HHS Public Access

Author manuscript

Value Health. Author manuscript; available in PMC 2016 March 01.

Published in final edited form as:

Value Health. 2015 March ; 18(2): 217–223. doi:10.1016/j.jval.2014.11.005.

Learning and Satisficing: An Analysis of Sequence Effects in Health Valuation

Benjamin M. Craig, PhD,
Health Outcomes and Behavior, Moffitt Cancer Center and University of South Florida, Tampa, FL, USA

Shannon K. Runge, MA,
Health Outcomes and Behavior, Moffitt Cancer Center and University of South Florida, Tampa, FL, USA

Kim Rand-Hendriksen, PhD,
Department of Health Management and Health Economics, University of Oslo, Oslo, Norway

Juan Manuel Ramos-Goñi, BSc, and
Canary Islands HealthService (SESCS)

Mark Oppe, PhD
Institute for Medical Technology Assessment (iMTA), Erasmus University Rotterdam, Rotterdam, The Netherlands

Abstract

Objective—This study estimates the effect of sequence on response precision and behavior in health valuation studies.

Methods—Time trade-off (TTO) and paired comparison responses from 6 health valuation studies—4 US, 1 Spanish, and 1 Dutch—were examined (22,225 respondents) to test whether task sequence influences response precision (e.g., rounding), response changes and median response times. Each study used a computer-based instrument that randomized task sequence among a national sample of adults, age 18 or older, from the general population.

Results—For both TTO and paired comparisons, median response times decreased with sequence (i.e., learning), but tended to flatten after the first 3 tasks. Although the paired comparison evidence demonstrated that sequence had no effect on response precision, the frequency of rounded TTO responses (to either 1-year or 5-year units) increased with sequence.

Address correspondence to: Benjamin M. Craig, PhD, Moffitt Cancer Center, 12902 Magnolia Drive, MRC-CANCONT, Tampa, FL 33612-9416. Phone: (813) 745-6710; Fax: (813) 745-6525. benjamin.craig@moffitt.org.

Conflicts of interest: There are no conflicts of interest.

Conclusion—Based on these results, randomizing or reducing the number of paired comparison tasks does not appear to influence response precision; however, generalizability, practicality, and precautionary considerations remain. Overall, participants learned to respond efficiently within the first 3 tasks and did not resort to satisficing, but may have rounded their TTO responses.

Keywords

QALY; Time Trade-off; Health Valuation; Preferences; Sequence Effects; Response Precision; Paradata

Introduction

Most economic evaluations summarize effectiveness using preference weights on a quality-adjusted life year (QALY) scale, as recommended by numerous health technology assessment agencies. Such QALY weights may be from societal or patient perspectives and derived using a wealth of preference elicitation tasks (e.g., best-worst scaling). Although valuation research has a well-established history, the use of online computer-based surveys for health valuation offers an array of new capabilities, such as quota-sampling at the task level; paradata on respondent behavior, device, and browser; and other interactive technologies. Compared to interview, postal, or telephone surveys, online computer-based experiments increase control in the randomization of tasks, while reducing cognitive burden and minimizing missing data and other data collection errors and biases.

Although online instruments typically randomize the order of presentation of tasks, response precision and behavior may change with sequence. For example, when a respondent is shown 2 alternatives and asked, “Which do you prefer?” he or she may take longer or change his/her responses on initial pairs while becoming acquainted with the valuation task as compared to later pairs. Furthermore, a respondent’s attention may wane in later pairs leading to satisficing (i.e., expediting selection among alternatives to minimize effort), reducing response precision.(1, 2) This paper examines whether response precision and behavior varies with number of tasks completed (i.e., sequence effect) in health valuation studies for 2 types of valuation tasks, time trade off (TTO) and paired comparisons.

Understanding the relationship between response precision and task sequence guides the number of tasks to be included in a valuation study, informs weights that place a greater emphasis on earlier or later tasks, and justifies the randomization of task sequence. Although studies have attempted to identify respondents who randomize all responses (i.e., shufflers and satisficers),(3) few studies to date have examined the effect of sequence on response precision in health valuation.(4)

Sequence effects have been identified in other forms of discrete choice experiments (DCEs) as a type of ordering effect specifically related to the order in which choice sets are presented (i.e., position-dependent order effects).(5) This type of order effect differs from those related to the order or position of attributes within a choice set.(5–7) Experimental design, such as the layout of questions, number of attributes, and number of tasks, can influence ordering effects and response time.(8–10) A key example in survey research is the primacy effect or the tendency for respondents to choose the first reasonable answer to a

survey question (e.g., first response option in a list of potential answers).(6, 11) This weak form of satisficing leads to non-random response; expedites response with minimum effort; reduces response quality and time; and is commonly cited by experimenters in order to justify randomization and reduction of the number of attributes, scenarios, and tasks.(12)

A wealth of studies have examined order effects in terms of perception and salience,(5, 7, 9, 10, 13–17) although the results have been somewhat inconsistent. For example, some evidence suggests that the order of attributes affects choice,(5, 7) yet other studies did not find this effect.(9, 14, 18) Additionally, the number and complexity of task sets within an experiment may induce order effects through respondent fatigue or boredom.(19) Evaluating the association between participant response behaviors (i.e., response times and changes) and task sequence has the potential to provide valuable insight regarding the influence of study design.

In complement to evidence on response precision, we examine response behaviors (i.e., response times and changes) that may indicate learning and added deliberative effort beyond that which is needed to satisfy the task requirements. Typically, response behavior is examined at the questionnaire level (e.g., the amount of time it takes a respondent to complete all tasks). In addition to evaluating response behavior at the questionnaire level, computerized software offers a unique opportunity to examine response behaviors at the level of individual questions (e.g., the amount of time it takes to complete a single task set or a series of different task sets). A better understanding of response behavior at each of these levels can aid in the interpretation of the empirical association between sequence and response precision and in the improvement of survey design (e.g., cognitive burden).

The present study contributes to an innovative evaluation of client-side paradata. Client-side paradata is the information recorded in web surveys by the respondent's computer (e.g., the number of times and locations of mouse clicks on a computer screen). Unlike server-side paradata, which refers to data management processes, client-side information allows researchers to interpret participant response behaviors in terms of changed responses and response time at the level of individual questions.(20) Evaluating response behavior patterns at such a specific level contributes to our knowledge of how sequence influences preferences. In this secondary analysis of health valuation data, we examine sequence effects, specifically whether response precision and behavior varies with number of tasks completed.

Methods

Preference Elicitation

In a paired comparison, respondents are asked "which do you prefer?" given 2 health episodes, and their choices define the relative value between these episodes. An original time trade-off (TTO) task is more involved, using an adaptive series of paired comparisons based on either time with no health problems or "immediate death." Specifically, each TTO begins with a paired comparison where the respondent must first decide whether the health episode is preferred to "immediate death." If so, an adaptive series of paired comparisons are presented to determine the number of years with no health problems that is equivalent to

the health episode (i.e., better-than-dead indifference statement). If the respondent prefers "immediate death," an alternative series of paired comparisons are completed to identify a worse-than-dead indifferent statement. The original adaptation procedure (25–27) is like a dose-response study in that it increases the duration of problems within an episode until it is equivalent to "immediate death" (e.g., how much poison is needed until it kills you). Thus, the TTO exercise is a matching task that produces an equivalence statement regardless of whether the original paired comparison response is better or worse than death.

Data

To test the effect of sequence on response precision and behavior, we examined paired comparisons and TTO responses from 6 health valuation studies—4 US, 1 Spanish, and 1 Dutch—totaling 259,318 responses from 22,225 respondents who completed 17 to 37 tasks.(2, 21–24) Table 1 summarizes the characteristics of these 6 studies. All studies used a computerized instrument that randomized task sequence using national samples of adults from the general population. For the US-based studies, respondents completed a set of paired comparisons trading improvements in health-related quality of life (HRQoL) for reduced lifespan (i.e., lifespan pairs) before completing a second set that traded alternative HRQoL scenarios of a common duration (i.e., health pairs). For the valuation of the EQ-5D-5L, respondents completed a set of TTOs before completing a set of paired comparisons that traded alternative HRQoL scenarios without a description of duration (i.e., health state pairs). Further description of the protocol of each study is provided online.(2, 21–24)

The TTO task in the Spanish and Dutch studies was an adaptive hierarchy of steps known as the composite TTO (Fig. 1).(24) The composite TTO is derived from both the original and lead-time TTO.(25–27) Each step displayed 2 scenarios and the respondent was asked, "Which is better?" If the respondent did not wish to choose, the respondent may instead state indifference (i.e., the scenarios were "about the same").

In this adaptive process, the task began with the choice between 10 years in full health and 10 years in the health state (i.e., step 1). If the respondent chose the health state scenario or stated indifference, the TTO response was +10 and the task ended. If the respondent chose the full health scenario in step 1, the task continue on to step 2 and displayed 0 years in full health (i.e., immediate death) instead of 10 years in full health.

If the respondent chose the full health scenario in step 2, the task continued to step 3 and displayed 5 years in full health instead of 0 years in full health. If the respondent chose the health state scenario in step 2, the task continued to step 3 and displayed −5 years in full health instead of 0 years in full health. If the respondent stated indifference in step 2, the TTO response was 0 and the task ended. This task continued for up to 9 steps until the respondent expressed indifference between the 2 scenarios (Fig. 1).

Aside from the highest possible response (+10), which required either 1 or 9 steps, each TTO response required a minimum number of steps (i.e., some TTO responses required more steps than others). The lowest possible response (-10) required the most effort (i.e., 9 steps). By construction, about half of any TTO sample should have been in half-year units.

Paired comparison tasks differed by the studies. The US-based paired comparisons began with 3 examples and asked “Which do you prefer?” showing 2 health scenarios with only 2 attributes and their durations. The Spanish and Dutch paired comparisons had no examples. Respondents completed between 7 and 37 paired comparisons. Unlike the TTO task, indifference was not allowed in any of these paired comparison tasks.

Econometrics

For each study, we graphed median response time and the relative risk of a changed response (CR) and a modal response (MR) by sequence. Response time was measured in seconds from the time that the task was first shown until the final response to the task.

A CR is when multiple responses were registered in the paradata for the task (e.g., a respondent may choose the first scenario in a paired comparison as the preferred scenario and then change his response to the second scenario). Changing a response may be related to the difficulty of the choice. For example, if 2 scenarios seemed similar, the probability of changing a response is higher than for a pair with dissimilar scenarios. Nevertheless, we hypothesize that sequence is unrelated to changed responses when pairs are randomly sequenced. Specifically, the relative risk of a CR is the risk of a CR at the location in the sequence divided by the overall risk of a CR. We did, however, investigate the impact of the difficulty of the choice on the CR in a sensitivity analysis.

In order to identify half-year unit responses in the TTO tasks, respondents may be required to complete additional steps to achieve the final response. These steps include overshooting the point of indifference by half a year and backtracking half a year. For example, in order for a respondent to achieve a final TTO of 6.5, he would first be presented with additional scenarios comparing 7 years in a health state (overshooting) and 6 years in a health state (backtracking). Therefore, a TTO CR requires added steps and responses, and a DCE CR implies just added responses. In either case, we hypothesize that sequence is unrelated to the relative risk of CR.

An MR is whether the respondent provided the same response as the mode response for the task. For example, in a choice between mild pain and mild depression, 80% may choose mild pain and this modal response should not vary by sequence. If respondent attention waned, however, the frequency of MRs should diminish until just 50% prefer mild pain. Specifically, the relative risk of an MR is the risk of an MR at the location in the sequence divided by the overall risk of an MR.

For a TTO task the responses are not binary but are integer and half-integer values on a scale ranging from +10 to −10. Therefore, the risk of a TTO MR may be lower than a risk of a DCE MR. In either case, we hypothesize that sequence is unrelated to the relative risk of MR, relative risk of CR, or median response times.

As ancillary measures of TTO response precision, we illustrated the frequency of 5-year and half-year unit TTO responses by sequence. A half-year unit response requires that the respondent complete at least 1 more step than a 1-year unit response. The frequency of half-unit responses represented a trade-off between added effort and greater precision, which

may have varied by sequence. Likewise, a respondent may have stopped the task early (i.e., within 3 steps: +10, 0, +5 or −5) and responded in 5-year units. Rounding to 1-year or to 5-year units was a tacit way to avoid added effort in the TTO task (i.e., satisficing).

All analyses were repeated using varying levels of difficulty (i.e., comparing different levels of severe health states) based on the assumption that decision difficulty increases as respondents compare health scenarios with similar levels of severity. For the TTO tasks, decision difficulty is assumed to peak at the point of respondent indifference between health scenarios. For the DCE tasks, this point occurs when the choice probability of two health scenarios is approximately 50%. Subsequently, we used posterior information about DCE pair probabilities to describe subgroups.

Results

Figure 2 illustrates median response times by sequence. At the beginning of each sequence, response times were reduced substantially. Each line exhibits the same downward sloping shape (i.e., learning) and shows a flattening out. The respondents in the Dutch study had a higher median time than the Spanish and US respondents, regardless of task. The Spanish paired comparisons had a higher response time than the US tasks, possibly due to differences in the number of attributes of each alternative (5 vs. 2). This pattern was also observed in the subgroup analysis, which confirmed that more time was needed when the task was more difficult. We examined, however, whether sequence effects (i.e., median response times, CR, MR and rounding) were similar among tasks with different levels of difficulty (e.g., greater effect seen in easier tasks) and found no differences. Figure 3 illustrates the relative risk of CR by sequence, which decreases over the initial tasks. Figure 4 illustrates the relative risk MR by sequence, and the MR lines appear flat (i.e., relative risks range from 1.1 to 0.9) aside from some wavering.

Unlike the paired comparison responses, TTO responses may be rounded to 1-year or 5-year units, possibly to reduce response effort (Fig. 1). Figure 5 illustrates the frequency of 5-year, 1-year, and half-year unit TTO responses. The results show that over 40% of the Spanish TTO responses were either +10, +5, 0, or −5, regardless of sequence, and that the frequency of these 5-year unit responses increased from 30% to 40% in the Dutch data, representing a reduction in TTO response precision with sequence. Half-year unit responses potentially indicated a small gain in precision and should be half of each sample. The frequencies of half-year unit responses were clearly less than 50% and decreased from 19% to 12% and from 14% to 12% in the Dutch and Spanish samples, respectively. Furthermore, all 86 modal TTO responses in the Dutch and Spanish studies were in 1-year units and most (77% and 87%, respectively) were in 5-year units. It should be noted, however, that even though the proportion of 5-year values and 1-year values is large across respondents, only a small amount of respondents give only 5-year values (2% and 36% in the Dutch study and 6% and 47% in the Spanish study).

Discussion

Using data from 22,225 respondents, we found that sequence had no effect on paired comparison response precision, but may induce greater rounding in TTO responses. The CR lines (Figure 3) decrease over the initial tasks, illustrating those respondents may be learning the task or establishing heuristics that govern their responses of all similar tasks. The first 6 tasks for each US study were lifespan pairs that involved the trade-off between reduced lifespan and HRQoL. This emphasis on a single attribute (i.e., lifespan) may have induced the formation of time-specific heuristics compared to latter pairs that traded 2 losses in HRQoL with common duration. Aside from some wavering, the MR lines (Fig. 4) appear flat (i.e., relative risks range from 1.1 to 0.9), illustrating that response precision was not associated with sequence. The greater variability seen in the TTO MR is likely attributable to its use of non-binary responses.

With TTO, it can be argued that the proportion of half-year responses, theoretically, should be similar to integer-year responses (1- or 5-year units), given the assumption that distribution of preferences could be considered continuous. The results show that the proportion of half-year responses was less than half and decreased with sequence, although at a different rate in the Spanish data than in the Dutch data. Nevertheless, such TTO rounding had no effect on the relative risk of a modal TTO response. This absence of effect may be attributed to the fact that most modal TTO responses are in 5-year units (i.e., rounding increases the likelihood of modal response).

The low and falling proportion of half-year unit TTO responses is striking, but the correct interpretation is not straightforward. The procedure used to identify a half-year unit response requires overshooting the point of indifference and backtracking half a year. For example, a respondent who has a TTO value of 6.5 for a health scenario would be offered 10 years of perfect health, followed by 0, 5, 6, and 7 (overshoot) before stating indifference at 6.5 years. Similarly, a respondent who has a TTO value of 3.5 for a health scenario is offered 10 years of full health, followed by 0, 5, 4, 3 (overshoot) before stating indifference at 3.5 years. The reduction in elicited half-year unit response could represent satisficing, but it could also reflect a reluctance to backtrack, reluctance to overshoot, or a genuine satisfaction with the level of precision offered by sticking to whole years. Which of these explanations is at play could possibly be determined through strategic manipulation of the routing, such as removing the half-year correction, altering the step size to half a year, or giving respondents multiple alternatives (i.e., more than 2 scenarios in a choice set) at each step. Regardless of explanation, the results show that sequence influences the frequency of half-year responses; however, the infrequency of half-year responses suggests that the potential loss of information is limited.

The apparent and increasing frequency of 5-year unit responses is more troubling, as the loss of information is large. The results suggest that the majority of respondents are attracted to these 5-year unit responses, increasing the risk of bias. The extent of these primacy effects and their attraction may be caused by digit preference, satisficing, or cognitive biases, such as anchoring and should be investigated further. Based on the paired comparison results, randomizing or reducing the number of paired comparison tasks does not appear to influence

response precision; however, generalizability, practicality, and precautionary considerations remain.

These considerations are largely related to the design of DCEs: What is the optimal number of tasks that should be included in a survey? Should later tasks be down weighted? Should tasks be randomized? It has been proposed that certain variations in survey design (e.g., increases in the number of tasks, scenarios, and attributes) increase respondent burden and fatigue, thus contributing to ordering effects and response variability.(10, 19, 28) Despite a growing interest in identifying the optimal design for DCEs, the existing literature remains inconclusive and the results of this study failed to identify any benefits from decreasing the number of tasks, down weighting later tasks or randomizing tasks.

Shortening a health preference survey may limit the breadth of the results (e.g., too few attributes) and collect insufficient data to calculate preferences on attributes, particularly if sample size is small.(18, 29) In their widely cited paper, Hensher et al (2001) found 4 and 8 tasks to be insufficient to estimate preferences for attributes which were selected less often, but concluded this could remedied by presenting 24 to 32 tasks without overburdening respondents.(18) Similarly, Carlsson and Martinsson (2008) compared the results of 12 and 24 tasks and found no evidence of sequence effects, but they did report a significantly higher dropout rate for the longer survey.(29) The results of these studies, however, contradict other findings. In a valuation of travel time, Hensher (2006) reported that increasing the number of tasks significantly decreased participant response time and significantly affected the outcome of the study.(28) These results were echoed by Chung et al (2011), who concluded that the ideal number of tasks to be 6 per survey.(30) Although it has been noted that researchers should use careful pretesting to identify the optimal number of tasks to include in a DCE,(30) our results did not find any sequence effects in the DCE, possibly due to their simplicity (2 alternatives with 2 attributes). Still, additional research is needed to rectify these discrepancies.

The primary limitation of this study is that each study included a maximum of 37 tasks because these components were designed to be completed in less than 30 minutes. Evidence, however, from health valuation studies with more than 40 tasks will be explored in future work. In fact, Craig and colleagues are currently in the beginning stages of a study that will allow respondents to complete hundreds of pairs.(31) Our sensitivity analyses on the time it takes to complete a task by difficulty indicated, however, that the time needed to answer a task is shorter for easy tasks than for difficult tasks. This should be taken into account in the design of a study.

Another limitation of the present study is that the trends in relative risk of MR may under-represent losses in TTO precision due to rounding, because most TTO MR are in 5-year units. The use of MR allowed for a uniform summary of trends in TTO and paired comparison response precision, but did not compensate rounding. The proportion of 5-year units and 1-year units is quite large across respondents. Only a few respondents, however, use only 5-year responses or 1-year responses. Future studies may investigate whether rounding is a greater concern in sub-groups of respondents, particularly those with low numeracy. The conclusion from this analysis is that sequence effects are present more in

TTOs than in DCEs, but both show some learning effect. In summary, the results of this study failed to identify any benefits from decreasing the number of DCE tasks, down weighting later DCE tasks, or randomizing DCE tasks.

Acknowledgments

Funding/Support: Funding support for this research was provided by an NCI R01 grant (1R01CA160104), the EuroQol Group (EQ Project 2013130), and Dr. Craig's support account at Moffitt Cancer Center.

The authors thank Carol Templeton and Michelle Owens at Moffitt Cancer Center for their contributions to the research and creation of this paper.

References

1. Barge S, Gehlbach H. Using the theory of satisficing to evaluate the quality of survey data. Res High Educ. 2012; 53:182–200.
2. Craig, B.; Reeve, Bb. Methods report on the promis valuation study: Year 1. Moffitt Cancer Center; 2012.
3. Craig, Bm; Ramachandran, S. Relative risk of a shuffled deck: a generalizable logical consistency criterion for sample selection in health state valuation studies. Health Econ. 2006; 15:835–48. [PubMed: 16532509]
4. Augestad, La; Rand-Hendriksen, K.; Kristiansen, Is, et al. Learning effects in time trade-off based valuation of eq-5d health states. Value Health. 2012; 15:340–45. [PubMed: 22433766]
5. Day B, Bateman Ij, Carson Rt, et al. Ordering effects and choice set awareness in repeat-response stated preference studies. Jenvironeconmanage. 2012; 63:73–91.
6. Malhotra N. Completion time and response order effects in web surveys. Public Opin Q. 2008; 72:914–34.
7. Kjaer T, Bech M, Gyrd-Hansen D, et al. Ordering Effect And Price Sensitivity In Discrete Choice Experiments: Need We Worry? Health Econ. 2006; 15:1217–28. [PubMed: 16786550]
8. Christian, Lm; Parsons, Nl; Dillman, Da. Designing scalar questions for web surveys. Sociol Methods Res. 2009; 37:393–425.
9. Farrar S, Ryan M. Response-Ordering Effects: A methodological issue in conjoint analysis. Health Economics. 1999; 8:75–79. [PubMed: 10082145]
10. Savage, Sj; Waldman, Dm. Learning and fatigue during choice experiments: A comparison of online and mail survey modes. J App Econ. 2008; 23:351–71.
11. Krosnick, Ja. Response strategies for coping with the cognitive demands of attitude measures in surveys. Appl Cogn Psychol. 1991; 5:213–36.
12. Schwarz, N.; Sudman, S.; Schuman, H., et al. Context Effects In Social And Psychological Research. Springer-Verlag; 1992.
13. Blumenschein K, Johannesson M. An experimental test of question framing in health state utility assessment. Health Policy. 1998; 45:187–93. [PubMed: 10338950]
14. Boyle, Kj; Ozdemir, S. Convergent validity of attribute-based, choice questions in stated-preference studies. Environ Resour Econ. 2009; 42:247–64.
15. Howard K, Salkeld G. Does attribute framing in discrete choice experiments influence willingness to Pay? Results from a discrete choice experiment in screening for colorectal cancer. Value Health. 2008 volume:page-range.
16. Kamoen N, Holleman B, Mak P, et al. Agree or disagree? Cognitive processes in answering contrastive survey questions. Discl Process. 2011; 48:355–85.
17. Yan T, Tourangeau R. Fast times and easy questions: The effects of age, experience and question complexity on web survey response times. Appl Cogn Psychol. 2008; 22:51–68.
18. Hensher DA, Stopher PR, Louviere JJ. An exploratory analysis of the effect of numbers of choice sets in designed choice experiments: An airline choice application. J Air Trans Manag. 2001; 7:373–79.

19. De Palma A, Myers Gm, Papageorgiou Yy. Rational choice under an imperfect ability to choose. Am Econ Rev. 1994; 84:419–40.
20. Heerwegh D. Explaining response latencies and changing answers using client-side paradata from a web survey. Social Science Comp Rev. 2003; 21:360–73.
21. Craig, B.; Owens, MA. Methods Report On The Child Health Valuation Study (Chv): Year 1. Moffitt Cancer Center; 2013.
22. Craig, B.; Owens, Ma. Methods Report On The Women's Health Valuation Study (Whv): Year 1. Moffitt Cancer Center; 2013.
23. Craig, B.; Owens, MA. United States Measurement And Valuation Of Health Study (2013 Us Mvh): Methods Report. Moffitt Cancer Center; 2013.
24. Janssen BMF, Oppe M, Versteegh MRN, et al. Introducing The Composite Time Trade-Off: A Test Of Feasibility And Face Validity. Euro J Health Econ. 2013; 14(Suppl):S5–S13.
25. Devlin, Nj; Tsuchiya, A.; Buckingham, K., et al. A uniform time trade off method for states better and worse than dead: feasibility study of the 'Lead Time' approach. Health Econ. 2011; 20:348–61. [PubMed: 21308856]
26. Gudex, C. Report Of The Centre For Health Economics. York, United Kingdom: University Of York; 1994. Time Trade-Off User Manual: Props And Self-Completion Methods.
27. Torrance, Gw; Thomas, Wh; Sackett, Dl. A utility maximization model for evaluation of health care programs. Health Serv Res. 1972; 7:118–33. [PubMed: 5044699]
28. Hensher, Da. Revealing differences in willingness to pay due to the dimensionality of stated choice designs: An Initial assessment. Environ Resour Econ. 2006; 34:7–44.
29. Carlsson F, Martinsson P. How much is too much? Environ Resour Econ. 2008; 40:165–76.
30. Chung C, Boyer T, Han S. How many choice sets and alternatives are optimal? Consistency in choice experiments. Agribusiness. 2011; 27:114–25.
31. Craig, BM.; Schell, MJ.; Brown, PM., et al. Hrqol Values For Cancer Survivors: Enhancing Promis Measures For Cer. H. Lee Moffitt Cancer Center: NiH; 2011.

Step	1	2	3	4	5	6	7	8	9
									+10
								+9.5	
							+9.0		
						+8.0		+8.5	
					+7.0		+7.5		
				+6.0		+6.5			
					+5.5				
			+5.0						
					+4.5				
				+4.0		+3.5			
					+3.0		+2.5		
						+2.0		+1.5	
							+1.0		
								+0.5	
Start	+10	0						-0.5	
							-1.0		
						-2.0		-1.5	
					-3.0		-2.5		
				-4.0		-3.5			
					-4.5				
			-5.0						
					-5.5				
				-6.0		-6.5			
					-7.0		-7.5		
						-8.0		-8.5	
							-9.0		
								-9.5	
									-10

Figure 1. Minimum Number of Steps Involved in Each Composite Time Trade-off Response*

* Numbers in the time trade-off (TTO) represent the value of 10 years in health state on a quality-adjusted life year (QALY) scale based on a statement of indifferences (e.g., 10 years in health state=+5 QALYs).

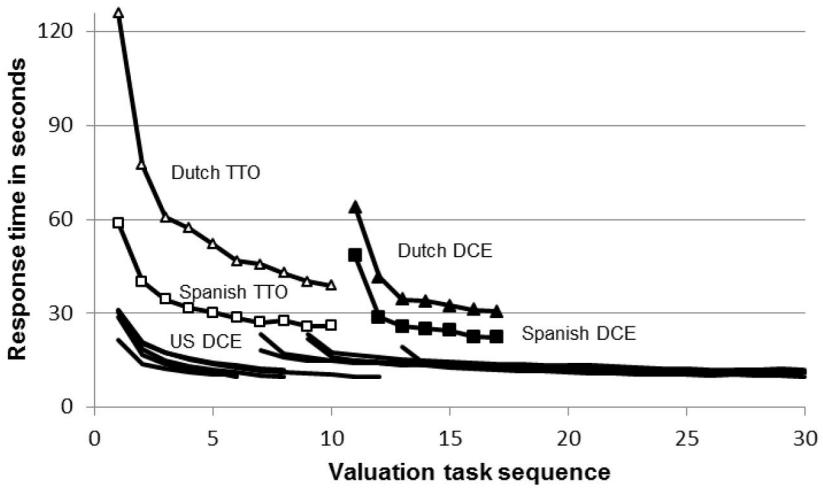

Figure 2. Median Response Time by Sequence

TTO = time trade-off; DCE = Discrete Choice Experiment

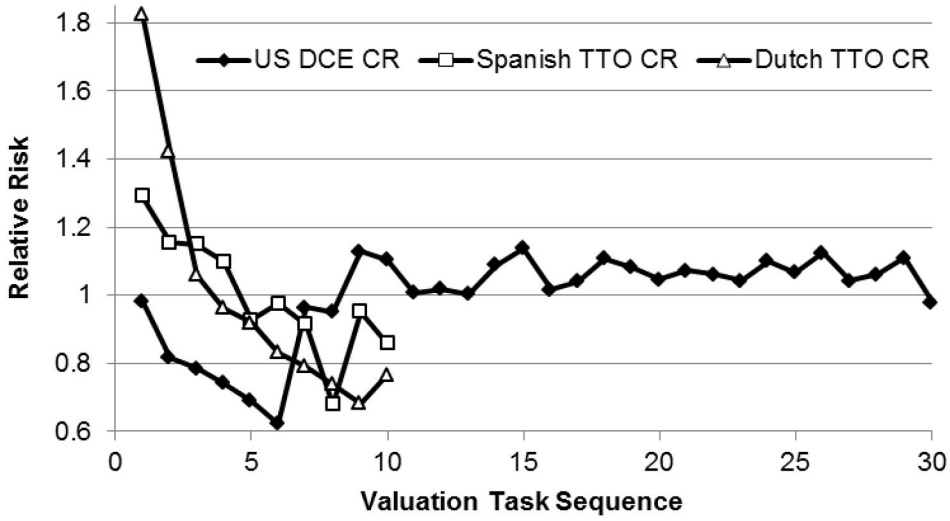

Figure 3. Relative Risk of Changed Responses (CR) by Sequence*

* The Dutch and Spanish studies did not collect CR data on paired comparisons.

TTO = time trade-off; DCE = Discrete Choice Experiment

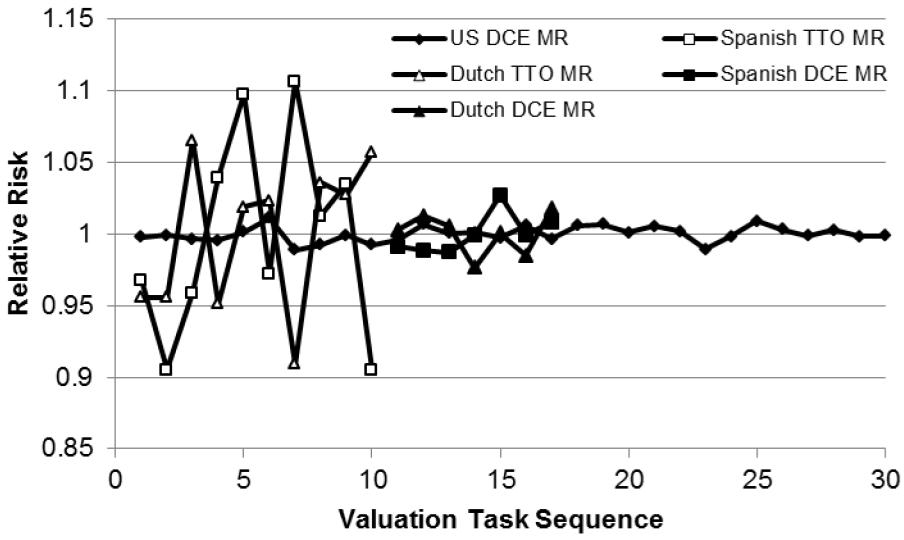

Figure 4. Relative Risk of Modal Response (MR) by Sequence
TTO = time trade-off; DCE = Discrete Choice Experiment

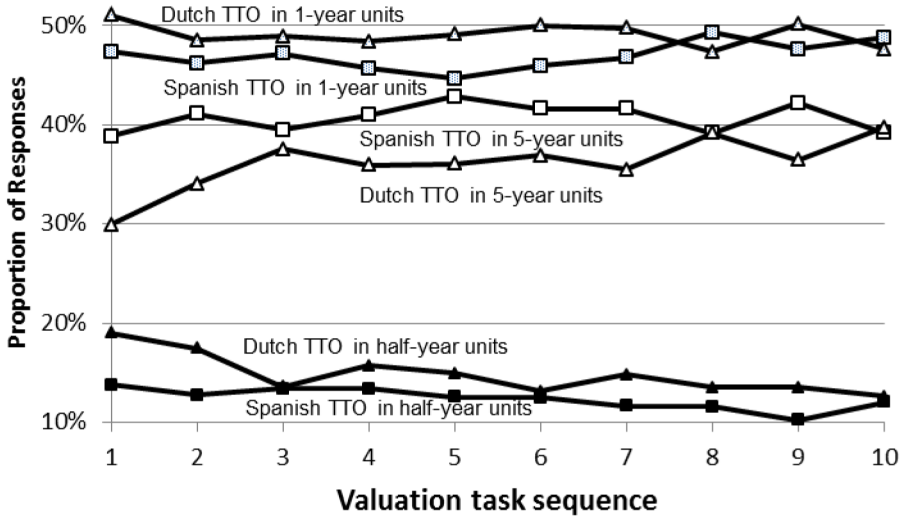

Figure 5. TTO Rounding by Sequence*
TTO = time trade-off

Table 1

Health Valuation Studies*

Study Title	Dates	#	First Set of Tasks	Second Set of Tasks
Patient Reported Outcomes Measurement Information System (PROMIS) Valuation Study - United States (2)	Mar–Jul 2012	7557	6 lifespan pairs	24 health pairs
EQ-5D-5L Valuation Study - Spanish	May–Jul 2012	986	10 time trade-offs	7 health state pairs†
Child Health Valuation Study - US, Wave 1 (21)	Jul–Aug 2012	2008	6 lifespan pairs	31 health pairs
EQ-5D-5L Valuation Study - Dutch (24)	Sep–Oct 2012	1052	10 time trade-offs	7 health state pairs†
Child Health Valuation Study - United States, Wave 2 (21)	Jan–Feb 2013	2147	12 lifespan pairs	18 health pairs
Women's Health Valuation Study - United States (22)	Apr 2013	3397	8 lifespan pairs	22 health pairs
Measurement and Valuation of Health Study - United States (23)	Nov–Dec 2013	5078	8 lifespan pairs	22 health pairs

*Each wave of the US Child Health Valuation Study is shown separately due to changes in the valuation tasks.

†Unlike health and lifespan pairs, health state pairs do not describe duration in the health state.