sample of the four respondent subgroups. The first model used to estimate the models was the 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 subgroups 1 to 3. Two variables: *i3-squared (i32)* and *d1* have negative signs. The *i2-square (i22)* is negative using subgroups 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 7.12 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 across the models from 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.

7.4.4 The comparison of scores estimated from all models

The scores estimated from all models using the four subgroups are compared in Tables 7.13, 7.14 and 7.15. The scores in the following tables are estimated from the full sample (that is, the modelling and validation samples combined). Table 7.13 compares the estimated scores estimated from the Dolan (1997) model across the four subgroups. None of the four subgroups estimates inconsistent scores. 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. Subgroup 4 has the highest Cohen effect size. The model estimated from subgroup 3 gives the greatest number of negative scores [68].

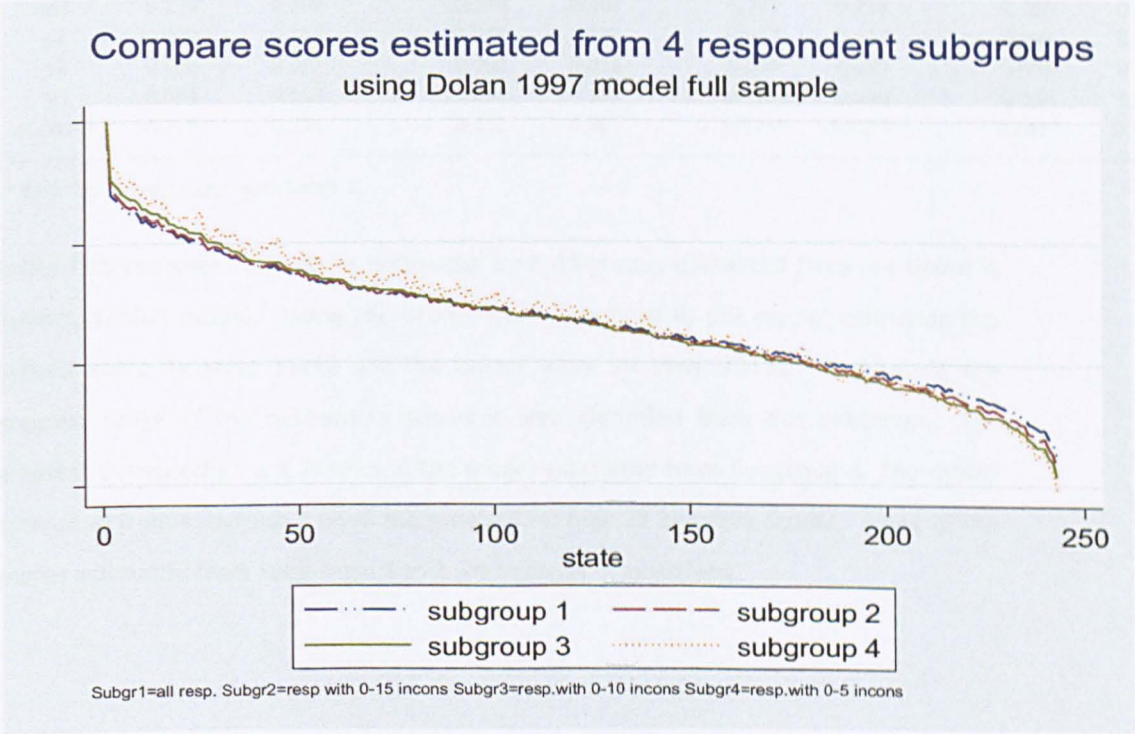
Table 7.13 Comparison of the scores estimated from the Dolan 1997 model by subgroup

	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 using a graphical illustration in Figure 7.2. The Y-axis represents the scores ranging from -0.50 to 1, while the 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 tended to be higher than those estimated from other subgroups. The scores estimated from the respondents in Subgroups 3 and 4 are quite similar to each other.

Figure 7.2 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 7.14. The confidence intervals for all coefficients estimated from Subgroup 2 overlap those estimated from Subgroup 1. The constant terms estimated from Subgroups 3 and 4 do not overlap that from 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 more *consistent* respondents (Subgroups 3 and 4), where the resulting models are slightly different from

that estimated from all respondents, but only the differences in constant terms are significant.

Table7.14 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 7.15 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 greatest Cohen effect size is seen in the model estimated from Subgroup 4. The model estimated from subgroup 3 gives the greatest number of negative scores. Some of the scores estimated from Subgroups 1 to 3 are logically inconsistent.

Table 7.15 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 7.16 compares the scores estimated from the Shaw *et al.* 2005 model across the 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 for state 33333. The models using the scores from Subgroups 1 to 3 estimated the lowest score 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 7.16 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

7.5 Discussion

This chapter reports the results from using the scores of 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. Compare with the main effects model (without an interaction term), by incorporating the interaction term (N3) into model estimations, the models perform slightly better. The different combinations of health states are unlikely to have significant effects on the estimated scores. To see the impacts from the choices of subgroups, the scores from the other three subgroups were also used in the model analysis. It was shown that excluding the scores from the inconsistent respondents has an impact on the coefficients and performance of the models. An arbitrary classification of respondents was applied in this study; had the respondents been classified using a different number of inconsistent responses, the resulting coefficients of the models would have 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. 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, as stated in Chapter 6, the scores from Subgroup 4 are not favoured to estimate the Thai scores because a considerably large proportion of respondents would have been excluded. Also, if the scores from Subgroup 3 were used to estimate the scores using the Dolan & Roberts 2002 model, there would be inconsistent responses in the estimated scores.

One may argue that if the scores from all respondents were used in the final model estimation, the model estimated from Subgroup 2 would 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 was 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 generate 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 a change was only seen in the constant term. This can be used as evidence to support the argument that by excluding the highly inconsistent scores at the "appropriate" number, the models are changed. The changes in the models can be regarded as both "justifiable" and "unjustifiable". It could be considered "justifiable" because the resulting models perform better (higher R-squared) and the estimated scores are systematically changed in a favourable fashion, in that the best ill health state is assigned a slightly higher score (0.766 using Subgroup 3 and 0.829 using Subgroup 4, compared with 0.707 and 0.729 using Subgroups 1 and 2, respectively) and the worst health state is assigned a slightly lower score (-0.452 using subgroup 3 and -0.513 using Subgroup 4, compared with -0.373 and -0.420 using Subgroups 1 and 2, respectively). The changes may be considered "unjustifiable" because the scores of some respondents were excluded and the data were lost from the model analysis. However, the exclusion was, in fact, justified here because the excluded scores were given by respondents who may have had difficulties participating in the TTO

interview. This can be used to support the selection of Subgroup 3 to model the Thai preference scores.

Regarding 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 were higher than those with higher numbers of inconsistent responses, implying that the model performance may be unsatisfactory. 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 markedly 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 could be used to support the view that inconsistent respondents were likely to assign scores at random with little correlation to the severity of the health states. This could be the result of respondents not understanding the health state descriptions or the tasks, or due to a 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 comparison 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 gained from health interventions. More research should be conducted to develop deeper insight regarding the responsiveness of the Thai preference scores.

Although a panel data model was applied to take into account the heterogeneity of individuals (age, gender, education, etc), heteroskedasticity still exists. This is in line with the models used to estimate the preference score for EQ-5D health states in other countries. Heteroskedasticity in this study cannot be accounted for by using a robust estimator approach. When adding the interaction terms into the Thai model, the problem still exists. According to the selection criteria, the Dolan 1997 model was chosen to estimate the Thai scores. Compared with the UK model, in which 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 variables were dropped in the Thai model because the models then performed slightly better. Moreover, both excluded coefficients that 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.

From the interactions model, there existed two patterns of interactions between the dimensions: [1] mobility and usual activities and [2] mobility and pain/discomfort. Variable *N3* was not significant in this model. These findings provide more insight into the impact of the interactions between the attributes of health on the Thai preferences of health.

The Thai algorithm estimates a score of 0.766 for the second best health state (11112). To take into account the prediction errors (95% CIs), the highest score that could have been predicted is 0.809. This implies that by moving away from full health, Thai preference drops by approximately 0.2. The reasons could be either that the score genuinely drops due to a deviation in this attribute (anxiety/ depression), or that EQ-5D health state descriptions cannot capture health states that would have had a score close to 1. 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, and therefore assigned a 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 studies should be aware of this problem and the interviewers should encourage the respondents to read the questions carefully and take all dimensions into consideration before assigning scores.

As far as the researcher can determine, this is the first study using the Stata program to detect logically inconsistent responses in estimated scores. The resulting estimates of logical inconsistency were re-examined by using the model coefficients to identify the possibilities of inconsistent responses if, for example, the coefficient of being in level 2 was higher than that of being in level 3. If this were the case, other things being equal, the resulting score for the better state would be lower than that of the poorer state. The researcher 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, in which one of the criteria to select the best model was logical inconsistency in the estimated scores, by using the Stata program to search for inconsistent responses, some logically inconsistent responses in the estimated scores were detected. Sixty states from the Dolan & Roberts 2002 were detected using the UK data and fifteen states from the Shaw *et al.* (2005) using the US data. Further results of the comparison of the Thai model with those estimated from other countries are provided in the next chapter.

7.6 Conclusion

The Thai algorithm is based on the Dolan (1997) model using data from respondents with fewer than eleven inconsistent responses. This model was chosen because the resulting algorithm produces no logically inconsistent scores, the model is the most parsimonious and highly robust, and the responsiveness is acceptably high. The effect of using the scores from the other subgroups was explored. The constant terms in the models were significantly changed after the scores from the highly inconsistent respondents were excluded from the model specifications. Applying a robust estimator approach in the model estimation, heteroskedasticity still exists. Interaction terms were added in the Thai model to eliminate the problem of heteroskedasticity, but it did not yield a superior algorithm. The specific combination of health states an individual faced is unlikely to have significant effects on the estimated scores. The models estimated from the lower number of inconsistent respondents predicted 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.

Chapter 8 A comparison of Thai preference scores with those from five other countries

8.1 Introduction

Economic evaluations are conducted in countries where preference scores are not yet available using health state values from other countries. Given that a considerable amount of money, time and expertise is required to estimate preference scores from the general population, researchers can undertake economic evaluations by obtaining the preference scores from other countries by either taking the health state values from studies of similar interventions or using existing algorithms. This is also the case in Thailand. An example of the former procedure is the cost-effectiveness analysis of inhaled corticosteroid therapy in Thai patients with asthma, where the preference scores are extracted from the US Asthma Policy model (160). The scores were obtained from a sample of 100 adults patients with asthma in Lexington, Kentucky who were interviewed using the Health Utility Index and the Asthma Symptom Utility index (161). Another example of the same procedure would be the cost-utility analysis of Erythropoietin treatment of anemia in Thai patients receiving chemotherapies, in which the QALYs were obtained from a literature review (162). An example of the latter approach is the cost-utility analysis of blood glucose control with metformin versus usual care in patients with type 2 diabetes mellitus in Beijing, China, where the preference scores were obtained using the UK algorithm (163). Thus it is possible to inform resource allocation decisions while preference scores from their own general population are awaited.

These approaches should be used with care. Given that there are differences in culture, religious beliefs, clinical practices and health systems across countries, preference scores derived from the general population in different countries might be expected to differ. In the study of inhaled corticosteroid in Thai patients with asthma, the researchers concluded that inhaled corticosteroids are cost-effective in the Thai health care context. However, this is not necessarily the case given that the preference scores used in the study come from US asthma patients. Resource allocation guidance from the study could be misleading because of the use of US rather than Thai patients' preferences to estimate QALYs.

EQ-5D scores have been elicited from the general population in many countries and the Thai preference scores are now available, as presented in Chapter 7. It is interesting to see to what extent Thai preference scores differ from those of other countries and to explore the implications of using scores from other countries in Thai cost-utility studies. This chapter aims to compare the Thai preference scores for EQ-5D health states with the scores from the UK, US, Japan, South Korea and Zimbabwe (67, 69, 115, 117, 121, 155). The reasons for choosing these five countries are as follows. Several researchers have chosen to compare their scores with those of the UK because they are the first preference scores estimated for EQ-5D health states and have been treated as a reference set. To see whether the population from the countries located in the same continent give similar values for health, the Thai scores are compared with those assigned by Japanese and Korean respondents. The other two countries, the US and Zimbabwe, are chosen to represent the North American and African continents. These two countries were chosen to compare the scores given by the respondents from different continents. This can be used to test an assumption that respondents from the neighbouring countries, in this case, Thailand, Japan and South Korea, would share some common values on health, and that preference scores elicited from the three countries would, to some extent, be more highly correlated with each other compared with the scores elicited from the population from other continents (UK, USA and Zimbabwe).

The outline of the chapter is as follows. Firstly, to compare the scale of the preference valuation studies, (i.e., number of respondents interviewed, number of health states used, and interview duration) the studies conducted in the US, UK, Zimbabwe, Japan and South Korea are used as a mean of comparison with the Thai study. Then, the methods used in this chapter to compare preference scores are described, followed by a discussion of differences and similarities between the Thai scores and the scores from the other countries and the impact of using these scores to calculate QALYs in Thai cost-utility studies.

8.2 Comparison of the preference valuation studies

An overview of the scale of these studies in terms of number of respondents, number of interviewers, mean age of the respondents, education attainment levels, mean overall interview duration and number of health states interviewed is reported in Table 8.1. All six countries where preference scores were estimated for EQ-5D health states, including Thailand, a face-to-face interview was conducted in a representative sample of the general population using the MVH protocol with ranking, VAS and TTO as the preference elicitation methods (109, 115, 117, 121, 164). The protocol was applied to the studies in all five countries represented in the table.

Table 8.1 Comparison of the overview of the preference studies in five countries

Characteristics	UK	US	Japan	South Korea	Zimbabwe
No.of respondents	3,395	4,048	621	500	2,384
No. of interviewers	92	109	62	19	9
Mean age (yrs.)	45	44.67	NA	41.3	NA
Education level (%)					
8 years or less	37	6.12	NA	3.8	19.2
12 years	40	46.59	NA	64.6	67.1
16 or more	20	47.29	NA	31.6	0.8
Mean overall interview duratic	54	NA	30*	NA	NA
No. of health states interviewe	13	15	17	15	7

NA=not available
* VAS duration not included

(69, 109, 115, 117, 121)

Note that mean age and education levels for the Japanese study were not provided. Mean overall interview duration was unavailable in the South Korean study. Mean respondent age and interview duration were not provided in the Zimbabwean study.

Comparisons of preference scores have mostly been conducted using the scores from study's country of the origin and the UK, the country in which the first-ever set of preference scores was estimated. Badia *et al.* compared the EQ-5D scores derived from the UK and the Spanish general population (23). Both the UK and Spanish studies elicited the scores using the same set of 43 health states by the TTO method. Compared with the UK scores, the Spanish scores are similar for milder states but tended to be lower for more severe states. British people seemed to place greater weight on

pain/discomfort and anxiety/depression; if they had a score of level 3 in pain/discomfort or anxiety /depression, their preferences were lower than those of Spanish. In contrast, the Spanish gave greater importance to mobility and self-care.

Tsuchiya *et al.* compared the Japanese scores with the UK scores (115). The estimated scores from the two countries were highly correlated when examined by Pearson's correlation coefficient, with a coefficient of 0.924. Except for the very mild states, Japanese utility seemed to be higher than the British for all health states. The maximum difference in the modelled scores and the directly elicited scores was 0.585 for state 11133 and 0.527 for state 23232, respectively. Compared with British utility, Japanese utility seemed to be affected more if there was any deviation from full health or problems with mobility and usual activities, but less affected if there was any extreme problem or problems in self-care, pain/discomfort and anxiety/depression.

Other studies included additional methods in the comparison of the preference scores. A comparison of preferences in Germany, Spain and the Netherlands (165) examined: [1] mean EQ-5D values of the three countries; [2] the association between socioeconomic factors and EQ-5D preference scores; and [3] differences in loss of QALYs using the different value sets where QALY loss is calculated by subtracting the age-specific and gender-specific index scores from 1. Luo *et al.* conducted not only head-to-head comparisons of the scores, but also explored responsiveness using Cohen effect size (ES) and Standardized response means (SRM) (166). The authors generated all possible (29,403) pairs of health states under the positive transition assumption in which the first state of the pair is assumed to be a baseline state and the second state of each pair a post-treatment state. ES and SRM of the US study are 1.58 and 1.59 respectively and those for the UK are 1.42 and 1.38 respectively. Although a smaller range between the highest and lowest scores is seen in the US scores compared to those of the UK, the responsiveness of both scores is fairly similar. However, attempts to compare the responsiveness of all possible scores are not without problems. One might consider that the EQ-5D is a responsive measure because the responsiveness is much higher than the threshold (0.8). However, responsiveness may be high in this study because the authors used all possible pairs of states including the pairs between the worst state (33333) and perfect health (11111). In practice, a treatment is unlikely to move the patients from the worst state to perfect health. To assess the responsiveness of the EQ-5D, the research should be conducted in a particular group of patients, for example, persons with hearing complaints in which ES and SRM of the UK

EQ-5D scores are 0.05, and the ES and SRM of the Dutch EQ-5D scores are 0.03 and 0.02 respectively (167).

Because the Thai preference scores were unavailable Sakthong *et al.* measured quality of life of a Thai sample with type-2 diabetes by applying the scores from the UK, US and Japan (53). To investigate which of the set of scores was appropriate to represent the Thai preferences on health, the authors examined the level of agreement between the US, UK and Japanese scores using the Pearson correlation coefficient. The Bland-Altman (BA) plot and the intraclass correlation coefficient (ICC) were used to test the one-two week test-retest reliability. Approximately twenty percent of the respondents were randomly selected to be interviewed by phone to test the test-retest validity. The authors concluded that, in the absence of Thai preference scores, the Japanese scores are recommended for use in economic evaluation in the Thai setting because the scores provided better test-retest reliability and validity, i.e., the ICC of the Japanese scores was slightly higher than those of the UK and US. However, this may not be a robust conclusion because the sample size was small and based only on one disease. It may not be true of patients with other diseases. Another example of cost-utility analysis on different diseases should be conducted, and given that the Thai preference scores are now estimated from Chapter 7, the Thai and Japanese preference scores will be used to confirm the similarity between them, as stated in the study by Sakthong *et al.*

Thai preference scores were established in Chapter 7. This chapter is going to compare the Thai scores with those estimated from the six models used in five countries, which is more extensively than has been previously attempted. Recall that the UK scores have been estimated using two models. The six sets of scores are compared with respect to differences between observed and modelled scores, level of agreement, and responsiveness. Furthermore, an example is presented using actual Thai cost-utility data. Details of the methods used in the comparisons are described in the next section.

8.3 Methods

8.3.1 Differences between the observed and modelled preference scores

It is interesting to see to what extent actual and estimated Thai scores differ from those of other countries. Through comparison of the coefficients of these models, the extent to which the estimated scores are correlated and the dimensions of the EQ-5D to which Thais attach greater weight can be identified. The number of states differs between countries, although several states have been valued in several countries. Mean actual scores of these states are compared; the mean absolute differences between the Thai scores and those of each country are calculated. Then the analytical models, variables, magnitude and signs of the model coefficients and mean absolute differences (MAD) of the final models are compared, including means and standard deviations (SDs). The modelled scores are examined in the following dimensions: [1] the number of states with negative scores; [2] the number of logically inconsistent pairs of health states; [3] the scores of the best ill state and the worst state; and [4] the range of scores for the best and worst states. The 243 scores estimated for all six countries are compared graphically.

8.3.2 Level of agreement

Level of agreement between the estimated scores is estimated using Pearson's correlation coefficient, the Intra-class correlation coefficient (ICC) and Bland-Altman (BA) plots.

Pearson's correlation

Pearson's correlation is used to test the correlation between two series of scores. The formula is:

$$r = \frac{\sum xy - n\bar{x}\bar{y}}{\sqrt{(\sum x^2 - n\bar{x}^2)(\sum y^2 - n\bar{y}^2)}}$$

In the case of the correlation between the Thai and the UK scores: x denotes the Thai scores and y denotes the UK scores. $\bar{x}$ and $\bar{y}$ represent means of the Thai and UK scores respectively (153).

To calculate the 95% confidence interval of r :

$$Z = \frac{1}{2} log_e(\frac{1+r}{1-r})$$

where Z denotes the transformed value of the correlation coefficient (r). The 95% confidence intervals (CI) range from F to G where:

$$F = Z - \frac{N_{1-\frac{\alpha}{2}}}{\sqrt{n-3}} \text{ and G= } Z + \frac{N_{1-\frac{\alpha}{2}}}{\sqrt{n-3}}$$

Where $N_{1-\frac{\alpha}{2}}$ denotes the standard Normal distribution for the $100(1-\frac{\alpha}{2})$ percentile. F and G denote the lower and upper limits of 95% CI respectively (168).

Intra-class correlation coefficients (ICC)

The ICC is a method used as the relative measure of reliability between subjects, in that if subjects differ little from each other, the ICC is small because the between-subject variability is small (169). In this chapter, "subject" means an individual country and the ICC is used to examine the variations of the scores rated by the respondents in different countries. If the ICC is low, it implies that the scores given by the two countries are less different. ICC can be interpreted as the measurement of inter-rate reliability at which the agreement between two raters is concerned (170). In the case of the reliability measurement between the UK and the Thai scores, one query may be how Thai raters, in relation to the UK raters, give values to EQ-5D states, taking into account the within-subject variances. ICC uses the variability of the scores (between-subjects variances) and the variability of errors (within-subject variances) to indicate the reliability. The ICC is calculated using the following formula.

$$ICC = \frac{between\ subject\ variances}{between\ subject\ variaces + error\ variances}$$

Analysis of Variance (ANOVA) is used to identify the error variances and between-subject variances. An ICC close to 0 indicates that the variations between the modelled preference scores are low, implying that the British and Thai scores are only slightly different. If the ICC is close to 1, then the differences between the Thai and British scores would be greater. Note that the interpretation of ICC in this sense is different from the test-retest reliability in that by measuring the test-retest reliability, the scores are given by the same subjects in different times, for example, one or two weeks apart.

The Bland - Altman plot (BA plot)

BA plots are used to examine whether a new measurement technique or an equipment can be used to substitute an old measurement technique (171). This method has been used to compare the performances of two different equipment, for example, to examine whether the Mini Wright peak flow meter can substitute the Wright peak flow meter in the lung function measurement (172). The plots are applied in this study to investigate the level of agreement between the Thai scores and the scores from the other countries. The BA plot is more informative than the correlation coefficient because the coefficient can be high if the scores are ranked with the same trends, but the scores for the same states can be largely different. The limits of agreement between the two scores are the means of the differences between the two scores $\pm$ 1.96 SDs and can be represented using graphical illustrations; the differences between the two scores are recorded on the Y-axis. The X-axis shows the mean of the two scores. Two sets of scores may be related (high Pearson's correlation coefficients), but may not "agree with" or be "close to" each other, thus the mean difference is large and the upper and lower levels of agreement are far apart.

8.3.3 Responsiveness

The responsiveness of the Thai scores is estimated using Cohen's effect size (ES) and Standardized response means (SRM) and is compared with the scores estimated from the other five countries. ES is the difference between post-treatment mean and baseline mean, divided by the standard deviation of the baseline (158). An effect size of 0.2 is considered as a small effect of minimal clinical importance, 0.5 as moderate and 0.8 as large (173). SRM is the difference between post-treatment mean and baseline mean, divided by the standard deviation of the difference of the two means. The benchmark for the SRM is similar to that for the effect size (0.2-small effect, 0.5-moderate effect and 0.8-large effect) (158). To see the responsiveness of the scores, all states are paired according to their scores, treating the score for the first state as a baseline score and that for the second state as a post-treatment score. Only positive transitions are examined, i.e., if the post-treatment score is higher than the baseline score of its pair. A total of 29,403 pairs $\left(\frac{243!}{2!(243-2)!}\right)$ are can be generated using the Stata program. Next, all pairs are divided into four subgroups (following Luo et al. (7)). Subgroup 1: *major improvement* where at least one dimension changes from level 3 to level 1 or 2 and no dimension worsens. Subgroup 2: *minor improvement* where at least

one dimension changes from level 2 to level 1 and no dimension worsens. Subgroup 3: *minor deterioration* where at least one dimension changes from level 1 to level 2. Subgroup 4: *major deterioration* where, at least one dimension changes from level 1 or 2 to level 3.

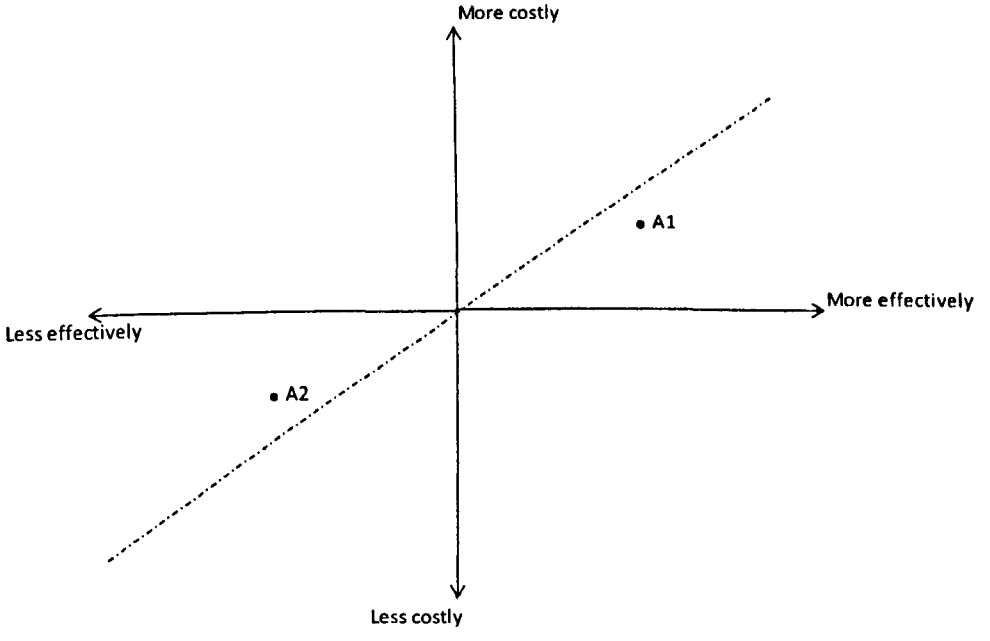
8.3.4 Example using a real Thai cost-utility analysis

In 2007, an economic evaluation was conducted to identify the most cost-effective prevention and control intervention for cervical cancer in Thailand (174). The interventions were Pap smear, visual inspection with acetic acid (VIA), Human Papilloma Vaccine (HPV) and HPV DNA test. A model-based cost utility analysis was conducted, adopting both societal and health care provider's perspectives. The health outcomes for the patients were measured using both Life Years (LY) and Quality-adjusted Life Years (QALYs) over a lifetime time horizon. Regarding the current situation, prevention programs are provided using either Pap smear or VIA, depending on the hospital. The researchers conducted a cost-utility analysis to identify whether providing the HPV vaccine would be more cost-effective, and if not, what alternative intervention should be implemented in Thailand. This analysis was used in this study because it was the only ongoing cost-utility analysis known to the researcher and was the best available real cost-utility analysis in which the researcher was allowed to access to the EQ-5D health states of the respondents.

In order to measure the quality of life of Thai patients with various stages of cervical cancer, 1,035 cervical cancer patients were asked to complete the EQ-5D questionnaire. Cervical cancers are classified into four stages and each stage has four categories: initial, remission, persistence and recurrence. The UK algorithm was used to transform the self-completed EQ-5D scores of the patients into preference scores. The scores are then used to estimate the QALYs for patients of the different screening methods. The seven interventions were: [1] 5-year interval Pap smear for women aged 30-60 years; [2] 5-year interval VIA for women aged 30-40 years; [3] HPV vaccine at the age 15 years; [4] 5-year interval Pap smear for women aged 45-60 and 5-year VIA for women aged 30-40; [5] HPV vaccine at the age 15 years + 5-year Pap smear for women aged 30-60 years ;[6] HPV vaccine at the age 15 years + 5-year interval VIA for women aged 30-45 years + 5-year interval Pap smear for women aged 30-60 years; and [7] HPV vaccine at the age 15 years + 5-year interval VIA for women aged 30-40 years + 5-year interval Pap smear for women aged 46-60 years.

The study reported that, compared with a "no prevention program" scenario, the most cost-effective intervention was the combination strategy of VIA and sequential Pap smear at five year intervals for women aged between 30 and 60 years. A comparison will be made between the results of CUA studies using QALYs estimated by the UK model and those estimated using the Thai model. The cost-effectiveness threshold in Thailand is 300,000 baht per QALY (approximately £5,500/QALY) (175). To prevent the problem of the interpretation of cost-effectiveness ratio where the same ratio can be obtained with different meanings, consider for example, the cost-effectiveness ratio of Interventions A1 and A2 where A1 is located in the north-east quadrant and A2 is in the south-west quadrant of the CE plane as shown in Figure 1. The dash line represents the willingness-to-pay threshold. The cost-effectiveness ratio of A1 is similar to that of A2 but the interpretations of both ratios are different. The ratio of A1 is the ratio of increasing cost per one unit of increased effect (better outcome) whereas in the south-west quadrant, the ratio is cost saving per one unit of decreased effect (worse outcome).

Figure 8.1 Cost-effectiveness ratio on the cost-effectiveness plan



To prevent problems of interpretation, the net monetary benefits (NMB) of the interventions are also presented. An intervention is considered to be cost-effective if NMB is positive (12). The following formula is used to calculate NMB.

$$NMB = R_T \Delta E - \Delta C > 0$$

NMB=Net monetary benefit

R_T = willingness to pay per unit of increased effectiveness

ΔE = increase of effectiveness

ΔC = increase of cost

(12).

8.4 Results

8.4.1 Observed and modelled preference scores.

A comparison of directly elicited scores is presented in Table 8.2. The Thai study has twenty health states in common with the UK, US and South Korean studies, 19 states with the Zimbabwean study and 9 states with the Japanese study. Mean absolute difference (MAD) of the differences between the Thai scores and the scores from the other five countries are reported at the bottom of the table.

Table 8.2 Comparison of the mean actual scores and differences of the mean scores from the five countries and the Thai scores

		Actual mean scores					Differences between the Thai scores and				
EQ-5D	Thai	UK	Zimbabwe	Korea	Japan	US	UK	Zimbabwe	Korea	Japan	US
11112	0.691	0.829	0.870	0.922	0.789	0.832	-0.138	-0.179	-0.231	-0.098	-0.141
11121	0.682	0.850	0.850	0.910	0.788	0.880	-0.168	-0.168	-0.228	-0.106	-0.198
11122	0.670	0.722	0.700	0.812	-	0.762	-0.052	-0.030	-0.142	-	-0.092
11211	0.658	0.869	0.840	0.906	0.816	0.867	-0.211	-0.182	-0.248	-0.158	-0.209
12111	0.640	0.834	0.810	0.908	0.807	0.842	-0.194	-0.170	-0.268	-0.167	-0.202
12121	0.478	0.742	0.690	0.798	-	0.789	-0.264	-0.212	-0.320	-	-0.311
12211	0.582	0.767	0.650	0.797	-	0.790	-0.185	-0.068	-0.215	-	-0.208
21111	0.667	0.878	-	0.902	0.777	0.870	-0.211	-	-0.235	-0.110	-0.203
21312	0.455	0.536	0.610	0.680	-	0.630	-0.081	-0.155	-0.225	-	-0.175
22112	0.453	0.662	0.690	0.751	-	0.703	-0.209	-0.237	-0.298	-	-0.250
22121	0.362	0.645	0.650	0.781	-	0.742	-0.283	-0.288	-0.419	-	-0.380
22233	-0.003	-0.142	0.250	0.358	-	0.201	0.139	-0.253	-0.361	-	-0.204
22323	0.178	0.042	0.430	0.252	-	0.359	0.136	-0.252	-0.074	-	-0.181
23232	0.019	-0.084	0.280	0.340	0.399	0.217	0.103	-0.261	-0.321	-0.380	-0.198
23321	0.126	0.147	0.370	0.295	-	0.376	-0.021	-0.244	-0.169	-	-0.250
32223	-0.212	-0.174	0.300	0.135	0.158	0.197	-0.038	-0.512	-0.347	-0.370	-0.409
32232	-0.122	-0.223	0.200	0.100	-	0.147	0.101	-0.322	-0.222	-	-0.269
33232	-0.298	-0.332	0.230	0.203	-	0.055	0.034	-0.528	-0.501	-	-0.353
33323	-0.268	-0.386	0.180	-0.161	-0.009	0.015	0.118	-0.448	-0.107	-0.259	-0.283
33333	-0.339	-0.543	-0.240	-0.708	-0.130	-0.103	0.204	-0.099	0.369	-0.209	-0.236
						Mean absolute differences	0.145	0.243	0.265	0.206	0.238

(67, 115, 117, 121, 164)

The highest MAD is seen in the comparision of Thai and South Korean scores, while the smallest difference is between the Thai and UK scores. Thais directly assigned the highest score to state 11112, as did the Zimbabweans and Koreans. The British and Japanese gave the highest score to state 11211 and the Americans to state 11121. The general population from all six countries assigned the lowest value to state 33333. The lowest score for state 33333 was -0.708 assigned by the Koreans and the highest score for state 33333 was -0.103 from the US population. Compared with the Thai study, the maximum difference in observed scores is 0.528 for state 33232 assigned by Zimbabweans. The minimum difference is 0.021 for state 23321 assigned by British respondents.

Table 8.3 compares the Thai and UK model coefficients (the only coefficients obtained from identically specified models). (67).

Table 8.3 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

(67)

The MAD of the Thai model is twice that of the UK model. The constant term and 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.

The models presented in Table 8.4 are not directly comparable because the variables are defined differently. The dependent variable in the UK-2 and Japanese models is 1 minus the score estimated from the models, whereas that of the South Korean model is the log of 1 minus the score estimated from the model. The dependent variable of the Zimbabwean model is the direct result from the model. The MAD of the Thai model, as seen in Table 8.3, is the highest among four countries: UK, UK-2, Japan, Zimbabwe and South Korea as illustrated in Table 8.4. The smallest MAD is seen in the Japanese model.

Table 8.4 Comparison of the model parameter estimates

UK-2		South Korean		Zimbabwean		Japanese		US	
variables	*coeff.*	*variables*	*coeff.*	*variables*	*coeff.*	*variables*	*coeff.*	*variables*	*coeff.*
a	-0.201	a	-2.680	a	0.900	a	0.148	m1	0.146
difmob1	0.320	m2	0.267	m2	-0.056	m2	0.078	m2	0.558
difmob2	0.391	sc2	0.471	sc2	-0.092	sc2	0.053	s1	0.175
difsc1	0.179	ua2	0.374	ua2	-0.043	ua2	0.04	s2	0.471
difsc2	0.280	pd2	0.318	pd2	-0.067	pd2	0.083	u1	0.140
difua1	0.084	ad2	0.313	ad2	-0.046	ad2	0.062	u2	0.374
difua2	0.156	m3	0.554	m3	-0.204	m3	0.418	p1	0.173
difpain1	0.372	sc3	0.819	sc3	-0.231	sc3	0.101	p2	0.570
difpain2	0.491	ua3	0.662	ua3	-0.135	ua3	0.128	a1	0.156
difmood1	0.271	pd3	0.488	pd3	-0.302	pd3	0.189	a2	0.450
difmood2	0.356	ad3	0.603	ad3	-0.173	ad3	0.108	i22	0.011
ANY13	-0.125	N3	-	N3	-	N3	0.014	i3	-0.122
								i32	-0.015
								d1	-0.140
MAD	0.030	MAD	0.074	MAD	0.049	MAD	0.014	MAD	0.035

(69, 111, 115, 117, 121)

The estimated scores from seven models in the six countries, as well as number of negative scores, number of logical inconsistencies, the second highest and the worst scores and the difference between the full health and the worst health state are presented in Table 8.5.

Table 8.5 Comparison of the estimated scores and other characteristics

	Estimated scores		No.of	No.of	2nd highest	Lowest score	Range
	Mean	SDs	neg. scores	logical incons.	score	(33333)	11111-33333
UK	0.144	0.309	81	0	0.883 (state 11211)	-0.572	1.572
UK2	0.130	0.348	88	60	0.859 (state 21111)	-0.744	1.744
US	0.370	0.224	8	15	0.86 (state 11211)	-0.102	1.102
Japan	0.420	0.217	7	0	0.812 (state 11211)	-0.106	1.106
S. Korea	0.590	0.247	6	0	0.910 (state 21111)	-0.562	1.562
Zimbabwe	0.451	0.206	5	0	0.857 (state 11211)	-0.145	1.145
Thai	0.172	0.264	68	0	0.766 (state11112)	-0.454	1.454

(67, 115, 117, 121, 164)

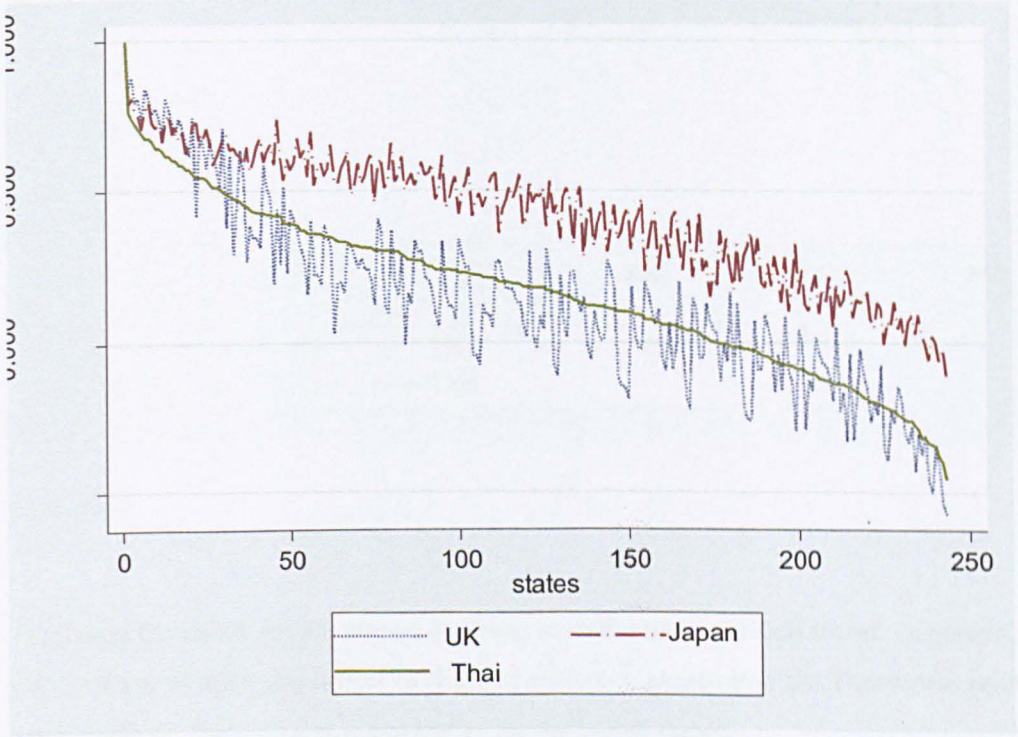
No logically inconsistent scores are presented in the tariffs except those based on the UK-2 and US models. The highest number of negative scores was predicted from the UK-2 model and the smallest number from the Zimbabwean model. The lowest mean of the estimated scores was 0.130 from the UK2 model and the highest mean was 0.590 from the South Korean model. Because the highest score for every model is determined to be 1 for full health, the second best scores are reported here. The second best states differed across the models. The highest second best score was 0.910 for state 21111 from the South Korean model and the lowest second best score was 0.766 for state 11112 from the Thai model. Every model estimated the lowest score for state 33333. The range was from -0.744 for the UK-2 model to -0.102 and -0.106 from the US and Japanese models. Note that the formulae used to transform the scores for states worse than death was different between the US model and the models of the rest of the countries. The minimum range between full health and state 33333 was seen with the US model and the maximum range with the UK-2 model.

Graphical illustrations of the score comparisons

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 five countries (6 models) are presented in Figures 8.2-8.4.

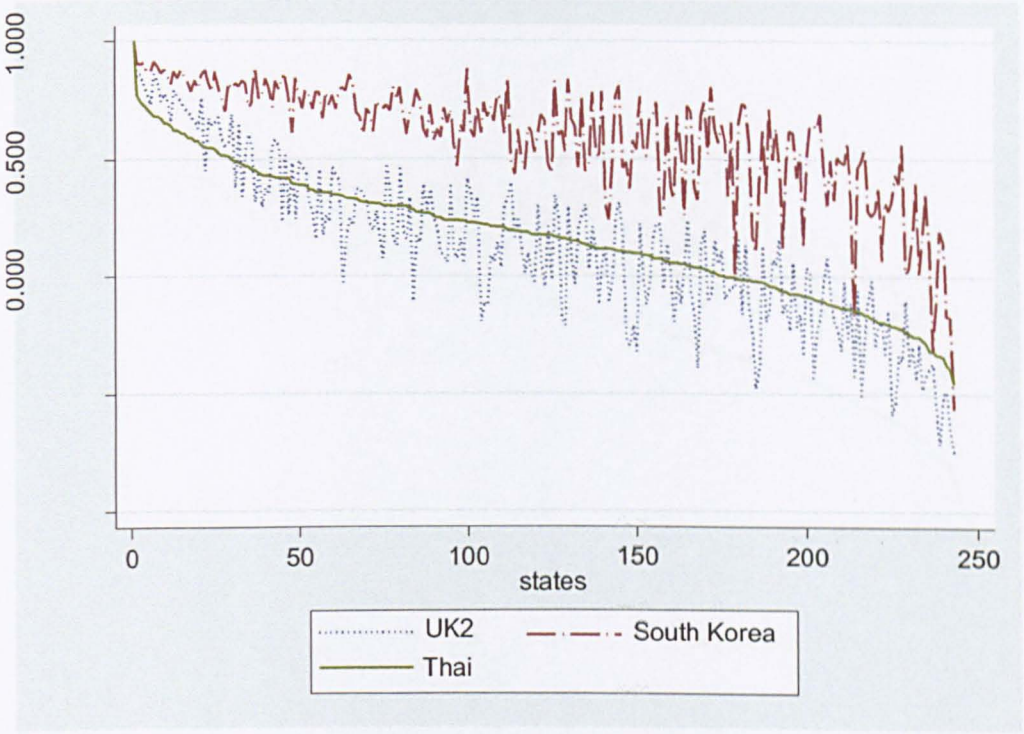
Figure 8.2 **Comparison of the UK, Japanese and Thai scores**



(67, 115)

In general, the Japanese scores are higher than those of the UK and Thai, 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 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 appear higher than those of the UK but are still much lower than the Japanese scores.

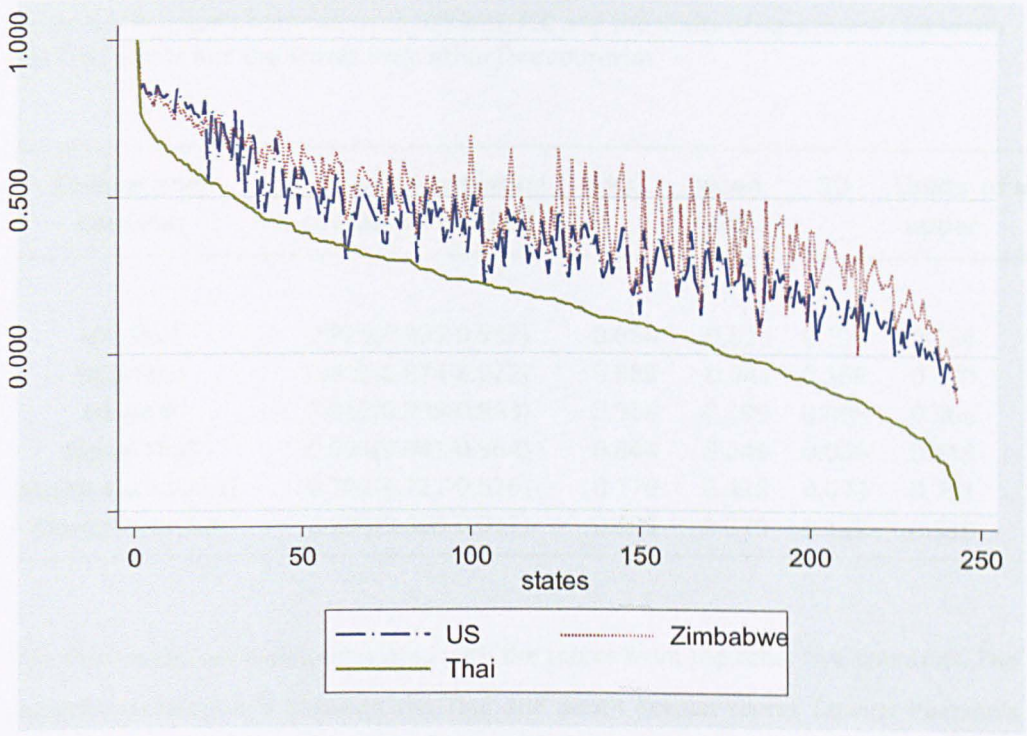
Figure 8.3 Comparison of UK2, South Korean and Thai scores



(111, 121)

The South Korean scores are almost all higher than the UK-2 and Thai scores. In general, the Thai scores are more similar to those of the UK-2, except that the Thai scores tend to be lower for the mild states and higher for the severe states.

Figure 8.4 Comparison of US, Zimbabwean and Thai scores



(69, 117)

The Thai scores are lower than the US and Zimbabwean scores. The US and the Zimbabwean scores are fairly similar for the mild states but the Zimbabwean scores appear to be higher than those of the US in the moderate and severe states.

8.4.2 Level of agreement

The results of the correlations between the estimated scores from Thailand are compared with those of the other five countries in Table 8.6.

Table 8.6 Pearson's correlation coefficient, ICC and the limits of agreements between the Thai scores and the scores from other five countries

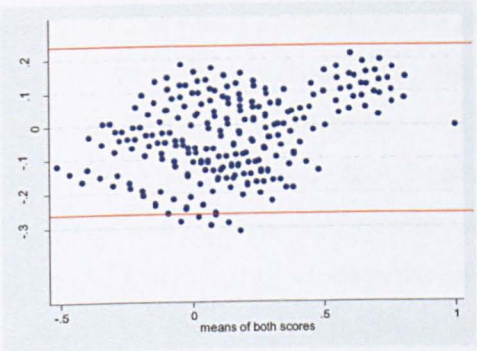
Comparison between	Pearson's correlation coefficient (95% CI)	ICC	Mean Diff.	SD	Limits of agreement upper	lower
UK-Thai	0.925(0.905-0.942)	0.656	-0.028	0.119	0.234	-0.261
UK2-Thai	0.901(0.874-0.922)	0.858	-0.041	0.159	0.270	-0.353
US-Thai	0.952(0.939-0.963)	0.938	0.199	0.085	0.366	0.031
Japan-Thai	0.954(0.941-0.964)	0.844	0.249	0.086	0.418	0.080
South Korea-Thai	0.781(0.727-0.826)	0.779	0.418	0.070	0.751	0.040
Zimbabwe-Thai	0.894(0.866-0.917)	0.925	0.279	0.122	0.519	0.039

The Thai scores are highly correlated with the scores from the other five countries. The weakest correlation is between the Thai and South Korean scores (lowest Pearson's correlation coefficient) and is significantly lower than those between the Thai and the remaining five sets of scores (p-level<0.05). The highest correlation coefficient is seen in the comparison between the Thai and Japanese scores. The highest ICC is identified in the comparison of the Thai and the US scores, whereas the lowest ICC is between the Thai and UK scores. This implies that Thai and British people are likely to vary the least from one another. The Japanese and South Korean scores appear to vary less from the Thai scores than do those derived from the US and Zimbabwean populations.

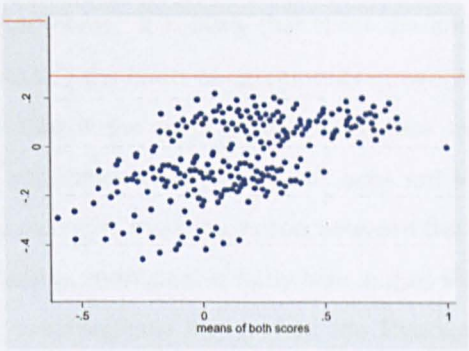
The greatest mean difference between the estimated scores is seen between the Thai and South Korean scores (0.418), with a possible range of the mean difference from 0.040 to 0.751. The smallest mean difference is between the Thai and the UK scores (-0.028), with a possible range from -0.261 to 0.234. The mean difference is negative because the UK scores (for poorer health states) are lower than the Thai scores. The limits of agreement presented in the 6th and 7th columns are presented as the BA plots in Figure 8.5.

Figure 8.5 BA plots of all comparisons

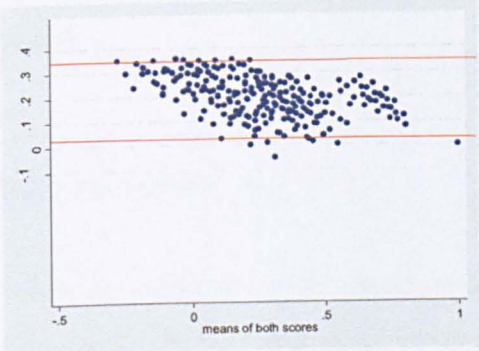
Thai-UK



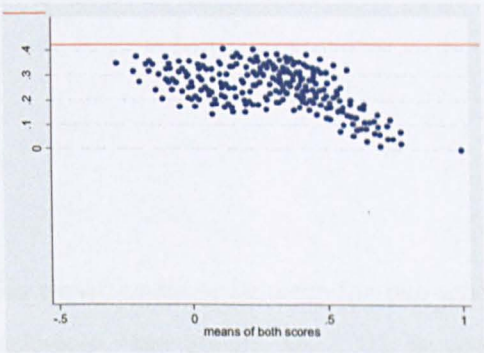
Thai_UK2



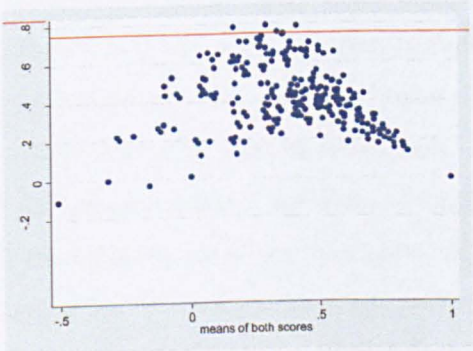
Thai-US



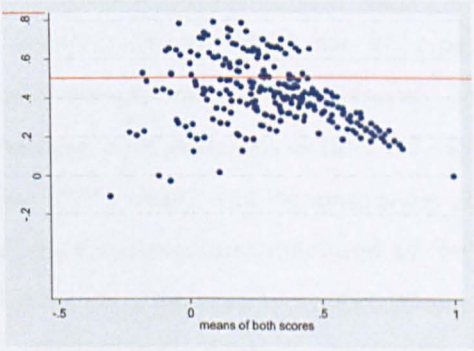
Thai-Japan



Thai-Korea



Thai-Zimbabwe



The X-axis represents the means of the scores from both countries for the same health states, and the Y-axis represents the differences between the comparator scores and the Thai scores. The plots demonstrate the level of agreement between the estimated scores of the two countries.

The horizontal lines in each graph indicate the upper and lower levels of agreement of the modelled scores for 243 states. Although the Thai-UK comparison has the smallest mean difference, as seen in Table 8.5 and the plots, the range of limits of agreement between the two scores was larger than those of the US-Thai and Japan-Thai, which have a greater difference between the mean scores. It is likely that there are greater numbers of health states that are located outside the limits of agreements between the Thai-UK-2 and the Thai-Zimbabwe. Note that if the pairs of the scores are highly correlated, as measured by the Pearson's correlation coefficients, this does not imply that the scores are close to each other. For example, the comparison between the Thai and Zimbabwean scores, in which the correlation coefficient is fairly high, but,as shown in Figure 8.3, the Zimbabwean scores are systematically higher than the Thai scores. The greatest ICC is seen from the Thai-US scores, implying that the Thais and Americansseem to be very different from each other when it comes to preferences for health.

8.4.3 Responsiveness

There are 29,403 pairs of potentially positive transformations between the two scores. The number of pairs with equal scores estimated from the UK, UK-2, US, Japanese, South Korean and Zimbabwean models are 14, 21, 49, 53, 39 and 70 pairs, respectively. The pairs with equal scores were dropped from the data before estimating the Cohen effect size (ES) and the standardized response mean (SRM) for all types of transformation: all positive, major improvement, minor improvement, minor deterioration and major deteriorations. The results are presented in Table 8.7. ES and SRM are presented in the last two columns of the table. The responsiveness scores from all six countries are very high. Of all the transformations measured by the Thai scores, the scores tend to be least responsive for minor improvements (ES 0.95 and SRM 1.11). The responsiveness of the Thai scores is very similar to those of the UK-2 scores for all positive transformations. Compared to the UK scores, the Thai scores seem to be slightly more responsive jn the deterioration transformations. The South Korean scores seem to be least responsive in the all positive and major and minor improvements. However, the scores are slightly more responsive for minor and major deteriorated

transformations. The UK and the US scores are highly responsive in all transitions and the major improvement transformations.

Table 8.7 Responsiveness of the scores from all six countries

All positive transitions

	No.of pairs	baseline mean (SD)	post-treatment mean (SD)	Mean (SD) of difference	Responsiveness ES	SRM
Thai	29,386	0.02(0.21)	0.32(0.22)	-0.30(0.22)	1.43	1.35
UK	29,389	-0.03(0.23)	0.32(0.28)	-0.35(0.27)	1.52	1.30
UK2	29,382	-0.07(0.28)	0.33(0.29)	-0.40(0.29)	1.41	1.36
US	29,354	0.24(1.67)	0.50(0.20)	-0.25(0.19)	1.52	1.33
Japan	29,350	0.30(0.19)	0.54(0.16)	-0.25(0.18)	1.29	1.37
South-Korea	29,364	0.46(0.26)	0.72(0.14)	-0.26(0.23)	1.00	1.13
Zimbabwe	29,333	0.33(0.17)	0.57(0.16)	-0.23(0.17)	1.33	1.36

Major improvement

	No.of pairs	baseline mean (SD)	post-treatment mean (SD)	Mean (SD) of difference	Responsiveness ES	SRM
Thai	7,980	-0.03 (0.24)	0.30(0.23)	-0.33(0.24)	1.40	1.38
UK	7,965	-0.09(0.26)	0.29(0.28)	-0.38(0.29)	1.45	1.33
UK2	7,979	-0.14(0.32)	0.30(0.30)	-0.44(0.32)	1.38	1.38
US	7,954	0.20(0.19)	0.48(0.20)	-0.28(0.20)	1.46	1.36
Japan	7,962	0.25(0.21)	0.53(0.17)	-0.28(0.20)	1.30	1.40
South-Korea	7,967	0.37(0.33)	0.70(0.16)	-0.33(0.30)	1.00	1.11
Zimbabwe	7,954	0.29(0.20)	0.55(0.17)	-0.27(0.19)	1.30	1.37

Minor improvement

	No.of pairs	baseline mean (SD)	post-treatment mean (SD)	Mean (SD) of difference	Responsiveness ES	SRM
Thai	1,432	0.17(0.18)	0.35(0.26)	-0.17(0.16)	0.95	1.11
UK	1,429	0.13 (0.21)	0.35(0.32)	-0.22(0.21)	1.06	1.05
UK2	1,431	0.13(0.24)	0.35(0.33)	-0.22(0.20)	1.13	0.93
US	1,417	0.36(0.15)	0.52(0.23)	-0.16(0.15)	1.03	1.05
Japan	1,425	0.43(0.16)	0.55(0.20)	-0.12(0.09)	0.73	1.22
South-Korea	1,430	0.62(0.16)	0.72(0.17)	-0.10(0.07)	0.63	1.40
Zimbabwe	1,419	0.46(0.14)	0.58(0.19)	-0.13(0.10)	0.88	1.20

Minor deterioration

	No.of pairs	baseline mean (SD)	post-treatment mean (SD)	Mean (SD) of difference	Responsiveness ES	SRM
Thai	4,890	0.35(0.22)	0.03(0.20)	0.32(0.21)	-1.46	-1.50
UK	4,890	0.36(0.28)	-0.02(0.22)	0.37(0.26)	-1.32	-1.44
UK2	4,890	0.37(0.29)	-0.05(0.27)	0.42(0.28)	-1.45	-1.50
US	4,890	0.52(0.20)	0.25(0.16)	0.27(0.18)	-1.36	-1.46
Japan	4,890	0.56(0.16)	0.31(0.19)	0.26(0.17)	-1.66	-1.49
South-Korea	4,890	0.74(0.14)	0.48(0.23)	0.26(0.20)	-1.89	-1.27
Zimbabwe	4,890	0.59(0.16)	0.34(0.17)	0.24(0.16)	-1.55	-1.50

Major deterioration

	No.of pairs	baseline mean (SD)	post-treatment mean (SD)	Mean (SD) of difference	Responsiveness ES	SRM
Thai	11,978	0.30(0.21)	0.02(0.20)	0.29(0.21)	-1.39	-1.39
UK	11,964	0.29(0.26)	-0.04(0.22)	0.33(0.25)	-1.25	-1.31
UK2	11,964	0.31(0.27)	-0.07(0.27)	0.38(0.28)	-1.40	-1.37
US	11,956	0.48(0.19)	0.24(0.16)	0.24(0.18)	-1.28	-1.33
Japan	11,964	0.53(0.15)	0.29(0.19)	0.24(0.17)	-1.60	-1.39
South-Korea	11,969	0.71(0.13)	0.46(0.24)	0.25(0.21)	-1.88	-1.19
Zimbabwe	11,961	0.56(0.15)	0.33(0.17)	0.23(0.16)	-1.50	-1.38

To summarize, compared with the scores from five other countries, the Thai scores are largely similar to those derived from the UK general population using the UK model generated from Dolan (1997). The actual scores assigned by Thais varied least from those of the British. The Thai and UK estimated scores were similar except that the Thai scores for mild states were lower and the scores for severe states were higher than the coordinating British scores. The two scores are highly correlated with the smallest variation between the Thai and UK general population. The range of the level of agreement between the Thai and UK scores is fairly small but still slightly greater than the level of agreement between the Thai-US and the Thai-Japanese scores. The scores for a few health states fell out of the level of agreement. The responsiveness of the UK scores is slightly higher than that of the Thai scores. The Thai and South Korean scores show the largest differences. Figure 8.2 shows that the South Korean scores were higher than the Thai scores. The Thai scores have minimum correlation with the South Korean scores, which suggests the variations between the Thai and South Korean general population are large.

8.4.4 Comparison of the CUA results

The number of Thai patients with cervical cancer in all stages is shown in the second column of Table 8.8. Health states of the patients were classified using an EQ-5D questionnaire and the algorithms from the six countries were used to transform EQ-5D health states into preference scores. Mean preference scores of the patients with respect to the stage of cervical cancer are illustrated in the 3rd column of Table 8.8, followed by the lowest and the highest preference scores in the 4th and 5th columns respectively. The largest number of patients is found in the remission state of CA stage 2 and the second largest number is in the initial state of stage 2 and the remission state of stage 3. Only two patients were categorised into the persistent state of stage 1 and 4 and the remission state of stage 4. The reason for this could be that the survey was administered only once at the out-patient department (OPD) of the university hospital. It is likely that the patients treated as out-patients are those with less serious cervical cancer. The patients in the persistent category of each stage would be in very poor

health, and therefore, they would have been admitted to the hospital or confined at home and not able to attend the OPD.

Table 8.8 Number of patients in each cancer state and mean preference scores

Scores	Obs	Mean	Min	Max	Scores	Obs	Mean	Min	Max
CA stage 1					***CA stage 2***				
Initial stage					**Initial stage**				
UK	125	0.765	0.118	1	UK	140	0.762	-0.016	1
UK2	125	0.748	0.158	1	UK2	140	0.762	0.110	1
US	125	0.814	0.432	1	US	140	0.811	0.307	1
Zimbabwe	125	0.810	0.504	1	Zimbabwe	140	0.815	0.361	1
S. Korea	125	0.870	0.477	1	S. Korea	140	0.874	0.477	1
Japanese	125	0.751	0.434	1	Japanese	140	0.761	0.416	1
Thai	125	0.699	0.195	1	Thai	140	0.712	0.117	1
Remission stage					**Remission stage**				
UK	136	0.858	0.291	1	UK	170	0.811	0.186	1
UK2	136	0.858	0.330	1	UK2	170	0.814	0.187	1
US	136	0.886	0.517	1	US	170	0.855	0.467	1
Zimbabwe	136	0.883	0.596	1	Zimbabwe	170	0.853	0.561	1
S. Korea	136	0.926	0.608	1	S. Korea	170	0.908	0.608	1
Japanese	136	0.841	0.536	1	Japanese	170	0.804	0.529	1
Thai	136	0.815	0.393	1	Thai	170	0.766	0.297	1
Persistence stage					**Persistence stage**				
UK	2	0.863	0.725	1	UK	20	0.719	0.088	1
UK2	2	0.863	0.726	1	UK2	20	0.725	0.086	1
US	2	0.900	0.800	1	US	20	0.787	0.397	1
Zimbabwe	2	0.894	0.787	1	Zimbabwe	20	0.796	0.453	1
S. Korea	2	0.936	0.871	1	S. Korea	20	0.878	0.710	1
Japanese	2	0.854	0.707	1	Japanese	20	0.747	0.469	1
Thai	2	0.847	0.694	1	Thai	20	0.682	0.238	1
Recurrence stage					**Recurrence stage**				
UK	28	0.759	-0.056	1	UK	49	0.741	-0.181	1
UK2	28	0.762	0.078	1	UK2	49	0.741	-0.161	1
US	28	0.809	0.217	1	US	49	0.807	0.205	1
Zimbabwe	28	0.817	0.356	1	Zimbabwe	49	0.812	0.234	1
S. Korea	28	0.864	0.304	1	S. Korea	49	0.872	0.379	1
Japanese	28	0.757	0.094	1	Japanese	49	0.768	0.370	1
Thai	28	0.714	-0.116	1	Thai	49	0.710	0.039	1

Table 8.8 Number of patients in each cancer state and mean preference scores (continued)

Scores	Obs	Mean	Min	Max	Scores	Obs	Mean	Min	Max
CA stage 3					***CA stage 4***				
Initial stage					**Initial stage**				
UK	124	0.692	-0.319	1	UK	34	0.583	-0.016	1
UK2	124	0.691	-0.294	1	UK2	34	0.605	0.11	1
US	124	0.764	0.053	1	US	34	0.696	0.307	1
Zimbabwe	124	0.767	0.121	1	Zimbabwe	34	0.708	0.361	1
S. Korea	124	0.836	0.014	1	S. Korea	34	0.802	0.477	1
Japanese	124	0.713	-0.012	1	Japanese	34	0.667	0.416	1
Thai	124	0.647	-0.253	1	Thai	34	0.559	0.117	1
Remission stage					**Remission stage**				
UK	139	0.825	0.124	1	UK	2	0.788	0.727	0.848
UK2	139	0.826	0.158	1	UK2	2	0.793	0.740	0.845
US	139	0.863	0.382	1	US	2	0.827	0.810	0.844
Zimbabwe	139	0.861	0.496	1	Zimbabwe	2	0.816	0.777	0.854
S. Korea	139	0.915	0.713	1	S. Korea	2	0.892	0.877	0.906
Thai	139	0.776	0.123	1	Japanese	2	0.741	0.691	0.790
Japanese	139	0.812	0.275	1	Thai	2	0.686	0.605	0.766
Persistence stage					**Persistence stage**				
UK	18	0.703	0.137	1	UK	2	0.056	-0.074	0.186
UK2	18	0.704	0.303	1	UK2	2	0.107	0.026	0.187
US	18	0.773	0.375	1	US	2	0.366	0.265	0.467
Zimbabwe	18	0.763	0.457	1	Zimbabwe	2	0.415	0.269	0.561
S. Korea	18	0.818	0.445	1	S. Korea	2	0.527	0.381	0.673
Japanese	18	0.704	0.474	1	Japanese	2	0.429	0.328	0.529
Thai	18	0.620	0.133	1	Thai	2	0.178	0.058	0.297
Recurrence stage					**Recurrence stage**				
UK	38	0.671	-0.380	1	UK	8	0.817	0.587	1
UK2	38	0.668	-0.589	1	UK2	8	0.819	0.567	1
US	38	0.748	0.002	1	US	8	0.853	0.687	1
Zimbabwe	38	0.755	0.086	1	Zimbabwe	8	0.842	0.642	1
S. Korea	38	0.822	0.302	1	S. Korea	8	0.898	0.713	1
Japanese	38	0.694	-0.005	1	Japanese	8	0.791	0.598	1
Thai	38	0.619	-0.256	1	Thai	8	0.737	0.425	1

The Thai algorithm was used to estimate preference scores. The mean scores of the patients were lowest in all cancer stages, except for the persistent stage of CA stage 4, in which the lowest mean score was calculated by the UK algorithm. Mean scores of the patients estimated by the South Korean algorithm were greatest across all stages. Obviously, the greatest difference among preference scores can be seen from those estimated using the Thai and South Korean algorithms. Mean scores estimated by the Thai algorithm were not much different from those estimated by the Japanese

algorithm in all stages of CA stage 1; all stages in CA stage 2, except those in the recurrence stage; the persistent stage in CA stage 3 and the remission and recurrent stages in CA stage 4. The differences were small between the mean scores estimated by the UK algorithm in the initial and recurrence stages in CA stage 2, all stages in CA stage 3 and the initial stage in CA stage 4.

Note that there were some patients who, although classified into the advanced stages of cancer, regarded themselves as in perfect health, hence, the scores are 1. It seems the mean scores of the patients in remission states are slightly higher than those in persistent states of stages 3 and 4, those in the recurrent states of every stage except stage 4 and those in the initial states of all stages. Surprisingly, the mean preference scores of the patients in CA stage 4 are quite similar to the scores of other milder stages. However, given the low number of patients in some categories, these preference scores cannot be regarded as representative of Thai patients with cervical cancer.

As discussed in the last section, the UK and Japanese scores are largely similar to the Thai scores, and the findings from Table 8.8 show that mean preference scores of the cancer patients estimated using the Thai algorithm were not much different from those estimated from the Japanese and UK algorithms. It is possible that the UK and Japanese scores would be preferred to estimate QALYs if the Thai scores were unavailable. This finding contrasts with the suggestion by Sakthong *et al.* that the Japanese scores are the best scores to approximate Thai preferences. To examine the results of using algorithms from the six different countries to estimate QALYs, the estiamted preference scores are presented in Table 8.9.

Table 8.9 Cost and QALYs of the various prevention interventions estimated from the algorithms from six countries

No.	Intervention	Costs (baht)	QALYs UK	Japanese	Thai	S.Korean	Zimbabwean	UK2	US
A	Pap q 5 yrs (45-60) : VIA q 5 yrs (30-40)	8,887.08	28.0738	28.0768	28.0692	28.0916	28.0828	28.0743	28.0828
B	Pap smear every 5 yrs. (30-60yrs.)	9,034.55	28.0726	28.0756	28.0679	28.0907	28.0817	28.0731	28.0818
C	VIA every 5 yrs.(30-45)	9,093.79	28.0711	28.0744	28.0665	28.0896	28.0805	28.0717	28.0805
D	Treatment only (no prevention)	9,605.55	28.0622	28.0661	28.0571	28.0827	28.0727	28.0628	28.0727
E	HPV vaccine at the age of 15+ VIA every 5 yrs.(30-45)+ Pap smear every 5 yrs. (46-60yrs.)	17,449.46	28.1263	28.1271	28.1251	28.1311	28.1287	28.1264	28.1287
F	HPV vaccine at the age of 15 + Pap smear every 5 years (45-60 yrs)+ VIA every 5 yrs (30-40)	17,464.19	28.1262	28.1270	28.1249	28.1310	28.1286	28.1263	28.1286
G	HPV vaccine at the age of 15 + Pap smear every 5 years (30-60 yrs)	17,509.86	28.1259	28.1267	28.1246	28.1304	28.1283	28.1260	28.1283
H	HPV vaccine at the age of 15	17,548.90	28.1232	28.1242	28.1218	28.0907	28.1260	28.1234	28.1260

(12)

Eight interventions (A-H) are ranked in Table 8.9 in terms of increasing cost. Intervention A is the least costly intervention and Intervention H is the most costly. The differences of QALYs of all eight interventions estimated from the algorithms from all six countries were small, ranging from 28.0571 in Intervention D using the Thai algorithm to 28.1311 in Intervention E using the South Korean algorithm. Based on this example, QALYs were similar regardless of which algorithm was used to estimate the preference scores. Compared with Intervention A, all interventions are dominated, with the exception of Intervention E. ICER s of Intervention E are presented in Table 8.10.

Table 8.10 ICERs of Intervention E using Intervention A as a comparator

No.	ICERs						
	UK QALYs	Japanese QALYs	Thai QALYs	S.Korean QALYs	Zimbabwe QALYs	UK2 QALYs	US QALYs
A	*Pap smear 5 years (45-60 yrs) + VIA 5 yrs (30-40) which is currently implemented in the Thai health system*						
B	Dominated	Dominated	Dominated	Dominated	Dominated	Dominated	Dominated
C	Dominated	Dominated	Dominated	Dominated	Dominated	Dominated	Dominated
D	Dominated	Dominated	Dominated	Dominated	Dominated	Dominated	Dominated
E	163,111.61	170,267.84	153,336.89	217,020.86	186,445.45	164,350.07	186,677.09
F	Dominated	Dominated	Dominated	Dominated	Dominated	Dominated	Dominated
G	Dominated	Dominated	Dominated	Dominated	Dominated	Dominated	Dominated
H	Dominated	Dominated	Dominated	Dominated	Dominated	Dominated	Dominated

* Comparator = Intervention A

Compared with Intervention A, the ICERs of Intervention E are 163,111, 170,267, 153,336, 217,020, 186,445, 164,350 and 186,667 baht/QALY using the UK, Japanese, Thai, South Korean, Zimbabwean, UK-2 and US scores, respectively. Note that all ICERs were lower than the Thai threshold at 300,000 baht/ QALY regardless of which preference scores were used. The lowest ICER was seen using the Thai preference scores and the highest ICER was seen using the South Korean scores. Had the Thai willingness-to-pay threshold been changed from 300,000 to 200,000 baht, Intervention E would still have been cost-effective using the scores from all countries except South Korea, where the ICER was greater than 200,000 baht.

The NMBs for Intervention E are 7,187.62 baht, 9,527.62 baht, 8,207.62 baht, 3,138.00 baht, 5,214.92 baht, 7,067.15 baht and 5,238.00 baht using the UK, Japanese, Thai, South Korean, Zimbabwean, UK-2 and US scores, respectively. All NMBs are positive.

8.5 Discussion

Unlike some other studies where the authors only present a comparison between their estimated scores and the UK scores, this chapter reports the comparisons of the estimated scores across seven models from six countries (including the Thai preference scores). The breadth of the comparison is also greater than previously reported; rather than comparing only the model coefficients and the estimated scores, the comparison in this study included correlations between the estimated scores, number of logical inconsistencies, responsiveness and level of agreement between the Thai scores and other countries' scores.

The size of the Thai study is in line with the range for other health state preference elicitations. The duration of interviews in Thailand was slightly longer than those in the UK with a smaller number of health states included. This may imply that the Thai respondents may have had greater difficulties in participating in the preference interview than their British counterparts. One reason might be that the education level of the Thai respondents tended to be lower, thus the respondents may have had difficulty reading the health state descriptions and making time trade-off decisions. This may have caused the Thai respondents to use more time to complete the interview. These findings could be used to assist future studies in planning for the interview, where more help could be provided for respondents with lower education attainment levels to ease the difficulties in participating without interfering with the interview process. Also, fewer health states could be used for the interview in this group of respondents.

Thai preference scores differ from those of other countries. Possible causes of the differences in preferences may be the differences in health systems, cultures and religious beliefs of the populations. Terminology used in the translated versions of EQ-5D may also play some role, because the vocabulary 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 the best scores to estimate the Thai QALYs of all the models presented in this chapter. This assumption is based on the correlations of the scores, differences of the mean estimated scores and the comparison of the predicted scores for all 243 states. The Thais and British seem to elicit fairly similar scores for the same health states, with high correlations between the scores. The states with large differences between the two sets of the scores can be found at the mild and severe states. In using the Thai and UK scores to calculate QALYs in cost-utility analysis, the ICERs are expected to be similar as well. The least favoured sets of scores to be chosen for use in Thai evaluations are the South Korean, Zimbabwean and US scores. Although the US and Zimbabwean scores are highly correlated with the Thai scores, it seems that the population among these four countries differ greatly with respect to their preferences on health, in addition, the differences between preference scores are large.

The differences between the scores may have considerable implications for using the scores from other countries to approximate Thai preferences; the results of economic evaluations could differ from the Thai public perspective. However, based on the real CUA analysis used in this study, the choice of preference scores used in QALYs estimation is unlikely to make a significant difference. All interventions were cost-effective, given that the current willingness to pay threshold is 300,000 baht in Thailand. From Table 8.8, the mean Thai utility scores of the patients with cervical cancer are more similar to the Japanese scores for eight states, more similar to the UK scores for 6 states and similar to both of the scores for 2 states. One reason for this could be that, out of all 243 scores, the Thai scores are more similar to the Japanese scores at the very mild states, but for the worse states, the Thai scores are more similar to the UK scores. For example, the Thai score is closer to the Japanese score for state 11112, but for state 11113, the Thai score is closer to the UK score. An obvious example is state 33322, where the Thai score is -0.237, the UK -0.144 and the Japanese score 0.046. This implies that the Thai and British regard this state as worse than death, whereas Japanese regard this state as better than death. More Thai patients with cervical cancer recruited in this cost-utility study tended to identify themselves as having milder health states; therefore, the ICERs using the Thai scores are more similar to those using the Japanese scores. This finding suggests that if the Thai scores were unavailable, then one may assume that for interventions provided to patients with milder health states, using the Japanese scores would have provided a better approximation to the Thai preferences on health.

However, if greater proportions of patients with poorer health states were recruited, using the UK scores would have provided a slightly better representation of the Thai preferences on health.

The results from this study showed that regardless of whether using the US, UK, Japanese, Zimbabwean, South Korean and Thai preference scores to estimate QALYs, the ICERs were not much different. This implies that although preference scores from different countries are, to some extent, systematically different, when they are used to calculate ICERs, the impact on ICERs is marginal.

The finding from this CUA example differs from the suggestion from Sakthong *et al.*, where the Japanese scores were recommended by the authors to be used to calculate QALYs in the Thai settings. It seems that by using the UK scores, rather than the Japanese scores, the ICERs may be similar to those estimated from the Thai scores. However, to definitely conclude this finding, further studies should be conducted using other examples with larger numbers of patients or using the data from randomized controlled trial (RCT) studies.

The results of this example should be treated with caution. Table 8.8 shows that the preference scores from patients in some poorer stages to be slightly higher than those in better stages. There could be two explanations for this finding. One would be that the scores were derived from a few patients who may not be representative. The other possible explanation would be that the patients in the poorer stage may have had some experiences of being in the milder states before, therefore, they had already "adapted" themselves to the symptoms and the consequences. Thus, their health states were not as affected as what may have been perceived by the general population. For example, they would have assigned level 2 to pain/discomfort dimension because they would have already been adapted to the symptoms experienced. A new patient or a person who has no previous pain experience would have assigned level 3 to the dimension, given that they had not adapted to the suffering. Please note that this result is based only on one example and before establishing a definite conclusion, further research is required in order to be more certain about the implications of using the scores from other countries in a Thai setting. Moreover, the methodology of deriving EQ-5D states from the patients may need to be changed from the current study in which the patients were surveyed cross-sectionally.

For researchers in the countries where preference scores are unavailable, by generating the criteria to select the "best alternative" scores to represent the preferences of their own settings, more caution should be paid in choosing the selection criteria. At the time of the study by Sakthong *et al.*, Thai scores was not yet available, so it could have been difficult for the researchers to develop the selection criteria for the preference scores to be used in Thailand. The researchers suggested that the level of agreement ought to be used to select the appropriate alternative set of preference scores with which to estimate Thai preferences. As shown in this study, this criterion may not be appropriate to substitute for the preference scores, given that utility scores have a limited range between 0 to 1 for states better than death and 0 to -1 for states worse than death. The range of mean differences between the two sets of scores may be, for example, from the lowest boundary at 0.040 to 0.751 in the Korea-Thai scores correlations. With this substantial range in preference scores, the CUA results could have been massively varied.

Results of the correlation tests between the Thai scores and the scores from other countries may contradict each other in that the Pearson correlation coefficients may be high, implying that the preferences scores between the two sets are highly correlated. However, the ICC may also be high, implying that the preferences surrounding health of the two countries are very different from each other. Compared with the UK, the Thai preferences seem to be more affected by moves away from full health, i.e., having extreme health problems in at least one dimension and having severe problems in mobility and self-care. Note that the coefficients of these two dimensions are equal. The UK preferences are likely to be affected mostly by having extreme problems in at least one dimension, having severe problems in mobility, and having extreme pain and discomfort.

The Thai scores also have a narrower range than those of the UK, but wider than those of the US, and therefore one may question the responsiveness of the Thai preference scores. Note that the ES of the US and UK calculated in this study are slightly different from those estimated in the study by Luo *et al.*. However, when comparing the ES and SRM using all possible pairs of EQ-5D states, the responsiveness of the Thai scores tends to be similar to that of the other countries. The reason would be because the differences between the baseline and the post-treatment means in the Thai scores are smaller, but the SDs of the Thai scores are also smaller. The Thai scores seem to be less

responsiveness when used in the minor improvement transformation. The assessment of responsiveness in this study must be treated with caution. By including all possible transitions, the responsiveness measures are high, because the difference between the baseline mean and the post-treatment mean is high. For example, the responsiveness measures include dramatic changes such as moving from 33333 to perfect health. This is unlikely to be justified since in a real evaluation a very restricted subset of transitions is generally present. The responsiveness of using EQ-5D to measure patients' quality of life in individual diseases or individual patients groups should be tested and compared with the responsiveness from other diseases. Interestingly, the South Korean scores have a wide range of the best-worst score but the responsiveness of the scores are lowest for the all positive transformations among the seven sets of scores.

8.6 Conclusion

The Thai preference scores differ from those estimated from other countries. These differences were explored using the following dimensions: observed and modelled scores, the model coefficients, graphical comparisons of the scores from all 243 states, Pearson's correlation coefficients, ICC, limits of agreement and responsiveness. It seems that Thai preferences are similar to those of the UK population, except for mild and severe states, where the Thai preferences seem to be lower for mild states and higher for severe states. . Therefore, the UK scores are the most appropriate scores to use to approximate Thai preferences. The Zimbabwean and the South Korean scores provide the poorest approximation to Thai preferences. The responsiveness of the Thai scores measuring the positive transformations of health states seems to be similar to other sets of scores. This chapter also shows that based on the analysis used in the study, when using the scores derived from the population of another country to represent Thai preferences, the ICERs do not differ greatly. However, further research should be conducted in order to offer more insight on this issue. Caution should be taken when using the scores elicited from other countries to represent preferences of health because a particular set of scores may be suitably applied to a particular group of patients, but other sets may not be.

This chapter suggests that it is important to establish preference scores from the general population in any given country. Using the scores elicited from other countries may

produce misleading results in a cost-utility analysis. Countries which have not yet established their own preference scores should be cautious when selecting the best alternative preference scores. The limits of agreement may not be appropriate to examine the similarity of the scores between two countries, given that preference scores are defined to range between 0 to 1 for states better than death and 0 to -1 for states worse than death.

Chapter 9 Discussion and summary

9.1 Introduction

This thesis reports 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.

9.2 Contribution

There are four major contributions offered by this research: it provides the first set of Thai population-based preference scores for health; it attempts to identify possible causes of logical inconsistency and the treatment of logical inconsistency; it details an international comparison of preference scores and it demonstrates successful administration of a population-based preference survey in Thailand. Details of each contribution are as follows.

9.2.1 The first set of Thai population-based health preference scores

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. 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, as stated in this study, differ from those used in the previous preference studies. The criteria used in other studies were: logical consistency

of the modelled scores; model parsimony and robustness. As stated in Chapter 7, 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 were consistent, an innovative method to detect logical inconsistencies in the modelled scores was proposed in Chapter 7 using Stata. Although the estimated scores were thoroughly examined only until a certain level, it is likely that almost all of the possible inconsistent 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 because it 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 Thailand in that QALYs of the Thai population can 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 (4). Researchers in the economic evaluation field can be reassured that the Thai preference scores were estimated using standard methodologies and that the scores were derived from a large population-based survey. The Thai preference scores could also contribute as an additional set of scores in the list of available preference scores provided by the EuroQol group.

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

9.2.2 Attempts to identify possible causes of, and alternative treatments for, logical inconsistency

Simply identifying the number of inconsistently valued health states is not sufficient to cast light on possible causes of logical inconsistency. This study offers two further approaches, quantitative and qualitative, to gain more insight into this issue. Regarding the quantitative analysis, two original attempts have been made in this study. Firstly, statistical analysis was used to estimate the determinants of the number of logical inconsistencies among the Thai respondents, which was demonstrated in Chapter 5. Additional knowledge has been provided by the analysis in that, apart from the respondent demographic characteristics (age and educational level), the combinations of health states in each health set, interviewers and interview methods have significant effects on the number of logically inconsistent responses. The results of this analysis can be used to guide the preparation of future preference elicitation research, so as to reduce the numbers of logically inconsistent responses. Secondly, an attempt to distinguish a "pattern" of "when" the inconsistencies occur is made in Chapter 6. It is likely that the more inconsistent respondents may reveal higher numbers of inconsistencies in the second half of the interview whereas those with fewer inconsistencies tend to have more inconsistencies near the beginning of the task. This finding cannot be strongly concluded because the differences between the first and the second half of the interview were not statistically significant. However, this addresses an interesting question on whether the respondents tended to "learn" how to respond to the questions at the beginning of the interview or were "overwhelmed" with the complexities of the tasks and became more "inconsistent" in the later part of the

interview. Further understanding of this issue would pave the way to discover strategies to minimise logical inconsistencies in the preference elicitation interview.

The Thai respondents seem to make more inconsistent responses than respondents in some studies, but fewer than those in other studies. The Thai respondents differed from those participating in other studies in that the Thai respondents tended to have a lower education level and most of them had no previous experience of a preference elicitation interview. One possible explanation for the lower numbers of inconsistency in other studies would be that fewer health states were used in these preference interviews, leading to fewer inconsistencies being generated. In addition, the perceived plausibility of the health states and the combinations of health states in the sets may have played some role in the generation of the inconsistent responses. The combinations of more plausible health states with explicit differentiation between health states could ease the interview tasks, thus producing fewer inconsistent responses. The types of interview methods also influence the degree of logical inconsistency; the more complex interview tasks tend to generate greater numbers of inconsistent responses.

This is the first study using a qualitative approach to explore the issue of logical inconsistency in a Thai context. Although the qualitative interview in this study is exploratory, the results provide an initial understanding on two important issues: the possible strategies used to cope with the complexities of the preference interview and what lies behind respondents' decision making on sacrificing time in poor health to live in full health. The respondents revealed that they may have had difficulties understanding the complex tasks and some may have "learned" how to respond the tasks. Moreover, the respondents tended to use only partial information from the EQ-5D health states, as well as external information, to help with their decision making when trading off time. These findings can be used as background to generate some hypotheses to explain coping strategies, and this could be explored further in future qualitative studies.

When considering the appropriate number of inconsistent responses upon which a decision to exclude respondents should be based, prior to the estimation of the models, the recommendations by Lamers *et al.* and Ohinmaa and Sintonen are inapplicable to the Thai study. Almost seventy percent of the respondents had more than three inconsistent values and would have been excluded following the recommendations by Ohinmaa & Sintonen. It would also be undesirable if the recommendations by Lamers *et*

al. were to be followed because the model would then be estimated using the scores assigned by the highly inconsistent respondents with a very high number of inconsistent values. Other studies would have encountered similar challenges if these recommendations were implemented. In other studies reasons for the exclusion of inconsistent scores were not stated explicitly. However, in this study the effects of logically inconsistent responses were thoroughly explored before a decision was made on the appropriate inconsistency threshold for exclusion. The models were used to estimate the preference scores only after the exclusion threshold was established. Additionally, after the effects of the exclusion on actual scores were rigorously explored, as shown in Chapter 6, the threshold used in this study was shown to be higher than that proposed by Ohinmaa and Sintonen. This is justified because in Chapter 5 it was shown that the highly inconsistent respondents tended to have greater difficulties in assigning scores to health states, therefore, the excluded respondents would be the respondents who probably assigned scores randomly. It is unlikely that the scores given by the highly inconsistent could actually represent their preferences for health.

Classifying the respondents in this study into more than two subgroups broadened the insights about the consequences of excluding respondents on the basis of logical inconsistency for mean scores, model specifications and modelled scores. This study classified the respondents into four subgroups and the models were estimated from the scores of these four subgroups. This approach is in line with the view of Devlin *et al.* that inconsistent responses should not be excluded until the consequences of doing so are thoroughly explored(115).

9.2.3 International comparison of preference scores

An extensive comparison of health state values from five other countries has been made in this study. Thai health values differ from those assigned by people in other countries. Using preferences from other countries to estimate QALYs gained by health interventions in Thailand could produce misleading results. Thai health state values drop dramatically if there is a deviation from "no problem" in mobility. The scores estimated from Thais for the second best health state are lower than those from the populations of the other five countries. This evidence can be used to support an

argument that the Thai scores suffer from a ceiling effect. Researchers from countries where preference scores are not yet available can use the results of this study to convince their policy makers to encourage the undertaking of preference elicitation studies. In the meantime, if it is necessary to use scores estimated from other countries, the selection process for the alternative scores should proceed with care. One set of scores might be more appropriate for the analysis of a disease involving more patients in mild health states, whereas another set of scores might be more appropriate for the analysis of those diseases that involve severe health states. Some criteria may not be appropriate to be taken into account, as shown in Chapter 8.

9.2.4 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 and 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 have been able to travel to the research office. The key factors for the successful completion of the preference elicitation interviews rest on a combination of the 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 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 to invite all the respondents to be interviewed in the researcher's offices. Moreover, if the interviews were scheduled to be conducted in the respondents' households, some may have felt 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 was 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 changes were made to 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 Thailand. 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.

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 7, indicate that there were a considerable number of states with differences between the actual and predicted scores; this can be used to support the argument that simply increasing the number of health states in the preference interview may not improve the model performance. This finding may shed more light on the limitations of the EQ-5D measure for describing very mild and 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 the Thai study, there was no level 3 in "mild" states, as opposed to the "mild" states categorised in Dolan, in which level 3 was seen in up to two dimensions. The poorest health state scored by the British in the "mild" category was state 21333 with an estimated score of -0.110, as opposed to the estimated score of 22211for the poorest state in the Thai study , with a score of 0.497. Regarding the "severe" state classification, there was no level 1 used in the "severe" states in the Thai study, whereas the best "severe" state was valued at 13332 in the UK study. The estimated score for this state in the UK study 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 all health states used in the interview were grouped into 12 sets with 11 health states in each set, unlike some previous studies where health states were randomly chosen to be presented to respondents. 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 in the UK study), it would be easier for the interviewers to simply choose at random one of the eleven health states already prepared in the sets. The third advantage is that the effects of the combinations of health states can be explored, as shown in Chapter 5, in regard to the determinants of the numbers of inconsistencies. Although the health states were combined to ensure similarity between the sets, it is likely that some health sets are more or less complex than others.

9.3 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; difficulties in accessing data from the National Statistical Office (NSO) Thailand; and number of observations per health state.

9.3.1 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 have understood the interview task and may have been 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 inconsistent responses to cause 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. This resulting model predicts the scores with some degree of error, in that approximately ten percent of all the health states have 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.

9.3.2 Modifications to the original MVH protocol

Regarding the respondent performance in the preference elicitation interview, it seems that some Thai respondents may not have understood 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 allowed 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 dead" 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 the chance 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 dead" 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.

9.3.3 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; or the illness experience of the respondents. It have been difficult for the respondents to imagine themselves being in the hypothetical health states. The descriptions of the health states may have increased the cognitive burden on respondents. The health state descriptions in Thai appeared to have been 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 have been 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 have confused the participants, especially those who are elderly with only primary education and poor reading ability. They may have thought that either mobility or self-care was at level 1 (positive "*Mai*") in state 31311 or that both dimensions wer at level 3 (negative "*Mai*").

As stated by Brazier *et al.*, values are sought from the "informed general population"(163). 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 have been "well informed", but when it came 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. An 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.

9.3.4 Time horizon for the TTO questions

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 has 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 may have been 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 studies 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.

9.3.5 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 in which 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 respondents with secondary level and university education were higher and the proportion with primary level education was lower. This implies that more of the respondents with a 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 have been different from the findings presented in this study. It is possible that other groups of elderly respondents, especially those with higher educational attainment, would have assigned consistent scores because they could have understood 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 have also had higher number of inconsistencies.

9.3.6 Number of interviewers and the interview sites

There are two major limitations regarding the interviewers, firstly, the numbers of interviews per interviewer were not equal. The second limitation, as seen in Chapter 5, is that there was a significant interviewer effect on the extent of logical inconsistency, although the interviewers had previously been trained in the interviewer training workshops. One possible cause of the first limitation is that a large proportion of the interviewers were recruited from Master's degree and first-degree students who were not available throughout the duration 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 at the universities. The reason why additional interviewers

were recruited in the middle of the fieldwork was to solve this 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 generated 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 these 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 workplace was a shop 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 could hardly have been 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 targeted respondents as possible 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, or they had to take care of their children, as well as their household chores.

9.3.7 Difficulties in accessing data from the NSO

Although good collaboration was given by the NSO regarding the random selection of respondents from the Health and Welfare Survey (HWS) 2007 to be interviewed in the current study, there were unfortunately a couple of difficulties when HWS data were requested for merger with the data from the preference survey. It was originally planned that only necessary data on respondent characteristics should be collected in the preference survey in order to allocate a greater proportion of interview time to the preference elicitation interviews. In addition, it would have made the respondents uncomfortable to have been asked the same questions that they had previously

answered in the HWS. Therefore, to shorten the interview, respondents were asked to supply only very limited personal information. It was planned to obtain other relevant data by linking with the NSO database, using the bridging codes created by the NSO to identify the respondents to the preference survey in the HWS database. Bridging codes are the codes assigned to the respondents' addresses, including: region; province; city; area; primary sampling unit; household number and household sequential number of the respondent.

The first difficulty arose as a result of a recently launched information confidentiality policy to protect the respondents' personal information after the fieldwork survey. The NSO decided to rearrange the data coding of the HWS in their data entry process. The re-arrangement occurred after the completion of the preference survey. The updated bridging codes of the HWS data now only include region, area and new household number. As a result, the respondents in the HWS data cannot be completely linked with the preference survey. This problem was alleviated by using other codes to merge the NSO data namely, region, area, gender, age and household sequential number. Unintentional mistakes could have been made if more than one respondent had the same new bridging codes. However, given the current situation, this was the best option available.

The second difficulty was the amount of time spent waiting for the HWS data to be sent. Although the data entry process for the preference survey was completed in Thailand by September 2007, the HWS data could not be merged until the researcher returned to the UK and started the data analysis process in the beginning of 2008, because the HWS data could not be sent until the data were ready to be analysed by the NSO statisticians. Had the data been sent earlier, this would have accelerated the segment of the data analysis procedure involving the demographic characteristics of the respondents.

9.3.8 Number of observations per health state

As seen in Chapter 4, the numbers of observations per health state ranged from 95 to 1,313. It is seen as a limitation that some health states had less than 200 observations; two-hundred observations per health state could not have been achieved in some states, given budget and human resource constraints in the fieldwork survey. The achieved number 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 (106). Another limitation is that the number of observations per health set was

not equal for all sets. This resulted from 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 is not much different than planned.

9.4 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: minimisation of the cognitive burden on the respondents; modelling health state values for further subgroups of respondents; recruitment of a sample that is more representative of the general population; improvement of fieldwork management; and the potential use of the new version of EQ-5D. Details are as follows.

9.4.1 Minimisation of 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 more tools. As stated in the self-completed questionnaire, some respondents could not 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 the questionnaires. Visual representations as described in

Hadorn *et al.* may be able to assist respondent comprehension(164). 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, test-retest reliability and numbers of counter-intuitive results should be examined before implementing the graphic illustrations. Some pictures may be emotionally offensive for some respondents. A friendly computer-based interview program will 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 feasibility and reliability before implementation in interviews. For those who may have difficulties or may be unfamiliar with using a computer, an assistant will be provided to help.

The second hypothesis is that 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 of health states. One of the limitations of this method for TTO is that the overall interview duration would be increased, as the respondents would be required to give scores for the warm-up session, in addition to the regular health set. This limitation could be resolved by using fewer health states in the Ranking and VAS stages.

9.4.2 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 as "all respondents" and one or more subgroups with various numbers of inconsistencies. If the 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 would 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 made in 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(39, 41). 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(114).

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 time and diminishing marginal utility. The methods used in the study by Attema & Brower could be adapted to address the latter issue to find a potential strategy to "correct" the Thai preference scores (80). The new TTO question format, as suggested by Robinson & Spencer and Devlin *et al.*, could be used to elicit the scores for states worse than death, and to place the negative scores on the same scale as the states better than death (72, 165).

9.4.3 Recruitment of a sample that is more representative of the Thai general population and improvement of fieldwork management

Female respondents tended to be over-represented in the sample. This may have resulted 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 appointments

with the respondents; 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 have been 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 of the interview. 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, with 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 into 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 then 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 will be used in the next study, full-time interviewers are needed who can 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 some interviewers to gain experience, who would otherwise have

less experience due to fewer numbers of interviews performed. 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 have completed 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 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.

9.3.4 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 will launch the new version of EQ-5D (EQ-5D-5L). One of the key topics of interest for the 26th EuroQol Group Scientific Meeting, September 2009, is the research on the five level descriptive system(166). The launching of the new EQ-5D version could be 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 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 and qualitative studies on how Thais respond to the elicitation interview, following what has been learnt from this study.

There could be more challenges waiting 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 questions 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 and how many respondents would be needed in the sample. More levels of health descriptions means greater complexity 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, in such a case, 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 was 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.

9.5 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, but 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 were 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. Close collaboration between the several organisations involved in the research was a key factor in conducting the fieldwork successfully. 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 semantic and conceptual equivalence in tune with 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. With a greater number of health states being valued, problems may arise with an increase in logical inconsistency.

Reference

1. WHO. The World Health Report 2008. Geneva, Switzerland: The World Health Organization, 2008.
2. Davis K, *et al.* Slowing The Growth of U.S. Health Care Expenditures: What Are The Options? . The Commonwealth Fund/Alliance for Health Reform 2007 Bipartisan Congressional Health Policy Conference 2007.
3. O'Connor S. US health reform. Financial Times, 2009.
4. Teerawattananon Y, *et al.* Historical development of health techonology assessment in Thailand. International Journal of Technology Assessment in Health Care. 2009; 25: 1-12.
5. Wibulpolprasert S. Thailand Health Profile 2005-2007. Bangkok: Bureau of Policy and Strategy, Ministry of Public Health, Thailand, 2007.
6. NSO. The 2007 Health and Welfare Survey. Bangkok, Thailand: National Statistical Office, Thailand, 2007.
7. Tangcharoensathien V, Teerawattananon Y, Prakongsai P. The Budget for capitation in National Health Insurance Scheme: Where 1,202 baht comes from? Journal of Health Science. 2001; 10.
8. Tisayaticom K, *et al.* Cost Analysis and Estimate of Capitation Rate for Universal Health Care Coverage Project for Fiscal Year 2004 Journal of Health Science. 2004; 12: 907-22.
9. NHSO. The National Health Security Office increases medical capitation to 12 baht per head next year. National News Bureau of Thailand. Bangkok: http://thainews.prd.go.th/en/news.php?id=255010050001, accessed 24 June 2009, 2008.
10. McPake B, Kumaranayake L, Normand C. Health Economics: An International Perspective. London: Routkedge, 2002.
11. Gold M, *et al.* Cost-Effectiveness in Health and Medicine. Oxford: Oxford University Press, 1996.
12. Drummond M, *et al.* Methods for the Economic Evaluation of Health Care Programmes. 3rd ed. Oxford: Oxford University Press, 2005.
13. 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.
14. Rutten F. Economic evaluation and health care decision-making. Health Policy. 1996; 36: 215-29.
15. Johannesson M. Economic evaluation of health care and policy making. Health Policy. 1995; 33: 179-90.
16. Chiawchanwattana A, *et al.* Cost utility of renal dialysis in Thailand Journal of The Nephrology Society of Thailand. 2003; 9: 158-69.
17. 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.
18. 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.
19. 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.

20. 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.
21. 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.
22. 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.
23. Badia X, *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.
24. NICE. Guide to the Methods of Technology Appraisal. London, UK, 2008.
25. Suksiriserekul S. The cost-utility analysis of some Thai public health programmes. Department of Economics and Related Studies. York: University of York, 1994.
26. Froberg D, Kane R. Methodology for measuring health-state preferences-I: Measurement strategies. Journal of Clinical Epidemiology. 1989a; 42: 345-54.
27. Johannesson M, Jonsson B, Karlsson G. Outcome measurement in economic evaluation. Health Economics. 1996; 5: 279-96.
28. Dolan P, Stalmeier P. The validity of time trade-off values in calculating QALYs: constant proportional time trade-off versus the proportional heuristic. Journal of Health Economics. 2003; 22: 445-48.
29. Gold M, *et al.* Identifying and Valuing Outcomes. In: Gold M, Siegel J, Russell L, et al., eds., Cost-Effectiveness in Health and Medicine. Oxford: Oxford University Press, 1996.
30. Froberg D, Kane R. Methodology for Measuring Health-State Preferences-II: Scaling Methods. Journal of Clinical Epidemiology. 1989b; 42: 459-71.
31. Cook K, *et al.* A psychometric analysis of the measurement level of the rating scale, time trade-off and standard gamble. Social Science&Medicine. 2001; 53: 1275-85.
32. Dolan P, Olsen JA. Distributing health care: Economic and ethical issues. Oxford: Oxford University Press, 2002.
33. WHO. Preamble to the Constitution of the World Health Organization as adopted by the Interanational Health Conference. Interanational Health Conference. New York, 1948.
34. Kaplan R, Anderson J. A General Health Policy Model: Update and Applications. Health Services Research. 1988; 23: 203-35.
35. Patrick D, Bush J, Chen M. Methods for Measuring Levels of Well-being for a Health Status Index. Health Services Research. 1973; 8: 228-45.
36. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. Journal of Health Economics. 2002; 21: 271-92.
37. Kaplan R. Chapter 5:Profile versus utility based measures of outcome for clinical trials. In: Staquet M, Hays R, Fayers P, eds., Quality of Life Assessment in Clinical Trials: Methods and Practice. Oxford: Oxford University Press, 1998.
38. Brazier J, *et al.* Deriving a Preference-Based Single Index from the UK SF-36 Health Survey. Journal of Clinical Epidemiology. 1998; 51: 1115-28.
39. Torrance G, Boyle M, Horwood S. Application of Multi-Attribute Utility Theory to Measure Social Preferences for Health States. Operations Research. 1982; 30: 1043-69.
40. Torrance G, *et al.* Multiattribute Utility Function for a Comprehensive Health Status Classficiation System: Health Utilities Index Mark 2. Medical Care. 1996; 34: 702-22.

41. McCabe C, Stevens K, Brazier J. Utility Scores for the Health Utilities Index Mark 2; An Empirical Assessment of Alternative Mapping Functions. Medical Care. 2005b; 43: 627-35.
42. Feeny D, *et al.* Multiattribute and Single-Attribute Utility Functions for the Health Utility Index Mark 3 System. Medical Care. 2002; 40: 113-28.
43. Hawthorne G, *et al.* The Assessment of Quality of Life(AQoL) Instrument: Construction, Initial Validation & Utility scaling. The Assessment of Quality of Life(AQoL) Instrument. Australia, 1997.
44. Richardson J, *et al.* The Assessment of Quality of Life(AQoL)2 Instrument: Overview and Creation of the Utility Scoring Algorithm. Melbourne: Monash University, 2007.
45. EuroQol. EuroQol - a new facility of health-related quality of life. Health Policy. 1990; 16: 199-208.
46. Brooks R. EuroQol: the current state of play. Health Policy. 1996; 37: 53-72.
47. Konig H-H, *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.
48. Kind P, Brooks R, Rabin R. EQ-5D concepts and methods: a development history. Netherlands: Springer, 2005.
49. Szende A, Oppe M, Devlin N. EQ-5D value sets: inventory, comparative review and user guide. London: Springer, 2007.
50. Brazier J, *et al.* A Review of the Use of Health Status Measures in Economic Evaluation. Health Technology Assessment. 1999; 3.
51. Savoia E, *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.
52. Misajon R, *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.
53. 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.
54. Rasanen P, *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.
55. Brauer C, *et al.* Trends in the Measurement of Health Utilities in Published Cost-Utility Analyses. Value in Health. 2006; 9: 213-18.
56. 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.
57. Scuffham P, *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.
58. Stein K, *et al.* What Value Health? Applied Health Economics and Health Policy. 2005; 4: 219-28.
59. Sakthong P. Measurement of Clinical Effect: Utility. Journal of Medical Association of Thailand. 2008; 91: S43-S52.
60. 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.

61. Leelakulthanit O. Quality of life in Thailand. Social Indicators Research. 1992; 27: 41-57.
62. 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.
63. Hunt S, *et al.* Cross-cultural adaptation of health measures. Health Policy. 1991; 19: 33-44.
64. Cheung YB, Thumboo J. Developing Health-Related Quality-of-Life Instruments for use in Asia. Pharmacoeconomics. 2006; 24: 643-50.
65. Ryan M, *et al.* Eliciting public preferences for healthcare: a systematic review of techniques. Health Technology Assessment. 2001; 5.
66. Torrance G, Thomas W, Sackett D. A Utility Maximization Model for Evaluation of Health Care Programs. Health Services Research. 1972: 118-33.
67. Dolan P. Modeling Valuations for EuroQoL Health States. Medical Care. 1997; 35: 1095-108.
68. Patrick D, *et al.* Measuring Preferences for Health States Worse than Death. Medical Decision Making. 1994; 14: 9-18.
69. 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.
70. Gudex C, Kind P, Dolan P. The Valuation of Death. The 9th Plenary Meeting of the EuroQol Group. Helsinki, Finland, 1992.
71. Charro Fd, *et al.* Some considerations about negative health states for EQ-5D health states. The 12th Plenary Meeting of the EuroQol Group. Barcelona, Spain, 1995.
72. 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.
73. Robinson A, Spencer A. Exploring challenges to TTO utilities: valuing states worse than dead. Health Economics. 2006; 15: 393-402.
74. Dolan P, *et al.* Valuing health states: A comparison of methods. Journal of Health Economics. 1996; 15: 209-31.
75. Bleichrodt H, *et al.* A consistency test of the time trade-off. Journal of Health Economics. 2003; 22: 1037-52.
76. Froberg D, Kane R. Methodology for measuring health-state preferences-III: Population and context effects. Journal of Clinical Epidemiology. 1989; 42: 585-92.
77. Buckingham K, Devlin N, Tabberer M. A Theoretical Framework for TTO Valuations and A Taxonomy of TTO Approaches: Results from a Pilot Study. London: Department of Economics, City University 2004.
78. Spencer A. The TTO Method and Procedural Invariance. Health Economics. 2003; 12: 655-68.
79. Dolan P, *et al.* An Inquiry into the Different Perspectives That Can Be Used When Eliciting Preferences in Health. Health Economics. 2003; 12: 545-51.
80. Sutherland H, *et al.* Attitudes Toward Quality of Survival: The Concept of "Maximal Endurable Time". Medical Decision Making. 1982; 2: 290-309.
81. 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.
82. Buckingham K, Devlin N. A note on the nature of utility in time and health and implications for cost utility analysis. Social Science and Medicine. 2009; 68: 362-67.
83. Ryan M, Farrar S. Using conjoint analysis to elicit preferences for health care. British Medical Journal. 2000; 320: 1530-3.
84. Lancsar E, Louviere J. Conducting Discrete Choice Experiments to Inform Healthcare Decision Making. Pharmacoeconomics. 2008; 26: 661-77.

85. Ratcliffe J, *et al.* Using DCE and Ranking data to estimate cardinal values for health states for deriving a preference-based single index from the sexual quality of life questionnaire. Health Economics. 2009; Published online. DOI. 10.1002/hec.1426.
86. Ryan M, *et al.* Using discrete choice experiments to estimate a preference-based measure of outcome-an application to social care for older people. Journal of Health Economics. 2006; 25: 927-44.
87. Hakim Z, Pathak DS. Modelling the EuroQol Data: A comparison of discrete choce conjoint and conditional preference modelling. Health Economics. 1999; 8: 103-16.
88. Thurstone L. A Law of Comparative Judgement. Psychological Review-New York. 1927; 34: 273-86.
89. Craig B, Busschbach J, Salomon J. Keep it simple: Ranking health states yields values similar to cardinal measurement approaches. Journal of Clinical Epidemiology. 2009; 62: 296-305.
90. Salomon J. Reconsidering the use of rankings in the valuation of health states: a model for estimating cardinal values from ordinal data. Population Health Metrics. 2003; 1.
91. Torrance G, Feeny D, Furlong W. Visual Analog Scales: Do They Have a Role in the Measurement of Preferences for Health States? Medical Decision Making. 2001; 21: 329-34.
92. Parkin D, Devlin N. Is there a case for using visual analogue scale valuations in cost-utility analysis? Heath Economics. 2006; 15: 653-64.
93. Brazier J, McCabe C. 'Is there a case for using visual analogue scale valuations in CUA' by Parkin and Dvelin, A Response: 'yes there is a case, but what does it add to ordinal data? Health Economics. 2007; 16: 645-47.
94. Rasanen P, *et al.* Use of the quality-adjusted life years for the estimation of effectiveness of health care: A systematic literature review. International Journal of Technology Assessment in Health Care. 2006; 22: 235-41.
95. Stiggelbout A, Vogel-Voogt E. Health State Utilities: A Framework for Studying teh Gap between the Imgained and the Real. Value in Health. 2008; 11: 76-87.
96. NSO. The key statistics of Thailand 2007. Bangkok: National Statistical Office, Thailand, 2007.
97. NSO. Survey of the extent of reading in the Thai general population. Bangkok, Thailand: National Statistical office, 2007b.
98. Fagerlin A, *et al.* Measuring Numeracy without a Math Test:Development of the Subjective Numeracy Scale. Medical Decision Making. 2007; 27: 672-80.
99. Woloshin S, *et al.* Assessing Values for Health: Numeracy Matters. Medical Decision Making. 2001; 21: 380-88.
100. Zikmund-Fisher BJ, *et al.* Validation of the Subjective NumeracyScale: Effects of Low Numeracyon Comprehension of Risk Communications and Utility Elicitations. Medical Decision Making. 2007; 27: 663-71.
101. Sornpaisarn S. Measurement of vision function and quality of life in patients under pharmacoemulsification with the intraocular lenses of different prices 2000. Bangkok, Thailand: MSc. thesis, Chulalongkorn University.
102. Maharattanviroj W. Quality of life of patients with end stage renal disease on hemodialyis and continuous ambulatory peritoneal dialysis. Epidemiology. Bangkok: MSc. Mahidol University, 1999.
103. Wee H, *et al.* Feasibility and acceptability of TTO and SG and factors influencing their acceptance for health valuation among Singaporeans. International Society for Pharmacoeconomics and Outcome Research. Shanghai, China, 2006.

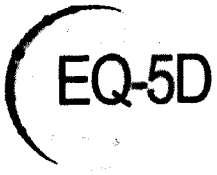
104. Wee H, *et al.* The impact of talking about death on health state valuation: A study among Chinese and Indian Singaporeans International Society for Pharmacoeconomics and Outcome Research. Shanghai, China, 2006.
105. Dolan P, Roberts J. To what extent can we explain time trade-off values from other informatin about respondents? Social Science&Medicine. 2002; 54: 919-29.
106. Dolan P, Roberts J. To what extent can we explain time trade-off values from other information about respondents? Social Science and Medicine. 2002; 54: 919-29.
107. Williams A. The Measurement and Valuation of Health: A Chronicle. York: Centre for Health Economics, The University of York, 1995.
108. Dolan P, *et al.* The Time Trade-Off Method: Results From a General Population Study. Health Economics. 1996; 5: 141-54.
109. Gudex C, *et al.* Valuing health states, Interviews with the general public. European Journal of Public Health. 1997; 7: 441-48.
110. Gudex C, Dolan P. Valuing Health States: The Effect of Duration. : Centre for Health Economics, The University of York, 1995.
111. Dolan P, Roberts J. Modelling Valuations for Eq-5d Health States: An Alternative Model Using Differences in Valuations. Medical Care. 2002; 40: 442-46.
112. Murti B, *et al.* Comparison of Finnish-and US-Based VAS Valuations of the EQ-5D. The 14th Plenary Meeting of the EuroQol group. 1997.
113. Johnson J, *et al.* Valuation of EuroQol(EQ-5D) Health States in an Adult US Sample. Pharmacoeconomics. 1998; 13: 421-33.
114. 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.
115. Tsuchiya A, *et al.* Estimating and EQ-5D population value set: the case of Japan. Health Economics. 2002; 11: 341-53.
116. Devlin N, *et al.* Logical inconsistencies in survey respondents' health state valuations - a methodological challenge for estimating social tariffs. Health Economics. 2003; 12: 529-44.
117. Jelsma J, *et al.* How do Zimbabweans value health states? Population Health Metrics. 2003; 1.
118. Greiner W, *et al.* Validating the EQ-5D with time trade-off for the German population. European Journal of Health Economics. 2005; 6: 124-30.
119. Lamers L, *et al.* The Dutch tariff: results and arguments for an effective design for national EQ-5D valuation studies. Health Economics. 2006; 15: 1121-32.
120. 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.
121. 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.
122. 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.
123. Gudex C. Time Trade-Off User Manual: Props and Self-Completion Methods. York: Measurement and Valuation of Health Group, Centre for Health Economics, University of York, 1994.
124. Busschbach J, Hessing D, Charro F. Observations on 100 students filling in th EuroQoL Questionnaire. Quality of life research. 1994; 3: 71-72.
125. Richardson J, *et al.* The Assessment of Qaulity of Life(AQoL)2 Assessment; Derivation of the Scaling Weights Using a Multiplicative Model and Econometric Second Stage Correction. Melbourne: Centre for Health Economics, Monash University, 2004.

126. Stevens K, *et al.* Multi-attribute utility function or statistical inference models: A comparison of health state valuation models using the HUI2 health state classification system. Journal of Health Economics. 2007; 26: 992-1002.
127. NSO. The 2000 Population and Housing Census. Bangkok, Thailand: National Statistical Office, Thailand, 2000.
128. O'Brien B, Drummond M. Statistical Versus Quantitative Significance in the Socioeconomic Evaluation of Medicines. Pharmacoeconomics. 1994; 5: 389-98.
129. Lenert L, Kaplan R. Validity and Interpretation of Preference-based Measures of Health-Related Quality of Life Medical Care. 2000; 38: II-138 - II-50.
130. Gigerenzer G. The Psychology of Good Judgement: Frequency Formats and Simple Algorithms. Medical Decision Making. 1996; 16: 273-80.
131. Gigerenzer G, Tod P. Simple Heuristics That Make Us Smart. USA: Oxford University Press, 1999.
132. Lanscsar E, Louviere J. Deleting 'irrational' responses from discrete choice experiments: a case of investiaging or imposing preferences? Health Economics. 2006; 15: 797-811.
133. Ryan M, Watson V, Entwistle V. Rationalising the "irrational": A think aloud study of discrete choice experiment responses Health Economics. 2009; 18: 321-36.
134. Ezzy D. Qualitative analysis: Practice and Innovation. London: Routledge, 2002.
135. Pliskin J, Shepard D, Weinstein M. Utility Functions for Life Years and Health Status. Operations Research. 1980; 28: 206-24.
136. Sarkadi K. The consistency of the Shapiro-Francia test. Biometrika. 1975; 62: 445-50.
137. Devlin N, *et al.* The Health State Preferences and Logical Inconsistencies of New Zealanders: A Tale of Two Tariffs. The 17th Plenary Meeting of the EuroQol Group. 2000.
138. Ohinmaa A, Sintonen H. Inconsistencies and Modelling of The Finnish EuroQoL(EQ-5D) Preference Values. The 15th Plenary Meeting of the EuroQoL Group. 1998.
139. Lamers L, *et al.* Inconsistencies in TTO and VAS Values for EQ-5D Health States. Medical Decision Making. 2006; 26: 173-81.
140. Dolan P, Kind P. Inconsistency and Health State Valuations. Social Science and Medicine. 1996; 42: 609-15.
141. Badia X, Roset M, Herdman M. Inconsistent responses in three preference-elicitation methods for health states. Social Science and Medicine. 1999; 49: 943-50.
142. Kind P, Dolan P. Inconsistency and health state valuations. In: Kind P, Brooks R, Rabin R, eds., EQ-5D concepts and methods: a developmental history Netherlands: Springer, 2005.
143. Craig B, Kind P. Logical consistency and the valuation of health: An analysis of US survey data. Value in Health. 2001; 4: 50-51.
144. Stalmeier P, Wakker P, Bezembinder T. Preference Reversals: Violations of Unidimensional Procedural Invariance. Journal of Experimental Psychology: Human Perception and Performance. 1997; 23: 1196-205.
145. Butler DJ, Loomes GC. Imprecision as an Account of the Preference Reversal Phenomenon. The American Economic Review. 2007; 97: 277-97.
146. Seidl C. Preference Revesal. Journal of Economic Surveys. 2002; 16: 621-55.
147. Miguel FS, Ryan M, Amaya-Amaya M. 'Irrational' stated preferences: a quantitative and qualitative investigation. Health Economics. 2005; 14: 307-22.
148. Cameron AC, Trivedi PK. Regression analysisi of count data. New York: Cambridge University Press, 1998.
149. Long JS, Freese J. Regression Models for Categorical Dependent Variables Using Stata. 2nd ed. Texas: Stata Press, 2006.

150. Hardin J, Hilbe J. Generalized Linear Models and Extensions. 2nd ed. Texas: Stata Press, 2007.
151. Vuong Q. Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. Econometrica. 1989; 57: 307-33.
152. Green W. Econometric Analysis. 6th edition ed. New Jersey: Prentice Hall, 2008.
153. Kirkwood B, Sterne J. Medical Statistics Oxford, UK: Blackwell, 2003.
154. 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.
155. 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.
156. Wittenberg E, *et al.* The effect of age, race and gender on preferences scores for hypothetical health states. Quality of life research. 2006; 15: 645-53.
157. Hsiao C. Analysis of Panel Data. New York: Cambridge University Press, 1986.
158. Crosby R, Kolotkin R, Williams G. Defining clinically meaningful change in health-related quality of life. Journal of Clinical Epidemiology. 2003; 56: 395-407.
159. Baum C. An Introduction to Modern Econometrics Using Stata. Texas: A Stata Press Publication, 2006.
160. Chulaporn L, *et al.* Cost-effectiveness analysis of inhaled corticosteroid for mild and moderate asthma in Thai health care context Isan Journal of Pharmaceutical Sciences. 2005; 1: 30.
161. Paltiel AD, *et al.* Cost-effectiveness of inhaled corticosteroids in adults with mild-to-moderate asthma: Results from the Asthma Policy Model. Journal of Allergy Clinical Immunology 2001; 108: 39-46.
162. Ruengrong J, Teerawatananon Y, Chaikledkaew U. Cost-utility analysis of Recombinant Human Erythropoietin to treat anemic condition of the Thai patients receiving chemotherapies. Bangkok: Health Impact and Technology Assessment Program, 2007.
163. Xie X, Vondeling H. Cost-utility analysis of intensive blood glucose control with metformin versus usual care in overweight type 2 diabetes mellitus patients in Beijing, PR China. Value in Health. 2008; 11: S23-S32.
164. Shaw J, *et al.* U.S. Valuations of hte EQ-5D Health States: Methods, Sampling, And Preliminary Analyses. The 20th Plenary Meeting of the EuroQol Group. Bled, Slovania, 2003.
165. Bernert S, *et al.* Comparison of Different Valuation Methods for Population Health Status Measured by the EQ-5D in Three European Countries Value in Health. 2009; 12: 750-58.
166. Luo N, *et al.* A Comparison of EQ-5D Index Scores Derived from the US and UK Population-Based Scoring Functions. Medical Decision Making. 2007; 27: 321-26.
167. Grutters J, *et al.* Choosing between measures: comparison of EQ-5D, HUI2 and HUI3 in persons with hearing complaints. Quality of life research. 2007; 16: 1439-49.
168. Altman D, *et al.* Statistics with confidence London: BMJ Books, 2000.
169. Weir J. Quantifying Test-Retest Reliability Using The Intraclass Correlation Coefficient and the SEM. Journal of Strength and Conditioining Research. 2005; 19: 231-40.
170. Fayers P, Machin D. Quality of Life: Assessment, Analysis and Interpretation. 2nd ed. Chichester: John Wiley & Sons, Ltd, 2007.
171. Bland JM, Altman DG. Statistical Methods For Assessing Agreement Between Two Methods of Clinical Assessment. Lancet. 1986; 327: 307-10.
172. Wee H-L, *et al.* Assessing Differences in Utility scores: A Comparison of Four Widely Used Preference-Based Instruments Value in Health. 2007; 10: 256-65.

173. Samsa G, *et al.* Determining Clinicially Important Differences in Health Status Measures: A General Approach with Illustrtation to the Health Utilities Index Mark II. Pharmacoeconomics. 1999; 15: 141-55.
174. IHPP. Research for Development of an Optimal Policy Strategy for Prevention and Control of Cervical Cancer in Thailand. Bangkok, Thailand: International Health Policy Program and Heath Intervention and Technology Assessment Program, 2008.
175. Roungrong J, Teerawattananon Y, Chaikledkaew U. Cost-Utility Analysis of Recombinant Human Erythropoietin in Anemic Cancer Patients Induced by Chemotherapy in Thailand. Journal of Medical Association of Thailand. 2008; 91 (Suppl 2) S119-25.

Appendix 1 The Thai EQ-5D translation certificate



The EuroQol Group

Certified Translation : EQ-5D Thai version

This is to certify that qualified translators under contract to the EuroQol Group translated the EQ-5D from UK English to Thai in 2002. The work was undertaken by the Center on Outcomes, Research and Education (CORE) at Evanston Northwestern Healthcare in the USA. CORE specializes in the translation and cross-cultural adaptation of QoL questionnaires and clinical scales.

The translation followed an established EuroQol Group translation methodology[1], which was developed with the aim of achieving semantic equivalence to the original and to be easily understandable to the people to whom the translated questionnaire is administered. This rigorous methodology requires two forward translations into the target language by native speakers, a reconciled version of the two forward translations and two back-translations of the reconciled version by a native English speaker fluent in the target language. The second reconciliation version was tested on 8 respondents. All translation steps were taken in full cooperation with members of the EuroQol Group's translation review team. The resulting translation was approved by the EuroQol Group Translation Committee in 2002. All translation work was performed by members of the Thai translation team to the best of their abilities as native speakers of Thai (or English in the case of the back-translators), and as translators and researchers experienced in the field of health-related quality of life research. This translation is, to the best of my knowledge, a valid and accurate translation of the corresponding original document.

Name: Rosalind Rabin

Title: Office Manager of the EuroQol Group Business Management and member of the EuroQol Group Translation Review Team

Signature:

Date : 17th May 2005

[1] Herdman M, Fox-Rushby J, Rabin R, Badia X, Selai C. Producing other language versions of the EQ-5D. In: Brooks R, Rabin R, de Charro F (eds). The measurement and valuation of health status using EQ-5D: A European perspective. Kluwer Academic Publishers. 2003.

Dr. Frank de Charro, EuroQol Business Manger PO Box 4443, 3006 AK Rotterdam, the Netherlands Phone +31 10 408 1545 Fax: +31 10 452 5303 E-mail: rabin@frg.eur.nl

Appendix 2 The interview manual in the Thai study

The Interview Manual

This interview manual was developed to be used in the Health Outcome Valuation Survey in Thailand, which is conducted during March-July 2007. The project fieldwork is sponsored by the Burden of Diseases Project; International Health Policy Program; and the Ministry of Public Health, Thailand. The manual includes 4 parts:

Section 1: Overall interview procedure

Section 2: Overview of the procedures constituted in the interview

Section 3: Dialogues for health state Ranking, VAS and TTO interview tasks

Section 4: Quick guidelines of the TTO interview questions

Section 1: Overall interview procedure

Respondents will be asked to imagine themselves to stay in the different hypothetical health states for 10 years. A number of years of life expectancy, living in the hypothetical states, will be sacrificed for living in perfect health. The interview procedures are as follows:

1. The supervisor introduces the aim and objectives of the project and the possible implications of the study to the Thai health care system. If respondents agree to participate in the interview, they will be asked to give their consent in the consent form.
2. The respondent is requested to report their own current health status within the past 24 hours using the Thai EQ-5D questionnaire and rate their health status on a thermometer scale.
3. A health set consisting of eleven health state cards (3 mild, 3 moderate and 3 severe states, plus 2 core states: 11111 and 3333) is shown to the respondent. After the respondent finishes reading each card, she is then requested to rank the cards according to her preference. The card perceived as the best state is ranked at the top whereas the state viewed as the worst state is ranked at the bottom.

4. A 20-cm thermometer scale with the lowest point (score 0) as "the worst health state imaginable" and the highest point (score 100) as "the best health state imaginable" is presented. A "bisection" method is used. The respondent is requested to assign a score to the best and the worst states (according to the rank). Then the state in the middle of the rank is used, followed by the health states remaining in the rank.
5. The next step is the TTO interview in which respondents will be asked to trade-off their 10-year life expectancy from being in the current health state for shorter life (less than 10 years) in full-health. Respondents can start from health states which are either better or worse than dead. To assist the respondent's thinking process, the interviewer is expected to use the provisional dialogue in the manual. Two TTO boards are also provided to help with the thinking process.
6. The interview ends by respondents filling in their personal details in part 3 of the questionnaire.

Section 2 Overview of interview procedures

Interview procedures and approximate durations of each procedure are presented in Table 1.

Table 1: Estimated duration for each step

Procedure	Estimated duration
1. Project introduction, objectives, informed consent in the consent form.	10 minutes
2. Fill in part 1 of the Thai EQ-5D questionnaire regarding respondent's own health. Then rate his/her own health on thermometer scale	10 minutes
3. Ranking 11 health state cards (2 core states, 3 mild, 3 moderate and 3 severe) and VAS	25 minutes
4. TTO questions for all 11 states (life expectancy of 10 years) to be in "full health" for each state.	25 minutes

Section 3: Dialogues for health state Ranking, VAS and TTO questions

The dialogues are classified into three steps according to the order of the tasks (i.e. Ranking, VAS and TTO) given to the respondents.

Step 1: The interviewer presents a Thai EQ-5D questionnaire to a participant. Dialogue for the questionnaire introduction and how to complete the questionnaire is as follows: *"Please read the health descriptions on this page. The health descriptions described in this questionnaire are comprised of 5 dimensions: mobility, self-care, usual activities, pain/discomfort and anxiety/depression. There are 3 degrees of severity for each dimensions including: no problem, some problems and major problems. Could you please tick (/) a box at the end of a description which best describes your health in the last 24 hours"*

Step 2: A thermometer scale is used in this step. The interviewer introduces the scale using the following dialogue.*"Please find this scale in which score 0 – 100 are drawn. Score 100 is for the most preferable health state possible, whereas score 0 is the worst preferable state possible. Which score would you like to assign for your health in the last 24 hours?"* The interviewer is requested to cross the line at the score given by a respondent and write the score on the left-hand side of the line and the word "Me" at the right-hand side.

Step 3: The interviewer starts by explaining the health state ranking procedure. The respondent is presented with 11 hypothetical health states in which the respondent is requested to imagine him/herself living in each state, without any change, for 10 years. *"There are 11 hypothetical health states described in 11 cards. Please imagine yourself being in each of these states for 10 years, without any changes. Then you will die. Please rank these 11 health cards according to your preference. The uppermost of a rank is the most preferable state, whereas the lowermost is the least preferable state. If you think any two states have equal severity, you could rank them at the same level. There is no right or wrong answers for the ranking process. Please use your own preference and you can change a rank order as often as you would like".* The interviewer records the starting and finishing time for this step. Results of the ranking state are recorded using the 2-alphabetical English code presented at the right end of a card.

Step 4: The participant is presented with a thermometer scale. The interviewer starts with the following dialogue. *"Please look at this scale. The scores are ranging from 0-*

100. Score 100 is for the most preferable health state whereas score 0 is the worst imaginable one. Please assign scores for the states you have ranked." The interviewer starts with the uppermost state. *"Which score would you like to assign for this state?"* Next, the lowermost state is asked.*"Then which score for this state?"* Then, the middle state of the rank is asked. *"Which score for this state?"* The procedure finishes by assigning scores for the rest of the states. Starting and finishing times for this step are recorded.

Step 5: The two TTO boards are used in this step. Firstly, the participant is introduced to the full health state (PH card). After the participant finishes reading the card, the interviewer places this card in Slot A in the TTO board for a state better than death. Then a second card is randomly selected before it is presented to the participant. The participant is asked to read the card before it is placed in Slot B. The interview starts with the TTO board for states better than death. The interviewer moves a marker to 10 years. *"There are two health states for you to choose from. Health state 1 in Slot A is living in this health state (point to Slot A) for 10 year,s then you will die. Health state 2 is living in this second health state for 10 years (point to Slot B). Then you will die. Would you prefer to live in health state 1 or health state 2 or there is no difference between these two states?"* If the participant chooses to stay in health state 1, a PH card is replaced with "immediate dead" card; then the interviewer continues the dialog as follows. *"Now health state 1 is to die immediately, health state 2 is to stay in this state (point to state in Slot B) for 10 years without any change, then you will die. Would you prefer to stay in health state 1 or health state 2 or is there no difference between these two states? "*If the participant chooses to stay in health state 1, that means the participant thinks state 2 is "worse than death". In this case, the TTO board for state worse than death is used. The interviewer ticks "/" the "worse than death" box. If the participant chooses to stay in health state 2, state 2 is "better than death", and the interviewer ticks "/" the "better than death" box.

Dialogue for a state better than death. The interview starts with sliding a PH card into Slot A and moves the marker to point at 5 years. The interviewer asks: *"Now there are two health states for you to choose. Health state 1 is to live in this health state for 5 years; then you will die. Health state 2 is to live in health state 2 (point to Slot B) for 10 years. Then you will die. Would you prefer to live in health state 1 or health state 2 or is there no difference between health state 1 and 2?"* If the participant chooses to stay in

health state 1, the interview ticks /" above the number 5 in the recording form. The marker is then moved to 4 years. The interviewer uses the above dialogue by using 4 years rather than 5. If the participant chooses to stay in health state 2, the interviewer ticks "X" above the number 5 in the recording form. The marker is, then, moved to 6 years. The interviewer repeats the above dialogue using 6 years rather than 5. If the participant reports no difference between the two states, the interviewer ticks "=" above the number 5 in the recording form. The interviewer fills "5" in a provisional box and starts a new state.

Score transformation. Preference weight for a state = $\frac{X}{10}$ Where: X = time being in a healthy state (PH)

If a "/" is next to "X", for example, the respondent prefers to live in state 1 for 9 years and in state 2 for 8 years. The marker is moved to the number 8.5. The interviewer asks: *"Now there are two health states for you to choose. Health state 1 is to live in health state 1 for 8.5 years; then you will die. Health state 2 is to live in health state 2 for 10 years. Then you will die. Would you prefer to live in health state 1 or health state 2 or is there no difference between health state 1 and 2?"* If the respondent reports no difference between the two states, the interviewer ticks "=" above the number 8.5, and fills "8.5" in the box provided. If the respondent prefers to stay in state 1, the interviewer ticks "/" above the number 8.5; then fills "8.25" in a box. If the respondent prefers to stay in state 2, the interviewer ticks "X" above the number 8.5; then fills "8.75" in the box.

Dialogue for a state worse than death. The TTO board for a state worse than death is used. The interviewer starts by introducing a PH card which is permanently fixed in the TTO board, and a selected state in Slot A. The "Immediate dead" card is permanently fixed in Slot B. A marker is placed at 5 years. The interviewer asks: *"There are 2 states for you to choose from. Health state 1 is to live in a selected state (point to the state in Slot A) for 5 years; you will stay in this state (point to the PH card) for 5 years, then you will die. Health state 2 is to die immediately. Would you prefer to live in health state 1 or health state 2 or there is no difference between the two states?"* If the respondent chooses to live in health state 1, the interviewer then ticks "/"above the number 5 in the recording form. The marker is moved to 6 years. The interviewer repeats the above dialogue using 6 years rather than 5. If a respondent chooses to live in health state 2, the

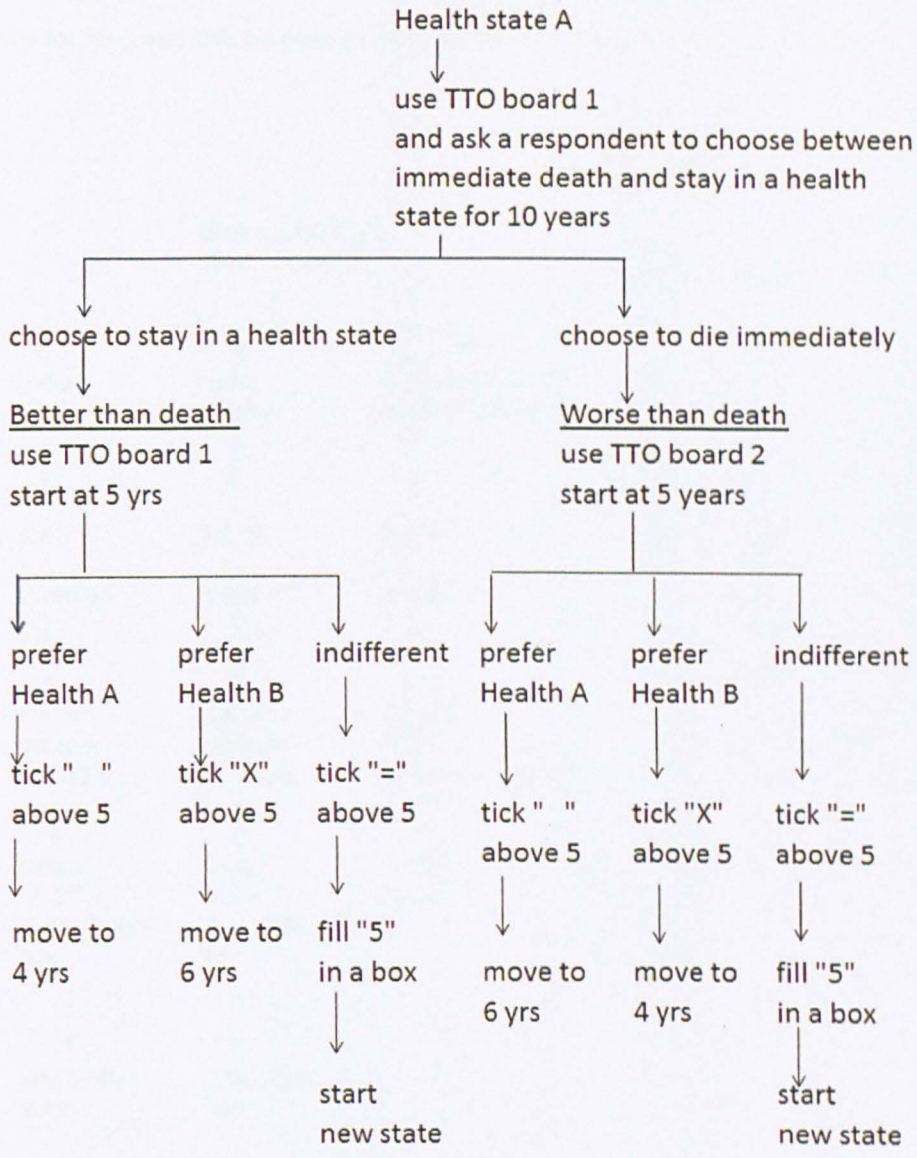
interviewer then ticks "X" above the number 5 in the recording form. The marker is moved to 4 years. The interviewer repeats the above dialogue using 6 years rather than 4. If the participant reports no difference between the two states, the interviewer ticks "=" above the number 5 in the recording form, fills "5" in a provisional box and starts a new state.

Score transformation. Preference weight for a state = $\frac{-X}{(10-X)}$ Where: X = time being in a healthy state (PH)

Dialogue for TTO questions for a very mild state. The respondent may not want to trade his or her life expectancy more than 6 months for a very mild state. In this case, the respondent is requested to trade in months. The interviewer asks: *"There are 2 states for you to choose. Health state 1 is to live in this state (point to box 1) for 9 years and 7 months. Then you will die. Health state 2 is to live in this state (point to box 2) for 10 years. Would you prefer to live in health state 1 or health state 2 or is there no difference between the two states?"*

Section 3: Diagram of TTO question guidelines

Diagram 1: TTO questions guideline

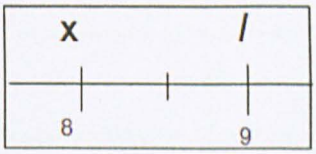


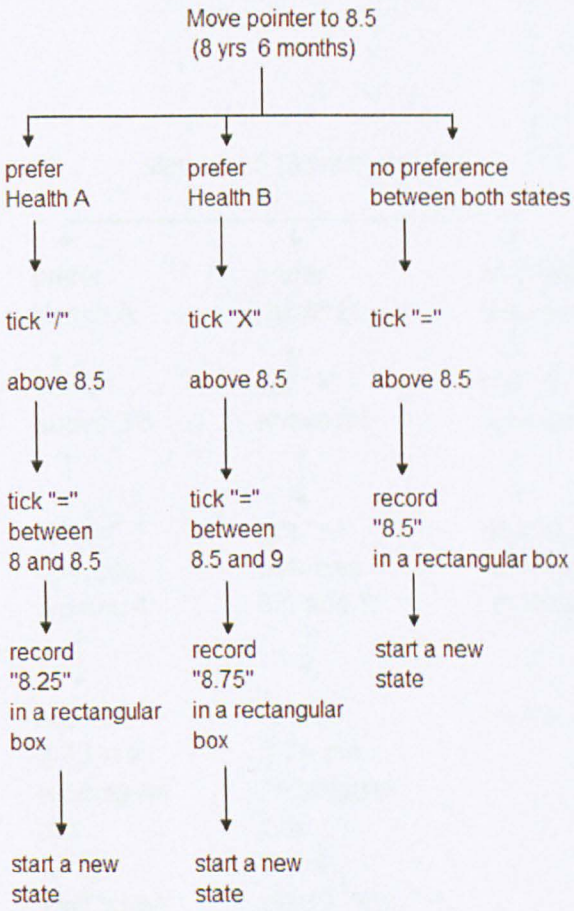
Note: Health A is living in full health for a shorter period

Health B is living in health state A for 10 years

Diagram 2: How to ask TTO questions for a state better than death

If a respondent prefer to live in Health A, tick"/" and if prefer to live in Health B, tick "X". In this diagram, a respondent prefer to live in Health A for 9 years but prefer to live in Health B for 10 years. This diagram shows how to ask TTO questions in this situation.



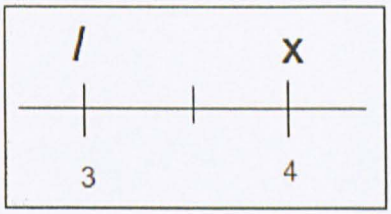


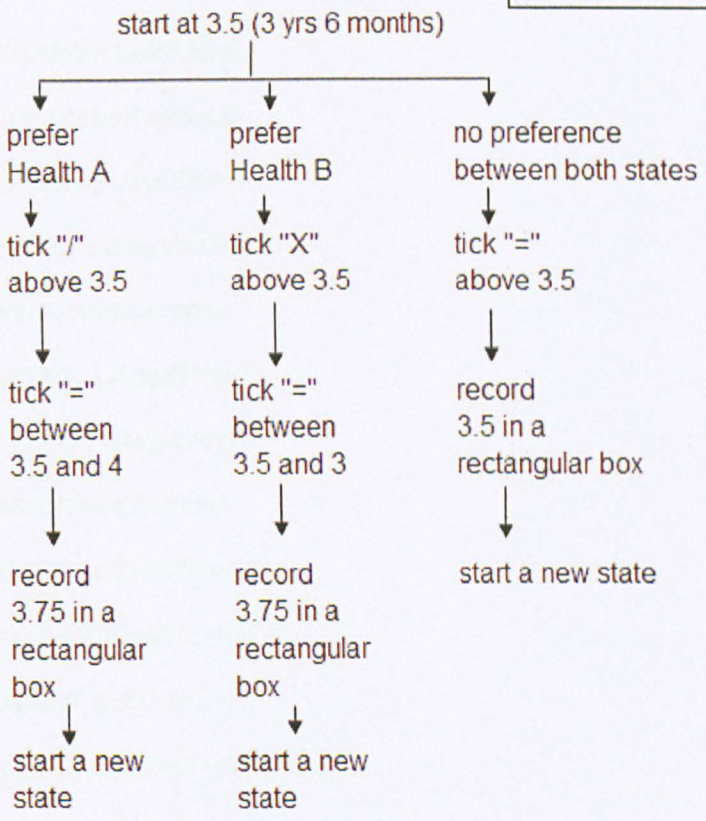
Note: Health A is living in full health for 8 or 9 years

Health B is living in health state A for 10 years

Diagram 3: How to ask TTO questions for a state worse than death

If a respondent prefer to live in Health A, tick "/" and if prefer to live in Health B, tick "X". In this diagram, a respondent prefer to live in Health A for 3 years but prefer to live in Health B for 4 years. This diagram shows how to ask TTO questions in this situation.





Note: Health A is living in health state A for 3 or 4 years, then living in full health for 7 or 6 years.

Health B is immediate dead.

Appendix 3 Example of a do-file to identify logical inconsistently TTO values

Set 1 is used in this example.

```
/* To find out the inconsistencies in the eligible health state pairs*/

gen incXTAE=0 if ttoXT==ttoAE

replace incXTAE=-1 if ttoXT>ttoAE

replace incXTAE=1 if ttoXT<ttoAE

gen incXTAJ=0 if ttoXT==ttoAJ

replace incXTAJ=-1 if ttoXT>ttoAJ

replace incXTAJ=1 if ttoXT<ttoAJ

gen incXTAO=0 if ttoXT==ttoAO

replace incXTAO=-1 if ttoXT>ttoAO

replace incXTAO=1 if ttoXT<ttoAO

gen incXTAT=0 if ttoXT==ttoAT

replace incXTAT=-1 if ttoXT>ttoAT

replace incXTAT=1 if ttoXT<ttoAT

gen incXTAY=0 if ttoXT==ttoAY

replace incXTAY=-1 if ttoXT>ttoAY

replace incXTAY=1 if ttoXT<ttoAY

gen incXTAD=0 if ttoXT==ttoAD

replace incXTAD=-1 if ttoXT>ttoAD

replace incXTAD=1 if ttoXT<ttoAD

gen incXTAK=0 if ttoXT==ttoAK

replace incXTAK=-1 if ttoXT>ttoAK

replace incXTAK=1 if ttoXT<ttoAK

gen incXTAP=0 if ttoXT==ttoAP

replace incXTAP=-1 if ttoXT>ttoAP

replace incXTAP=1 if ttoXT<ttoAP
```

```
gen incXTAV=0 if ttoXT==ttoAV
replace incXTAV=-1 if ttoXT>ttoAV
replace incXTAV=1 if ttoXT<ttoAV
gen incAVAE=0 if ttoAE==ttoAE
replace incAVAE=1 if ttoAV<ttoAE
replace incAVAE=-1 if ttoAV>ttoAE
gen incAVAJ=0 if ttoAV==ttoAJ
replace incAVAJ=1 if ttoAV<ttoAJ
replace incAVAJ=-1 if ttoAV>ttoAJ
gen incAVAO=0 if ttoAV==ttoAO
replace incAVAO=1 if ttoAV<ttoAO
replace incAVAO=-1 if ttoAV>ttoAO
gen incAVAT=0 if ttoAV==ttoAT
replace incAVAT=1 if ttoAV<ttoAT
replace incAVAT=-1 if ttoAV>ttoAT
gen incAVAY=0 if ttoAV==ttoAY
replace incAVAY=1 if ttoAV<ttoAY
replace incAVAY=-1 if ttoAV>ttoAY
gen incAVAK=0 if ttoAV==ttoAK
replace incAVAK=1 if ttoAV<ttoAK
replace incAVAK=-1 if ttoAV>ttoAK
gen incAPAE=0 if ttoAP==ttoAE
replace incAPAE=1 if ttoAP<ttoAE
replace incAPAE=-1 if ttoAP>ttoAE
gen incAPAJ=0 if ttoAP==ttoAJ
replace incAPAJ=1 if ttoAP<ttoAJ
replace incAPAJ=-1 if ttoAP>ttoAJ
```

```
gen incAPAO=0 if ttoAP==ttoAO
replace incAPAO=1 if ttoAP<ttoAO
replace incAPAO=-1 if ttoAP>ttoAO
gen incAPAY=0 if ttoAP==ttoAY
replace incAPAY=1 if ttoAP<ttoAY
replace incAPAY=-1 if ttoAP>ttoAY
gen incAPAK=0 if ttoAP==ttoAK
replace incAPAK=1 if ttoAP<ttoAK
replace incAPAK=-1 if ttoAP>ttoAK
gen incAKAE=0 if ttoAK==ttoAE
replace incAKAE=1 if ttoAK<ttoAE
replace incAKAE=-1 if ttoAK>ttoAE
gen incAKAJ=0 if ttoAK==ttoAJ
replace incAKAJ=1 if ttoAK<ttoAJ
replace incAKAJ=-1 if ttoAK>ttoAJ
gen incAKAO=0 if ttoAK==ttoAO
replace incAKAO=1 if ttoAK<ttoAO
replace incAKAO=-1 if ttoAK>ttoAO
gen incAKAY=0 if ttoAK==ttoAY
replace incAKAY=1 if ttoAK<ttoAY
replace incAKAY=-1 if ttoAK>ttoAY
gen incADAE=0 if ttoAD==ttoAE
replace incADAE=1 if ttoAD<ttoAE
replace incADAE=-1 if ttoAD>ttoAE
gen incAYAJ=0 if ttoAY==ttoAJ
replace incAYAJ=1 if ttoAY<ttoAJ
replace incAYAJ=-1 if ttoAY>ttoAJ
```

```
gen incAOAJ=0 if ttoAO==ttoAJ
replace incAOAJ=1 if ttoAO<ttoAJ
replace incAOAJ=-1 if ttoAO>ttoAJ
replace  incXTAE=. if  incXTAE < 1
 replace  incXTAJ=. if  incXTAJ < 1
 replace  incXTAO=. if  incXTAO < 1
  replace  incXTAT=. if  incXTAT < 1
  replace  incXTAY=. if  incXTAY < 1
  replace  incXTAD=. if  incXTAD < 1
  replace  incXTAK=. if  incXTAK < 1
  replace  incXTAP=. if  incXTAP < 1
  replace  incXTAV=. if  incXTAV < 1
  replace  incAVAE=. if  incAVAE < 1
  replace  incAVAJ=. if  incAVAJ < 1
  replace  incAVAO=. if  incAVAO < 1
  replace  incAVAT=. if  incAVAT < 1
  replace  incAVAY=. if  incAVAY < 1
  replace  incAVAK=. if  incAVAK < 1
  replace  incAPAE=. if  incAPAE < 1
  replace  incAPAJ=. if  incAPAJ < 1
  replace  incAPAO=. if  incAPAO < 1
  replace  incAPAY=. if  incAPAY < 1
  replace  incAPAK=. if  incAPAK < 1
 replace  incAKAE=. if  incAKAE < 1
 replace  incAKAJ=. if  incAKAJ < 1
  replace  incAKAO=. if  incAKAO < 1
  replace  incAKAY=. if  incAKAY < 1
```

```
 replace  incADAE =. if  incADAE < 1

 replace  incAYAJ=. if  incAYAJ < 1

 replace  incAOAJ=. if  incAOAJ < 1

/* nmis=number of inconsistencies in each respondent*/  /*to get nmis use*/

 egen nmis=rmiss2(inc*)

label var nmis "number of inconsistent pairs"

 tab nmis
```

Appendix 4 Parameter estimates using the Fixed effects model

Dolan 1997			Dolan & Roberts 2002		
Variables	Coeff.	SE.	Variables	Coeff.	SE.
mo	0.120	0.012	difmob1	0.308	0.012
sc	0.120	0.007	difmob2	0.455	0.014
ua	0.060	0.007	difsc1	0.126	0.012
pd	0.072	0.012	difsc2	0.272	0.014
ad	0.039	0.012	difua1	0.068	0.012
m2	0.175	0.019	difua2	0.161	0.013
p2	0.082	0.019	difpd1	0.169	0.012
a2	0.042	0.019	difpd2	0.268	0.014
N3	0.142	0.016	difad1	0.101	0.012
cons.	0.195	0.012	difad2	0.171	0.013
			ANY13	-0.075	0.010
			cons.	-0.053	0.010
			Mean score of state 33333	-0.419	
R2 (overall)	0.448			0.447	
RMSE	0.102			0.106	
MAD	0.080			0.085	
Number of states with absolute difference >0.1	27			28	
Numbe of logical inconsistencies in the estimated 243 states	0			15	
Cohen effect size	1.076			1.076	
scores for state					
11112	0.766			0.785	
33333	-0.458			-0.472	

Note: RMSE= Root mean squared errors, MAD=Mean absolute difference

Appendix 5 The Thai preference scores

EQ-5D states	scores	95% CIs		EQ-5D states	scores	95% CIs	
		lower	upper			lower	upper
1 1 1 1 1	1.000	-	-	1 2 2 2 2	0.513	0.430	0.597
1 1 1 1 2	0.766	0.723	0.809	1 2 2 2 3	0.295	0.139	0.453
1 1 1 1 3	0.548	0.432	0.665	1 2 2 3 1	0.269	0.130	0.408
1 1 1 2 1	0.726	0.682	0.769	1 2 2 3 2	0.237	0.079	0.395
1 1 1 2 2	0.693	0.631	0.756	1 2 2 3 3	0.158	-0.048	0.365
1 1 1 2 3	0.475	0.340	0.612	1 2 3 1 1	0.419	0.339	0.501
1 1 1 3 1	0.449	0.331	0.567	1 2 3 1 2	0.387	0.288	0.488
1 1 1 3 2	0.417	0.280	0.554	1 2 3 1 3	0.309	0.161	0.458
1 1 1 3 3	0.338	0.153	0.524	1 2 3 2 1	0.347	0.247	0.448
1 1 2 1 1	0.739	0.704	0.774	1 2 3 2 2	0.315	0.196	0.435
1 1 2 1 2	0.707	0.653	0.761	1 2 3 2 3	0.236	0.069	0.405
1 1 2 1 3	0.489	0.362	0.617	1 2 3 3 1	0.210	0.060	0.360
1 1 2 2 1	0.666	0.612	0.721	1 2 3 3 2	0.178	0.009	0.347
1 1 2 2 2	0.634	0.561	0.708	1 2 3 3 3	0.099	-0.118	0.317
1 1 2 2 3	0.416	0.270	0.564	1 3 1 1 1	0.417	0.348	0.486
1 1 2 3 1	0.390	0.261	0.519	1 3 1 1 2	0.384	0.297	0.473
1 1 2 3 2	0.358	0.210	0.506	1 3 1 1 3	0.306	0.170	0.443
1 1 2 3 3	0.279	0.083	0.476	1 3 1 2 1	0.344	0.256	0.433
1 1 3 1 1	0.540	0.470	0.612	1 3 1 2 2	0.312	0.205	0.420
1 1 3 1 2	0.508	0.419	0.599	1 3 1 2 3	0.234	0.078	0.390
1 1 3 1 3	0.430	0.292	0.569	1 3 1 3 1	0.207	0.069	0.345
1 1 3 2 1	0.468	0.378	0.559	1 3 1 3 2	0.175	0.018	0.332
1 1 3 2 2	0.436	0.327	0.546	1 3 1 3 3	0.096	-0.109	0.302
1 1 3 2 3	0.357	0.200	0.516	1 3 2 1 1	0.357	0.278	0.438
1 1 3 3 1	0.331	0.191	0.471	1 3 2 1 2	0.325	0.227	0.425
1 1 3 3 2	0.299	0.140	0.458	1 3 2 1 3	0.247	0.100	0.395
1 1 3 3 3	0.220	0.013	0.428	1 3 2 2 1	0.285	0.186	0.385
1 2 1 1 1	0.677	0.643	0.711	1 3 2 2 2	0.253	0.135	0.372
1 2 1 1 2	0.645	0.592	0.698	1 3 2 2 3	0.174	0.008	0.342
1 2 1 1 3	0.427	0.301	0.554	1 3 2 3 1	0.148	-0.001	0.297
1 2 1 2 1	0.605	0.551	0.658	1 3 2 3 2	0.116	-0.052	0.284
1 2 1 2 2	0.572	0.500	0.645	1 3 2 3 3	0.037	-0.179	0.254
1 2 1 2 3	0.354	0.209	0.501	1 3 3 1 1	0.298	0.208	0.390
1 2 1 3 1	0.328	0.200	0.456	1 3 3 1 2	0.266	0.157	0.377
1 2 1 3 2	0.296	0.149	0.443	1 3 3 1 3	0.188	0.030	0.347
1 2 1 3 3	0.217	0.022	0.413	1 3 3 2 1	0.226	0.116	0.337
1 2 2 1 1	0.618	0.573	0.663	1 3 3 2 2	0.194	0.065	0.324
1 2 2 1 2	0.586	0.522	0.650	1 3 3 2 3	0.115	-0.062	0.294
1 2 2 1 3	0.368	0.231	0.506	1 3 3 3 1	0.089	-0.071	0.249
1 2 2 2 1	0.546	0.481	0.610	1 3 3 3 2	0.057	-0.122	0.236

EQ-5D states	scores	95% CIs		EQ-5D states	scores	95% CIs	
		lower	upper			lower	upper
1 3 3 3 3	-0.022	-0.249	0.206	2 2 2 2 1	0.425	0.342	0.507
2 1 1 1 1	0.677	0.635	0.719	2 2 2 2 2	0.392	0.291	0.494
2 1 1 1 2	0.645	0.584	0.706	2 2 2 2 3	0.175	0.000	0.350
2 1 1 1 3	0.427	0.293	0.562	2 2 2 3 1	0.148	-0.009	0.305
2 1 1 2 1	0.605	0.543	0.666	2 2 2 3 2	0.116	-0.060	0.292
2 1 1 2 2	0.573	0.492	0.653	2 2 2 3 3	0.037	-0.187	0.262
2 1 1 2 3	0.355	0.201	0.509	2 2 3 1 1	0.299	0.200	0.398
2 1 1 3 1	0.328	0.192	0.464	2 2 3 1 2	0.266	0.149	0.385
2 1 1 3 2	0.296	0.141	0.451	2 2 3 1 3	0.188	0.022	0.355
2 1 1 3 3	0.217	0.014	0.421	2 2 3 2 1	0.226	0.108	0.345
2 1 2 1 1	0.618	0.565	0.671	2 2 3 2 2	0.194	0.057	0.332
2 1 2 1 2	0.586	0.514	0.658	2 2 3 2 3	0.115	-0.070	0.302
2 1 2 1 3	0.368	0.223	0.514	2 2 3 3 1	0.089	-0.079	0.257
2 1 2 2 1	0.546	0.473	0.618	2 2 3 3 2	0.057	-0.130	0.244
2 1 2 2 2	0.513	0.422	0.605	2 2 3 3 3	-0.022	-0.257	0.214
2 1 2 2 3	0.295	0.131	0.461	2 3 1 1 1	0.296	0.209	0.383
2 1 2 3 1	0.269	0.122	0.416	2 3 1 1 2	0.264	0.158	0.370
2 1 2 3 2	0.237	0.071	0.403	2 3 1 1 3	0.185	0.031	0.340
2 1 2 3 3	0.158	-0.056	0.373	2 3 1 2 1	0.223	0.117	0.330
2 1 3 1 1	0.419	0.331	0.509	2 3 1 2 2	0.191	0.066	0.317
2 1 3 1 2	0.387	0.280	0.496	2 3 1 2 3	0.113	-0.061	0.287
2 1 3 1 3	0.309	0.153	0.466	2 3 1 3 1	0.086	-0.070	0.242
2 1 3 2 1	0.347	0.239	0.456	2 3 1 3 2	0.054	-0.121	0.229
2 1 3 2 2	0.315	0.188	0.443	2 3 1 3 3	-0.025	-0.248	0.199
2 1 3 2 3	0.236	0.061	0.413	2 3 2 1 1	0.237	0.139	0.335
2 1 3 3 1	0.210	0.052	0.368	2 3 2 1 2	0.204	0.088	0.322
2 1 3 3 2	0.178	0.001	0.355	2 3 2 1 3	0.126	-0.039	0.292
2 1 3 3 3	0.099	-0.126	0.325	2 3 2 2 1	0.164	0.047	0.282
2 2 1 1 1	0.556	0.504	0.608	2 3 2 2 2	0.132	-0.004	0.269
2 2 1 1 2	0.524	0.453	0.595	2 3 2 2 3	0.054	-0.131	0.239
2 2 1 1 3	0.306	0.162	0.451	2 3 2 3 1	0.027	-0.140	0.194
2 2 1 2 1	0.484	0.412	0.555	2 3 2 3 2	-0.005	-0.191	0.181
2 2 1 2 2	0.452	0.361	0.542	2 3 2 3 3	-0.084	-0.318	0.151
2 2 1 2 3	0.234	0.070	0.398	2 3 3 1 1	0.178	0.069	0.287
2 2 1 3 1	0.207	0.061	0.353	2 3 3 1 2	0.145	0.018	0.274
2 2 1 3 2	0.175	0.010	0.340	2 3 3 1 3	0.067	-0.109	0.244
2 2 1 3 3	0.096	-0.117	0.310	2 3 3 2 1	0.105	-0.023	0.234
2 2 2 1 1	0.497	0.434	0.560	2 3 3 2 2	0.073	-0.074	0.221
2 2 2 1 2	0.465	0.383	0.547	2 3 3 2 3	-0.006	-0.201	0.191
2 2 2 1 3	0.247	0.092	0.403	2 3 3 3 1	-0.032	-0.210	0.146

EQ-5D states	scores	95% CIs		EQ-5D states	scores	95% CIs	
		lower	upper			lower	upper
2 3 3 3 2	-0.064	-0.261	0.133	3 2 2 2 2	-0.058	-0.231	0.115
2 3 3 3 3	-0.143	-0.388	0.103	3 2 2 2 3	-0.137	-0.358	0.085
3 1 1 1 1	0.226	0.113	0.340	3 2 2 3 1	-0.163	-0.367	0.040
3 1 1 1 2	0.194	0.062	0.327	3 2 2 3 2	-0.195	-0.418	0.027
3 1 1 1 3	0.116	-0.065	0.297	3 2 2 3 3	-0.274	-0.545	-0.003
3 1 1 2 1	0.154	0.021	0.287	3 2 3 1 1	-0.013	-0.158	0.133
3 1 1 2 2	0.122	-0.030	0.274	3 2 3 1 2	-0.045	-0.209	0.120
3 1 1 2 3	0.043	-0.157	0.244	3 2 3 1 3	-0.124	-0.336	0.090
3 1 1 3 1	0.017	-0.166	0.199	3 2 3 2 1	-0.085	-0.25	0.080
3 1 1 3 2	-0.015	-0.217	0.186	3 2 3 2 2	-0.117	-0.301	0.067
3 1 1 3 3	-0.094	-0.344	0.156	3 2 3 2 3	-0.196	-0.428	0.037
3 1 2 1 1	0.167	0.043	0.292	3 2 3 3 1	-0.222	-0.437	-0.008
3 1 2 1 2	0.135	-0.008	0.279	3 2 3 3 2	-0.254	-0.488	-0.021
3 1 2 1 3	0.057	-0.135	0.249	3 2 3 3 3	-0.333	-0.615	-0.051
3 1 2 2 1	0.095	-0.049	0.239	3 3 1 1 1	-0.015	-0.149	0.118
3 1 2 2 2	0.063	-0.100	0.226	3 3 1 1 2	-0.048	-0.200	0.105
3 1 2 2 3	-0.016	-0.227	0.196	3 3 1 1 3	-0.126	-0.327	0.075
3 1 2 3 1	-0.042	-0.236	0.151	3 3 1 2 1	-0.088	-0.241	0.065
3 1 2 3 2	-0.074	-0.287	0.138	3 3 1 2 2	-0.120	-0.292	0.052
3 1 2 3 3	-0.153	-0.414	0.108	3 3 1 2 3	-0.199	-0.419	0.022
3 1 3 1 1	0.108	-0.027	0.244	3 3 1 3 1	-0.225	-0.428	-0.023
3 1 3 1 2	0.076	-0.078	0.231	3 3 1 3 2	-0.257	-0.479	-0.036
3 1 3 1 3	-0.003	-0.205	0.201	3 3 1 3 3	-0.336	-0.606	-0.066
3 1 3 2 1	0.036	-0.119	0.191	3 3 2 1 1	-0.075	-0.219	0.070
3 1 3 2 2	0.004	-0.170	0.178	3 3 2 1 2	-0.107	-0.270	0.057
3 1 3 2 3	-0.075	-0.297	0.148	3 3 2 1 3	-0.185	-0.397	0.027
3 1 3 3 1	-0.101	-0.306	0.103	3 3 2 2 1	-0.147	-0.311	0.017
3 1 3 3 2	-0.133	-0.357	0.090	3 3 2 2 2	-0.179	-0.362	0.004
3 1 3 3 3	-0.212	-0.484	0.060	3 3 2 2 3	-0.258	-0.489	-0.026
3 2 1 1 1	0.105	-0.018	0.229	3 3 2 3 1	-0.284	-0.498	-0.071
3 2 1 1 2	0.073	-0.069	0.216	3 3 2 3 2	-0.316	-0.549	-0.084
3 2 1 1 3	-0.005	-0.196	0.186	3 3 2 3 3	-0.395	-0.676	-0.114
3 2 1 2 1	0.033	-0.110	0.176	3 3 3 1 1	-0.134	-0.289	0.022
3 2 1 2 2	0.001	-0.161	0.163	3 3 3 1 2	-0.166	-0.340	0.009
3 2 1 2 3	-0.078	-0.288	0.133	3 3 3 1 3	-0.244	-0.467	-0.021
3 2 1 3 1	-0.104	-0.297	0.088	3 3 3 2 1	-0.206	-0.381	-0.031
3 2 1 3 2	-0.136	-0.348	0.075	3 3 3 2 2	-0.238	-0.432	-0.044
3 2 1 3 3	-0.215	-0.475	0.045	3 3 3 2 3	-0.317	-0.559	-0.074
3 2 2 1 1	0.046	-0.088	0.181	3 3 3 3 1	-0.343	-0.568	-0.119
3 2 2 1 2	0.014	-0.139	0.168	3 3 3 3 2	-0.375	-0.619	-0.132
3 2 2 1 3	-0.064	-0.266	0.138	3 3 3 3 3	-0.454	-0.746	-0.162
3 2 2 2 1	-0.026	-0.180	0.128				