ARTIGOS
MEASURING PATIENT PREFERENCES THROUGH THE TIME TRADE-OFF METHOD FOR ORTHOPEDIC CONDITIONS ON LARGE SAMPLES

Talitha Yen¹, Clarissa Garcia Rodrigues¹, Aline Chotte de Oliveira² Paulo Rafeal Sanches Calvo³ Richard Mather¹ Jonathan Routh¹ João Ricardo Nickenig Vissoci¹

1- Department of Surgery. Duke University. Durham/NC, USA
2- Faculdade Ingá
3- Faculdade Ingá. Department of Medicine. Maringa/PR, Brazil
4- Department of Surgery, Division of Orthopaedics. Duke University. Durham/NC, USA
5- Assistant Professor. Department of Surgery, Division of Urology. Duke University. Durham/NC, USA
6- Division of Emergency Medicine, Department of Surgery. Duke University Medical Center, Durham/NC, USA.

ABSTRACT

Background: In cost-effectiveness analyses, Quality-Adjusted Life Years (QALY) remains one of the most widely used health effect measure. Among the various methods of estimating utility values, time trade-off (TTO) has traditionally been one of the dominant methods for eliciting utilities, however it has been presenting several practical impediments to provide a high and fast collecting process.

Objective: To test a method of collecting TTO-derived utilities using a platform called Amazon’s Mechanical Turk (MTurk) that provides reliable, fast and inexpensive data.

Methods: A pre-programmed interactive questionnaire was design to simulate a live TTO interview using Qualtrics. To validate the results members of the Research on Research (RoR) Group not aware of the research agreed to answer the same questions on a videoconference live interview. We determined feasibility through assessment quality and cost/benefit relation indicators. In addition, this paper followed the framework for reproducible research reports proposed by our group.

Results: Results: Our results showed that the MTurk population is representative of the US population (based on 2012 census) and there were no differences on the willingness to live when comparing the MTurk sample and the live interview sample, and also no differences of the WTL when comparing the different questionnaire designs developed. Preference results showed differences only for race (between others and African-Americans, and other and white), and overall median values of 0.83 (Q1=0.83;Q3=0.90).

Conclusions: MTurk is a reliable web place to collect large sample using the TTO method, and should be used to collect utility data for CEA.

Keywords: cost-effectiveness, Quality-Adjusted Life Years, time trade-off.
Introduction

Because the effect of medical interventions may be difficult to measure and compare, it is unsurprising that there has been considerable interest and research into developing methods to quantitatively measure the health status of individuals and population. Panel on Cost-Effectiveness in Health and Medicine, a nonfederal panel with expertise in clinical medicine, ethics, and health outcomes measurements, concluded that although cost-utility analyses do not reflect every element of health care decisions, they do provide critically important information to apply to decisions about health care resource allocation. The Panel recommended that CEAs be conducted from a societal perspective in order to allow readers and policymakers to judge the relative magnitudes of health effects. More pertinently, the Panel recommended that researchers incorporate quality-of-life into the denominator of the cost/effectiveness ratio.

The most commonly used method of quantitatively comparing these interventions is cost-utility analysis, which measures the benefits of competing health care interventions using quality-adjusted life years (QALYs). The QALY model provides a straightforward way to combine the two main outcomes of health care interventions: 1) quality of life in a given health state, and 2) duration of life in that health state into a single index measure. Though other alternatives certainly exist, the QALY has been the mainstay of incorporating the valuation of health outcomes in economic evaluation. The QALY is based on a utility value, which assigns a weight between 0 (for death) and 1 (for full health). This weight is the relative value of time spent in different health states.

As is widely known, utility values may be elicited either by direct or indirect methods. Direct methods are considered by some experts to be more accurate, as they incorporate the subject’s judgment of his own health state compared to perfect health and death. In addition, indirect methods typically use community-established health profiles that provide scores on impairments on several health dimension; however, these scores usually offer little information on the significance of the impairment.

Currently, the dominant methods for directly eliciting utility values are the
standard gamble (SG) and the time trade-off (TTO). These preference techniques produce values anchored by full health and being dead. The TTO was originally designed as a simpler alternative to the SG and is used to identify the point of indifference between a fixed length of life in an impaired health state and a shorter life span in perfect health. However, there are practical impediments to collecting these direct measurements of utility values (SG and TTO); for example, using people’s individual opinion leads to significant variance, requiring a large sample size to reach statistical significance. In addition, direct methods are complex to administer and can be quite burdensome to participants and site staff. Thus, although direct methods of utility elicitation have significant benefits, there are equally significant challenges to incorporating them into standard CEA.

Therefore, the aim of this study was to develop and validate a novel method for directly eliciting utility weights, in large scale in an efficient way. We therefore used a novel for decision analysis and cost effectiveness analysis, making use of time trade-off indirect method and the Mechanical Turk platform.

**Methods**

**ETHICS**

This study received ethical approval from the Institutional Review Board at Faculdade Ingá, in Brazil. Informed consent was presented in the first page of the online questionnaire with a description of the survey and its purpose. If the respondent did not agree with the terms of the informed consent he/she was automatically taken to the end of the survey. Data collection was based in an electronic website called Amazon.com’s Mechanical Turk (MTurk) (from January 30 to March 3, 2012).

**DATA COLLECTION MECHANISMS**

MTurk is a web platform that allows researchers to obtain a relatively large number of responses in a relatively short period of time. MTurk distributes tasks requiring human intelligence to a large pool of online workers. Recent research suggests that the platform has a similar validity - both internally and externally - as laboratory and field experiments, and that MTurk produces data of similar reliability as those obtained via traditional methods consistent with standard decision-making biases. MTurk has been previously reported to be significantly less expensive.
and more rapid than more traditional survey instruments and sources. In order to design our survey instrument, we used web-based research survey software called Qualtrics. The resulting survey was then distributed to MTurk workers via the web.

**QUESTIONNAIRE DEVELOPMENT AND DESIGN**

Our objective was to design a TTO questionnaire on MTurk that would reproduce the same result as a live time trade-off interview. Typically, TTO participants are asked to decide how much time spent in a state of perfect health they would be willing to give up in order to escape or prevent the health state in question. Specific values for the amount of time that respondents are willing to “trade-off” are then varied until the respondent is indifferent between the two alternatives. The TTO preference (the indifference point) is thus the length of remaining life in perfect health divided by the length of remaining life with the evaluated health state. For example, a respondent who is indifferent between living with the described health state for 10 more years and living with perfect health for 5 more years has a TTO utility of .514.

Based on this model, our instrument was developed with the following features:

- **a) Health State Description:** We chose knee osteoarthritis (OA) as the health condition of interest because it is one of most prevalent joint diseases in the US (CDC). Because the health state description has been noted as a source of bias in previous TTO studies, we presented respondents with a testimonial video of an actual patient suffering OA followed by a written description: “Imagine you’re 50 years old, and you have knee osteoarthritis, you have daily pain, difficulty climbing stairs, lifting heavy weights, walking long distances, you can't do heavy domestic duties, you experience morning stiffness, you take daily pain killers and you need a cane or a walker to attend social activities.”

- **b) Trade-off questions:** After the health description was presented, we asked respondents to imagine that there was a magical pill capable of restoring his/her knee to normal, but in doing so it would take away some of the remaining 30 years of the respondent’s life. TTO questions were then presented in two different formats: the jumping questions and the slider (Figure 1).
c) Jumping questions: Initially, respondents were presented with two yes/no questions: a) if he or she would take the pill taking away 5 years and b) 25 years. Based on these answers, a series of follow-up questions would be presented to each respondent until there were no more logical options and a final decision was reached.

d) Slider: Each worker would then again be presented with the magical pill option, but instead of repeating the option of years willing to give up the respondent was asked to select the maximum number of years he would be willing to give up using a visual analog scale.

e) Assurance question: Because the TTO method is a challenging cognitive task, it is conceivable that gaining experience with the method may influence the resulting values. Therefore, we gave respondents the opportunity to reevaluate their response by including the assurance sentence: “So based on your previous answers, you could say: I'M WILLING TO LIVE ONLY X YEARS WITHOUT KNEE PAIN COMPARED TO Y YEARS OF LIFE WITH KNEE DEGENERATION? Yes/No”. If the respondent answered yes, they would then move on to the next segment of the interview; if they answered no, they would be taken to the beginning of the questionnaire for the
opportunity to answer the questionnaire again.

f) Attention question: We also included screening questions that gauge attention in order to identify the disinterested respondent, consistent with previously published recommendations\textsuperscript{17}. Specifically, we asked an attention question directly related to the video, as we deemed that source was particularly crucial that the worker understood and watched the video carefully to make a conscious and right trade-off.

g) Demographic questions: We included demographic questions to characterize our sample, specifically: age, sex, race, ethnicity, marital status, education, and income.

h) Comorbidity question: Since the objective of this project was to validate a TTO method for eliciting the utility of knee OA with a virtual environment, we used comorbidity as a variable to assess validity by analyzing the variability of willingness to live according to comorbidity.

**PAYMENT**

In previous research using MTurk, the amount of participation is directly related to financial incentives\textsuperscript{18}. The amount paid for each TTO session varied throughout the data collection phase. Initially we paid 20 cents for the Jumping questions; however we noticed a low rate of adherence with this amount. We therefore decided to increase the payment to 40 cents in the Slider questionnaire.

**TECHNICAL FEATURES**

We a priori determined that the jumping questions questionnaire required approximately 7.5 minutes to be answered (3.5 min of watching the video, 3 minutes answering the jumping questions and 1 minute for the demographic questions) and was made up of 76 items displayed in 12 sequential screens. The slider version was shorter (21 items divided into 5 screens); it took approximately 5.5 minutes to be answered (3.5 min for the video, 1 minute for the slider and 1 min for the demographic questions).

**ELIGIBILITY CRITERIA**

For the MTurk population the inclusion criteria were age greater than 18 years old (MTurk workers are required to be 18 years of age or older), US residency, and acceptance rate of 95% or higher in previous MTurk work assignments. Respondent data was excluded if they submitted incomplete survey instruments, or if their slider
responses were inconsistent with the assurance question. In addition, we excluded any respondents whose time to completion of the questionnaire was less than 3.5 minutes, given that the video alone lasted 3.5 minutes. However, we did include these last two items for validation analysis.

**Traditional Questionnaire**

Sample included members of the RoR group that were not participating directly with this project. Exclusion criteria included non-agreement with the informed consent term (there were no exclusions). Fifteen participants were interviewed. Using a video chat platform the participants were initially asked if they would agree to participate in this research according to the terms of the same informed consent presented to the MTurk population. The type of questionnaire was selected randomly, and the video was presented using the YouTube link. Then the questions were read by interviewer TY and explained if necessary. Note that, on the jumping questions the same order and logic of the online questionnaire was followed, according to the flowchart, and on the slider questionnaire the actual scale was presented on the screen for the respondent to answer.

**Feasibility and Validation Measures**

We determined feasibility through two factors: assessment quality and cost/benefit relation indicators. Assessment quality was analyzed through the attention question and the time that each respondent took to answer the questionnaire. We evaluated the number of respondents who answered the questionnaire in less than 3.5 minutes in addition to the number who missed the attention question (which should have been apparent after watching the video). We assessed the cost/benefit by determining the number of respondents compared to the time and cost for each.

To assess validity, we compared the utility values we obtained between MTurk and the traditional model of measurement through interviews. Validity was also tested through the analysis of the outcome (Years to Live) in relation to socio demographic and comorbidity variables. Good validity was indicated if the results followed the results previously reported with traditional measures.

**Data Analysis**

We initially performed an exploratory data analysis. Descriptive statistics were presented as relative frequencies, median
and interquartile ranges. We compared the data distributions using the Anderson-Darling normality test; due to a nonparametric distribution, the Kruskal-Wallis test was used for multiple group comparisons and the Mann-Whitney test was used for pairwise group comparison. We defined the level of significance as \( p=0.05 \), and all statistical procedures and graphs were performed using R language software.

**REPRODUCIBLE RESEARCH FRAMEWORK**

This paper followed the framework for reproducible research reports. The dataset (in CSV format) and figures are available in our open repository and all data analysis codes are shared through our Github project page. The codes are linked to the data set and are functional. All documents are licensed with Creative Commons Attribution - Non commercial 3.0 License.

**Results**

**SAMPLE DESCRIPTION**

The total amount of respondents in the MTurk platform was 404, most of whom were white (81.00%), women (53.70%), between the ages of 18 and 44 years (72.42%). A plurality of respondents reported having 2 to 4 years of college education (40.61%), with income levels of <$20,000 (30.82%) or $20,000-50,000 (30.71%). Half of respondents were never married (50.00%). Figure 2 demonstrates the comparison among MTurk sample, US Census and our validation sample. We observed that roughly all socio demographic characteristics from the MTurk respondents mirrored US census data. Only respondent age, education and marital status showed differences, with MTurk respondents being younger, better educated, and less more likely to be married than the average US citizen (Figure 2).
FEASIBILITY AND VALIDITY OF THE METHOD

Regarding feasibility, 352 (90.95%) of the original sample completed the instrument within a priori estimated acceptable levels. From those patients, 61 (15.76%) missed the Attention question. In relation to Observing Duration and Attention Loss regarding socio demographic variables, the group’s 18 to 44 years (35 respondents, 12.54%), 2 to 4 years of college (17, 10.83%), male (21, 11.93%), larger income (17, 16.00%), never married (21, 11.00%), asian and other races (4, 16.67% and 4, 18.18%) showed larger relative answers below the minimum time expected. Older respondents (10, 20.00%), asian (3, 12.50%) and African americans (4, 16.67%), with high school or less of education (11, 22.00%), with income from 20 to 50 (30, 18.86%) or less (20, 17.54%), showed higher frequencies of mistakes in the attention question. Gender (12%, 28 Female and 31 Male), marital status (11 to 13%, 23 married, 31 never married and 6 separated or widow) showed similar patterns of sample loss concerning Attention (Figure 3).
As for those who answered the Slider question, we observed a sample loss of 84 respondents (30.50%). From those, the larger frequencies of sample loss due to Inconsistency were 25 with 65 years or older (66.66%), 7 from African American race (41.17%), 15 with high school or less as education (40.50%), 30 with income less than 20,000$ (35.71%), 39 from married (37.86%) and 9 from separated and widow (34.61%) marital status groups, and 49 from female gender (33.79%). A comparison between the sample with and without the exclusion with the Observing Duration, Attention Loss and Inconsistency criteria showed a improvement in the outcome variable (Willingness to Live) behavior, detecting a diminish of outliers frequency and a statistically difference between sample groups (p=0.05) (Figure 4).
Validity testing showed no statistical differences between the evaluation methods (Jumping Questions vs. Slider) \( (p<0.05) \). Median values for the Willingness to Live evaluated by TTO and slider methods were similar to the MTurk sample (Median 24.00; Quartilic Range 20.00 to 28.00; and, Median 25.00; Quartilic Range 16.00 from 27.00, respectively) and in Traditional sample (Median 20.50; Quartilic Range 16.50 to 23.50; and, Median 20.00; Quartilic Range 20.00 to 25.00, respectively). We found no significant differences between the TTO method and the slider; specifically, we found no statistical differences between MTurk or traditional interview respondents (Figure 5).
Knee OA Utility Value Results

The overall median utilities value for knee OA was 0.83 (Interquartile range, 0.83 to 0.90). We found no significant differences in TTO utilities based on age (p=0.80), education (p=0.53), income (p=0.38), marital status (p=0.83) and gender (p=0.86). We did find a statistically significant difference among racial/ethnic groups (p<0.01). Differences in race were observed between white and others, african-americans and others (p<0.05). No statistical differences were observed among subjects who identified themselves as having a comorbidity vs. those who did not self-identify (p=0.90) (Figure 6).
DISCUSSION

To the best of our knowledge this is the first study to validate a TTO instrument using the MTurk platform. Previous study suggested that it would be feasible to conduct quality of life research in patients via the Internet25, and other internet based...
Instruments have been previously published, although not with this same methodology.

The best method to derive utility values and patient preferences is still a source of some debate. These range from direct elicitation methods such as visual analog scales, TTO or SG to indirect methods, which convert QOL instrument responses into utility values. Each method of calculating utilities has its own particular risks and benefits, but direct elicitation methods have been suggested to more closely approximate the “true” utility value than a utility value derived from a health state classification system. It was previously reported that although SG may be the best direct method of health state measurement for decision modeling, TTO provides good reliability and better acceptability when compared to SG, particularly on a computerized (though not web-based) method.

However, some experts have previously considered indirect elicitation methods to be superior to direct elicitation methods for use in cost-effectiveness studies. This preference is in part due to the ease with which they can be collected; one of the principal drawbacks of direct elicitation methods is the time and expense involved in their collection. In this study, we found that MTurk significantly reduced the time and expense involved in directly eliciting utility values. This would seem to imply that the use of MTurk may increase the practical usage and applicability of direct elicitation methods in cost-effectiveness studies.

**Validation**

Importantly, our study was able to validate this methodology by demonstrating that the results of live interviews were similar to responses collected from MTurk in either questionnaire design, both jumping questions and sliders. We also noted our response pattern to have similar results to other articles where there were no differences of Willingness to Live related to gender, age, income, or education. High variability or ‘noise’ is common when collecting preferences using a direct method such as TTO or SG, and like our results, other articles have demonstrated that demographic respondent characteristics such as sex, age or education could not explain TTO individual response patterns.

Other studies, however, show that these characteristics tend to influence the TTO results, but different studies do not present the same TTO weights for the same demographic subgroups. To neutralize the effects of other, non-health-related factors that may influence the TTO, very large and
randomly picked samples for each combination of health problems are required.

**EXPLAINING WHY USE THIS METHOD**

The design and report of this questionnaire comes from the idea that collecting utilities using the TTO method through MTurk provides a series of solution to several limitations of utilities collection in particular the traditional TTO method, such as:

**Comparison with Prior Work**

Fast and low cost and high number of responses. We have concluded that MTurk is a fast and low cost way to collect reliable utility since TTO method requires a high number of responses and interviewers are costly and may delay the process. The time spent on each of the instruments, and the average time taken for the questionnaire to be answered by each worker, even after excluding the ‘inappropriate’ answers. As presented in the results the TTO method using the MTurk platform provided a high number of responses in a short amount of time on the other hand the live interview proved to be time consuming not only during the interview process as well as the recruiting process therefore the low number of live interviewed participants. Several authors have reported the ability of retrieving more than 1000 responses in less than 3 days. However it was also noticeable that the jumping question version took longer to be answered and to gather a higher amount of answers. This shows us a already reported feature of the MTurk population, they are attracted to faster and interesting questionnaires and a higher payment and although representative of the Us population it is important to know this particular sample.

Representativeness. The MTurk sample in this article proved to be representative of the US population and as other articles have reported before, the MTurk population has particular feature, but is more representative of the US population than college samples. Several authors have already highlighted the importance of population preferences, time trade-off method used in the general public to elicit utility and disutility values that can also be
used to support the assessment of QALY outcomes in economic models for healthcare decision making. Individual and social TTO values are different, when TTO values are based on individuals who experience the health state it’s guaranteed that they represent a best informed decision on the specific health state, however general population TTO valuation is more valid for health policies and societal intervention, because they represent the whole population instead of a diagnostic group allowing these values to be comparable among different health states.

**INTERFACE POTENTIAL OF IMMERSION, INTERACTIVITY, AND VARIATION OF QUESTION FRAME, AND SIMULATION.**

Also by using a video with patients testimonials and assurance questions and other measurements to guarantee understandability and quality of response we increase the results reliability as some authors have described before, self-administering questionnaires in general population may provide bias responses. It has been reported that that assessments of affect may not provide a fully adequate description of the effects of states of health and illness on experienced utility itself or even its labeling affects health state values. Also even attention may bias the response therefore to have attentions checks and a audiovisual as well as written presentation of the health state increase the accuracy of the utility value.

**STUDY AND TTO LIMITATIONS**

Our findings should be interpreted in light of this study’s limitations. Notably, MTurk respondents are not necessarily representative of the US population; MTurk workers tend to be younger, better educated (though of lower income), and are more likely to be unmarried and to be Caucasian than the US population as a whole. However, the MTurk population is substantially more representative of the US population than most other convenience-based samples, such as the canonical experimental cohort of undergraduate college students. As previously described, TTO is a well-established method of collecting patient preferences for decision analysis and cost effectiveness analysis purposes, for both practical and theoretical reasons. Similarly, its methodological problems and biases have already been well studied and reported. There are several limitations within the TTO process, including respondents who do not wish to trade any length of life for a quality of life improvement. Though risk-averse behavior...
is not as significant with TTO as SG, this is a potential problem in any TTO study39. Similarly, there is a possible lack of validity of constant proportional trade-offs (CPTO)3. In addition, the TTO method is based on a rigidly rational and logical interpretation of human behavior. Because human psychology may occasionally be irrational or illogical, TTO and other direct elicitation methods may not always hold to the basic principles of utility theory.

**FUTURE RESEARCH AND IMPLICATIONS FOR PRACTICE**

For future research, we have found this tool to be extremely efficient as a method to collect preferences for decision analyses and CEA; use of MTurk could conceivably be applied to build a utility database in a fast and inexpensive manner. More specifically, we are planning to integrate the specific values we estimated within the context of a knee OA Markov Model. In addition, the reproducible research framework in which our study was conducted is specifically designed to allow the use of the same methodology in similar utility collecting projects involving other disease processes and health states.

**ACKNOWLEDGEMENTS**

The authors thank the Research on Research Group (http://researchonresearch.duhs.duke.edu/) for the templates for writing introduction and discussion sections of the manuscript40 as well as templates for Literature matrix, Duke University Health System41
REFERÊNCIAS BIBLIOGRÁFICAS


Torrance GW. Health index and utility models: some thorny issues. Health Serv; Res (8) 12:14.


Creative Commons Attribution - Non commercial 3.0 License [Internet] http://creativecommons.org/licenses/by-nc/3.0/us/. Accessed 8 November 2012.


