VALUE IN HEALTH ▮ (2018) ▮▮▮–▮▮▮

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/jval

Test-Retest Reliability of EQ-5D-5L Valuation Techniques: The Composite Time Trade-Off and Discrete Choice Experiments

Fredrick Dermawan Purba, MSc, MPsi[1,2,], Joke A.M. Hunfeld, PhD[1], Reinier Timman, PhD[1], Aulia Iskandarsyah, PhD[3], Titi Sahidah Fitriana, MPsi[1,4], Sawitri S. Sadarjoen, PhD[3], Jan Passchier, PhD[5], Jan J.V. Busschbach, PhD[1]*

[1]*Department of Psychiatry, Section Medical Psychology and Psychotherapy, Erasmus MC University Medical Center, Rotterdam, The Netherlands;* [2]*Department of Developmental Psychology, Faculty of Psychology, Universitas Padjadjaran, Jatinangor, Indonesia;* [3]*Department of Clinical Psychology, Faculty of Psychology, Universitas Padjadjaran, Jatinangor, Indonesia;* [4]*Faculty of Psychology, YARSI University, Jakarta, Indonesia;* [5]*Department of Clinical, Neuro and Developmental Psychology, VU University, Amsterdam, The Netherlands*

ABSTRACT

Objectives: To explore the test-retest reliability of the composite time trade-off (C-TTO) and discrete choice experiment (DCE) used in the Indonesian five-level EuroQol five-dimensional questionnaire (EQ-5D-5L) valuation study. **Methods:** A representative sample aged 17 years and older was recruited from the Indonesian general population by stratified quota sampling with respect to residence, sex, and age. Trained interviewers conducted computer-assisted face-to-face interviews using the EuroQol valuation technology. Each respondent valued 10 health states using C-TTO and 7 pairs of health states in a DCE exercise. The retest interview was conducted after 2 weeks by the same interviewer. The Wilcoxon matched-pairs signed-rank test, intraclass correlation coefficient, and multilevel regression were applied in comparing the C-TTO test and retest data. For DCE, the analysis of proportions was used. **Results:** A total of 226 respondents with characteristics similar to the Indonesian population completed the retest interview. For C-TTO, 82 (95.3%) of 86 health states had no significant mean value differences between test and retest. The mean value of the second test was statistically significantly higher than that of the first test by 0.042. For DCE, 72.5% of responses were identical. DCE retest showed a different pattern concerning the relative importance of the dimensions, whereas the C-TTO remained the same. **Conclusions:** C-TTO is stable over time, whereas in DCE the relative values of the dimensions shift. The results support the use of the C-TTO, in particular the Indonesian EQ-5D-5L value set, and suggest a critical examination of the reliability of DCE results over time.

Keywords: discrete choice experiment, EQ-5D-5L, Indonesia, test-retest, time trade-off.

Introduction

The five-level EuroQol five-dimensional questionnaire (EQ-5D-5L) is a generic measure of health outcomes. It comprises two parts: a descriptive system and a visual analogue scale (EQ-VAS). The descriptive system consists of five dimensions: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. The EQ-5D-5L has five levels for each dimension: no problems, slight problems, moderate problems, severe problems, and unable/extreme problems [1].

To be used in economic evaluations, EQ-5D descriptive responses should be converted into an index score using a value set representing societal preferences [2,3]. This value set is usually obtained using techniques such as the standard gamble, time trade-off (TTO), and discrete choice experiment (DCE). In 2014, the EuroQol Group developed a valuation protocol for the EQ-5D-5L, together with the EuroQol valuation technology (EQ-VT) computer platform, to facilitate and standardize EQ-5D-5L valuation studies across the world. In this protocol, the composite TTO (C-TTO) and DCE were the chosen elicitation techniques [4].

Several studies [5–11] concluded that the clinical administration of the descriptive part of the EQ-5D-5L was valid and reliable. Three studies [12–14] evaluated the reliability of the valuation techniques used to obtain value sets. This reliability could not be tested using the Cronbach α, or any other "internal test," because no questionnaire items related to the traits of the respondent; instead, the respondent valued health states. Hence, reliability could be tested only with test-retest reliability. The reliability of the values elicitation technique is important, because otherwise it would be difficult to advocate the use of a national value set in health care for budgeting decisions, if the values provided by the respondent were to change over time. Van Agt et al. [12] investigated test-retest reliability of value sets based on the EQ-VAS.

* *Address correspondence to:* Fredrick Dermawan Purba, Department of Psychiatry, Section Medical Psychology and Psychotherapy, Erasmus MC University Medical Center, Wytemaweg 80, Rotterdam 3015 CN, The Netherlands.

E-mail: f.purba@erasmusmc.nl, fredrick.purba@unpad.ac.id.

1098-3015$36.00 – see front matter

https://doi.org/10.1016/j.jval.2018.02.003

Badia et al. [13] used the EQ-VAS and TTO, whereas Robinson [14] used TTO and person trade-off. All of them used the old three-level version of the EQ-5D, and in those days the valuation protocol was not fully standardized. The studies by Van Agt et al. and Badia et al. applied generalizability analysis and found that the variability of the values was mostly due to differences between individuals and differences between health states, whereas the variability attributable to the timing of the tests was almost 0. This very low value is difficult to interpret, because one would expect at least some unreliability, given the difficulty of the task. The low value was probably the result of the high variance between the states on the scale, and the variance between individual respondents, which would diminish the variance between the test and the retest exercises. Multilevel analysis would be more appropriate, because this could account for dependencies between health state values provided by one respondent and the dependencies of the values of the test and retest exercises, because they were also given by the same respondent. Moreover, we could include an interviewer level, because interviewers have an effect on values [15]. Robinson's study confirmed the stability of TTO on the basis of intraclass correlation coefficient (ICC) analysis, which would overcome some of the problems just mentioned, but not all. Robinson also concluded that the use of the TTO for the valuation of the three-level EQ-5D was highly reliable.

There is no published test-retest study for the new EQ-VT with respect to the new EQ-5D-5L using C-TTO and DCE. Such a study would be timely, because the new EQ-VT protocol has already been implemented in numerous valuation studies across the globe [16–23]. Therefore, the objective of our study was to measure the test-retest reliability of C-TTO and DCE used in the Indonesian EQ-5D-5L valuation study.

Methods

This study was part of a larger study that focused on the Indonesian national valuation of the EQ-5D-5L, using a face-to-face setting. See the study report of this valuation study also for details of the current test-retest study [23]. The study was approved by the Health Research Ethics Committee, Faculty of Medicine, Padjadjaran University, Indonesia.

Sampling and Data Collection

A quota method stratified with respect to residence (urban/rural), sex, and age (17–30, 31–50, and >50 years) resulted in 12 quota groups. The predefined quotas were based on data from the Indonesian Bureau of Statistics [24]. Fourteen interviewers conducted interviews in six cities and their surroundings, located in different parts of the country: Jakarta, Bandung, Jogjakarta, Surabaya, Medan, and Makassar.

After the first interview, the interviewer judged the respondent's problems concerning the interview, on the basis of completion time (e.g., more than 2 hours), verbally expressed comment(s) indicating frustration, or frequent interruptions indicating tiredness. When these signs were absent, the interviewer asked for the respondent's consent to be interviewed again (retest). The interval between the first test and the retest ranged from 10 days to 2 months. The retest interview was held by the same interviewer.

Valuation Interview Protocol

The valuation protocol (EQ-VT) consists of six components [4,25]:

1. A general welcome and informed consent.
2. Completion of the descriptive system page, the EQ-VAS, and background questions.
3. The C-TTO (see later) tasks. Three debriefing questions regarding the difficulties of the C-TTO tasks were added at the end of this section.
4. A DCE (see later) task followed by three debriefing questions regarding this task.
5. A roundup.
6. Country-specific questionnaires: paper-and-pencil version of the World Health Organization Quality of Life instrument abbreviated version (WHOQOL-BREF) and the Family Resilience Assessment Scale.

All the sections, except for the country-specific questionnaires, were administered via computer-assisted face-to-face interviews using the EQ-VT platform.

Composite time trade-off

The C-TTO exercise applied conventional TTO to obtain better-than-dead (BTD) values, and lead-time TTO to obtain worse-than-dead (WTD) values (see Appendix Figure A1 in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2018.02.003). Details of the C-TTO approach can be found in the studies of Purba et al. [23] and Oppe et al. [25].

The EuroQol valuation protocol included 86 EQ-5D-5L health states to be valued using C-TTO. The 86 health states were distributed into 10 blocks with a similar level of severity. Each block contained eight unique EQ-5D-5L health states and one very mild state (i.e., only one dimension at level 2 and all others at level 1, e.g., 11112) plus the most severe/"pits" state (55555) [4]. Respondents were randomly assigned to 1 of the 10 C-TTO blocks. Each state in the block was presented in random order to respondents.

Discrete choice experiment

DCE tasks were conducted by presenting two health states and asking the respondent to select the preferred state (see Appendix Figure A2 in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2018.02.003). The DCE design consisted of 196 pairs of EQ-5D-5L health states distributed over 28 blocks, each consisting of 7 pairs with a similar severity. The seven paired comparisons were presented in random order using the EQ-VT framework; the right-left order of the two health states presented was also randomized [4]. Each respondent was given one block of DCE tasks to complete.

Statistical Analysis

We first described the current sample's characteristics in comparison with the Indonesian general population using percentages for discrete variables and data from the Indonesian Bureau of Statistics [24].

The test-retest exercises for C-TTO and DCE were analyzed from two perspectives, the respondent perspective and the health state perspective, and also at an aggregate level, using pooled data.

For the C-TTO respondent perspective analysis, we investigated any significant change in the mean of 10 health state values given by each respondent at test and retest applying the Wilcoxon matched-pairs signed-rank test. For the health state perspective analysis, three calculations were made. First, we investigated any significant change in the mean of 86 health states valued applying the Wilcoxon matched-pairs signed-rank test. Second, we evaluated the ICC (two-way random effects, absolute agreement) of each health state. We applied the following reliability guideline for the strength of the ICC values: less than 0.40 = poor, 0.40 to 0.59 = fair, 0.60 to 0.74 = good, and more than 0.74 = excellent [26]. Finally, we investigated the proportions of respondents per health state who were consistent in