In search of a common currency: A comparison of seven EQ-5D-5L value sets

Keywords: EQ-5D-5L, preference weighting, Western preference pattern

1. Introduction

The EQ-5D is the most widely used generic preference-based measure of health. A comprehensive review showed the instrument had been applied in 63% of 1,682 studies [1]. Another review found that 77% of all cost-utility analyses published in 2010 were based on the EQ-5D [2]. Given the increasing influence of such analyses in resource allocation decisions, it is timely to inquire into the value sets (sometimes referred to as 'tariffs') that form the basis for calculating QALYs (quality-adjusted life years). When a value set is not available for a country where a cost-utility analysis is undertaken, a value set from another country would normally be chosen. However, the more the value sets in other countries differ, the more sensitive are the estimated QALY-gains, and the subsequent cost-effectiveness ratio, to the choice of value set.

The EQ-5D descriptive system consists of five dimensions: mobility (MO); self-care (SC); usual activities (UA); pain/discomfort (PD); anxiety/depression (AD). The original EQ-5D-3L is increasingly being supplanted by the EQ-5D-5L with its five severity levels. As of January 2017, value sets for the 5-level version have been developed for seven countries: Canada, England, the Netherlands, Spain, Uruguay, Japan and Korea [3-9]. Given the differences in culture, language, modelling and data collection, some systematic variation across the value sets should be expected, despite a common protocol.

The first aim of this paper is to highlight some methodological (dis)similarities before we compare the index values in each country. The second aim is to identify some characteristics in preference patterns. We distinguish between: i) the relative importance of the five dimensions; ii) the relative utility decrement, or distances between each of the five levels, and; iii) the 'scale-length' differences that reflect the quality vs quantity trade-offs. The third aim is to develop a simplified model that is based on similarities in preference patterns that are revealed by a comparison of the value sets of the four Western countries: Canada, England, the Netherlands and Spain. Lastly, we compare our model with these four value sets, using a large international survey that includes six countries and seven diagnostic groups (N = 7,933).

2. Descriptive comparisons

All countries used a digital aid called the EuroQol valuation technology (EQ-VT), which is the EuroQol standard protocol for EQ-5D-5L valuation studies and a word for word transcribed interview protocol [10]. In each country, around 1,000 respondents from representative population samples expressed their preferences through the composite time-trade-off (cTTO) method and discrete choice experiments (DCE). The EuroQol introduced this composite approach (cTTO), which combines the use of conventional TTO for states better than dead, and the lead-time TTO for states worse than dead to derive values less than zero [11]. The econometric modelling differed. See Table A1 in the Appendix for details on the methodological differences across value sets in the seven countries.

Comparing the published models is difficult when scoring algorithms are non-additive. Some models include a fixed decrement associated with any move from perfect health (all dimensions at level 1; 11111) referred to as *N1* in the modelling to generate the UK 3L value set [12]. Additionally, for some countries there are further subtractions depending on how many dimensions are at levels 4 and 5, or the use of further rescaling. Hence, to facilitate comparison of value sets, we compute and report the index values assigned to all the 20 health states that involve partial decrements along one dimension only, i.e. four dimensions are symptom free (level 1), while the remaining dimension has level 2, 3, 4, or 5. Where alternative value sets based on different models are available, we use the model recommended by the authors.

Table I reports index values for each of these 20 partial health state combinations, including: i) the range between the highest and the lowest value for the same health state across the seven countries and across the four Western countries, referred to as the CENS countries (Canada, England, the Netherlands, Spain); ii) median index values for the CENS countries, and; iii) the index values from our suggested 'amalgam model' WePP (Western Preference Pattern), to be explained below.

Table I: Index values for each country

Generally, Uruguay has the highest values, and Korea the lowest. Japan tends to have low values for the non-severe levels 2 and 3, and lies close to the Korean values for these levels. The differences across the three European and the Canadian value sets are much smaller (compare the total range with the CENS-range in Table I). The Canadian index values generally lie within the European ranges, except for the SC dimension where Canadian values lie below the European values. The CENS-ranges are generally small, except for levels 4 and 5 of the PD and AD dimensions, where the Dutch values are much lower than those for Canada, England and Spain.

Figure 1 provides an illustration of Table I. For Canada, England, the Netherlands and Spain, three striking similarities in preference patterns emerge on: i) the relative importance of the five dimensions; ii) the relative utility decrements along the five levels, and; iii) the 'scale-length' that reflects the quality vs quantity trade-off.

Figure 1a-b: Pattern of dimensions and levels

3. Three characteristics of preferences

3.1. The relative importance of the dimensions

An examination of the relative importance of each dimension shows that similar patterns exist for the CENS countries, where the last two dimensions (PD and AD) generally have the highest relative importance, i.e. lowest index values in Table I. Japan and Korea are different, in that the first dimension (MO) has the highest relative importance.

In addition to measuring the dimension importance by cardinal weights, Appendix Table A2 provides the ordinal ranking of the decrements at each level, for each country. These importance rankings differ only slightly depending on which level is considered. The same general pattern is confirmed in the three European countries and to some extent Canada: PD and AD are ranked highest, representing highest relative importance, and have similar ranking scores; MO and SC are ranked next, also with similar ranking scores, and; UA is generally ranked lowest. Canada differs slightly, in that SC has a higher ranking than MO and UA.

The relative importance weights (reported in Appendix Figure A1) are obtained by identifying which of the five dimensions involves the largest utility decrement at a given level, which is then assigned a relative importance of 1.00. The corresponding decrements for the other dimensions are then divided by this largest decrement at that level, and assigned a relative importance weight (< 1), as illustrated by the length of the coloured lines. Interestingly, the magnitude of the relative differences in the weighting of dimensions is much smaller for Japan and Korea, than in the CENS countries.

The CENS value sets are broadly similar to the English TTO data [4], which suggest that: i) the MO, SC and UA dimensions have quite similar weights; ii) the PD and AD dimensions also have similar weights, and; iii) the sum of the first three's weights is about the same as the sum of the last two's weights. The DCE data in England indicated similar patterns. Based on these observations, Table II compares the aggregate of the three 'functioning dimensions' (MO, SC, UA) with the two 'symptom dimensions' (PD, AD) at each level. Again, a similar pattern is confirmed for the CENS countries: the aggregate importance of the first three dimensions is similar to the aggregate importance of the last two dimensions. As for the other three countries (Uruguay, Japan, Korea) the 'functioning dimensions' have consistently much higher relative importance.

Table II: The relative importance of functioning vs symptoms

3.2. The relative utility decrements between the levels

There is no theoretical reason why a move from one level to the next one down involves the same marginal disutility. This is simply because the EQ-5D is a descriptive system, whereby the levels under each dimension are *described* as opposed to having a numerical or visual scale with identical intervals or space in between. Hence, the utility decrement from one level to the next would reflect respondents' interpretation of the severity differences associated with the words used at each particular level. For example, to the ears of most (English speaking) people the distance from 'moderate' to 'severe' is larger than the distance from 'severe' to 'extreme'. However, given cultural differences and linguistic nuances in translations of the descriptive system, the same relative distances between the five levels should not be expected across countries. Still, Figure 1 illustrates some striking similarities across the three European countries, each with their own language.

Table III compares the values assigned to the different levels of the scale, when each dimension is described at the same level. The drop from 11111 to 22222 includes the constant term, N1, which is part of the value sets in all countries except Canada and England. Generally, the decrements from level 1 to level 2 are larger than from level 2 to level 3. The largest falls occur from level 3 to 4.

The three European countries appear to be quite similar. Canada follows the same pattern, except for the drop from 11111 to 22222 being identical to that from 22222 to 33333. The latter drop for the European countries is much smaller. Interestingly, in all four CENS countries, half the scale length is located between levels 3 and 4. The differences between levels 4 and 5 are fairly small. This similar pattern of scale length distribution, observed in the CENS countries, is not observed in the other three countries' value sets.

Table III: The proportion of the scale length occurring between levels

3.3. Scale length differences

Beyond the (dis)similar pattern of relative utility decrements, the scale lengths differ; with 55555 having its lowest value -0.446 in the Netherlands and its highest -0.025 in Japan. For the CENS countries, the median value at 55555 is -0.25. To facilitate comparison across different scale lengths , Table III includes the proportions of the total scale length across the four intervals. Note that the major part of the differences in the length of scales occurs in the bottom half of the scale (between levels 33333 and 55555).

3.4. An underlying preference pattern

Based on the above, some characteristics of the seven value sets are extracted and compared in Table IV. Some striking similarities can be observed across the value sets in Canada, England, the Netherlands and Spain.

Table IV: Key characteristics of the value sets

To reiterate the similarities in the CENS countries' value sets: the aggregate weight of the two 'symptom dimensions' (PD and AD) is about the same as the three 'functioning dimensions' (MO, SC, UA). PD and AD are quite similar in importance. MO and SC are also quite similar, and UA has generally the lowest importance. Decrements from level 1 to level 2 are generally larger than from level 2 to level 3. The drop from level 3 to 4 is large, particularly so for dimensions PD and AD. The differences in index values are fairly small in the upper part of the descriptive system (levels 2 and 3) but larger when health problems become more severe (levels 4 and 5).

4: Towards a common Western currency

4.1. Western Preference Pattern

The similarities observed in the CENS values sets reveal a preference pattern that can provide a basis for developing a common currency. More specifically, the WePP (Western Preference Pattern) model is derived based on the following observations:

The ordinal ranking of weights across the five dimensions

From Table I, Figure 1a and Figure A2:

i) PD = AD
ii) MO = SC at levels 2 – 4; MO > SC at level 5
iii) MO ≥ SC > UA

From Table II:

(1) PD + AD = MO + SC + UA

Thus, the ordinal ranking can be summarized:

(2) PD = AD > MO ≥ SC > UA

The relative utility decrements along the five levels

Figure 1a illustrates a distinct pattern of relative utility decrements for each of the five dimensions in each of the CENS-countries: The smallest drops occur between levels 2 and 3 and the largest between levels 3 and 4. Furthermore, drops from levels 1 to 2 appear to be larger than those between levels 4 and 5.

This pattern of relative utility drops is supported by Table III. As a general approximation, we aim for the relative decrements in WePP to correspond as closely as possible with those observed for the CENS-median, i.e. third last column of Table III.

The total scale length:

Two of the four CENS countries' value sets include a small fixed constant (N1) subtracted for all health state combinations other than 11111. A closer look at the CENS-median values in Tables I and III suggests an implicit N1 term of about 0.03. This would have implied a 55555 value of – 0.23, which comes close to the CENS-median for 55555 (– 0.25) reported in Table III. Hence, we seek a scale length for WePP similar to that for the CENS median.

The WePP model was derived based on an observed underlying preference pattern, as expressed by: i) the ordinal importance weights in equations (1) and (2); ii) the relative utility decrements for the CENS median in Table III, and; iii) the scale length of the CENS median (i.e. 55555 in Table III). In addition, a crucial premise was to minimize discrepancies between the modelled values and the CENS-median values. Hence, with identical level for all dimensions, the CENS-median column in Table III suggests equations (3a-d) should be satisfied, where the sub-scripts refer to levels, and the numbers refer to 1 minus the CENS-median values reported in Table III:

(3a) $N1 + MO_2 + SC_2 + UA_2 + PD_2 + AD_2 \approx 0.29\ (1 - 0.71)$
(3b) $N1 + MO_3 + SC_3 + UA_3 + PD_3 + AD_3 \approx 0.44\ (1 - 0.56)$
(3c) $N1 + MO_4 + SC_4 + UA_4 + PD_4 + AD_4 \approx 1.04\ (1 - (-0.04))$
(3d) $N1 + MO_5 + SC_5 + UA_5 + PD_5 + AD_5 \approx 1.25\ (1 - (-0.25))$

After several alternative value sets were explored, the WePP value set presented in Box 1, using two decimals only, came closest to equations (1) – (3a-d).

Box 1: The WePP (Western Preference Pattern) model

Dimensions	MO	SC	UA	PD	AD
Level 2	0.04	0.04	0.03	0.06	0.06
Level 3	0.07	0.07	0.06	0.10	0.10
Level 4	0.17	0.17	0.16	0.25	0.25
Level 5	0.22	0.20	0.19	0.30	0.30
Full health (11111): 1.00; Constant (N1): 0.03					

Equation (1) is satisfied at levels 3 and 4. At level 2, PD + AD = 0.12 > MO + SC + UA = 0.11, and at level 5, there is a reverse absolute difference (0.01). Thus, when considering the aggregate of the four comparisons, the equation holds.

As for Equation (2), PD = AD at all levels. MO = SC at levels 2 – 4, while MO > SC at level 5. UA has slightly lower values than SC at each level.

Generally, when comparing WePP with the CENS median values, we observe a very close correspondence: Among the 20 index values reported in Table I, 7 are identical. For 6 combinations WePP is 0.01 higher than CENS, while the reverse discrepancy is observed for 5 combinations. In the remaining two combinations, the discrepancy is 0.02, one in each direction. Among the discrepancies between *symptoms vs functioning* items at different levels, Table II reports a similarly close correspondence between WePP and CENS median values. As to the proportions of the scale lengths between levels, Table III shows a two percentage point discrepancy in each direction in two of the four level intervals. Finally, with respect to the total scale length, there is a 0.01 difference between WePP and the CENS median for the 55555 combination.

Note that the CENS values are medians, whereas, the WePP values are the values implied if the value set is to conform to a series of rules suggested by characteristics shared by the four value sets. The WePP model is based on some striking similarities in the preference patterns revealed from each of the four value sets, emphasizing the ordinal ranking and the relative differences across dimensions and levels. From these extracted patterns, an amalgam model is developed.

4.2 How does WePP perform?

The proposed WePP model and each of the value sets in the CENS countries were compared using data from the Multi Instrument Comparison (MIC) study, which includes seven major 'disease groups' (arthritis, asthma, cancer, depression, diabetes, hearing loss, heart diseases) and a 'healthy' group (who did not have any known diagnosis) in six OECD-countries (Australia, Canada, Germany, Norway, UK, US) [13].

In the total sample of 7,933 respondents, as many as 1,530 described their health at 11111. Of the remaining 6,403 in non-perfect health: 50% reported only level 2 in one or more dimensions; 30% had levels 3, 2 or 1; 15% had levels 4, 3, 2 or 1, and the remaining 5% had level 5 in one or more dimensions. In other words, in this large study including seven chronically ill patient groups, only 20% had reported a health state combination that includes levels 4 or 5 in at least one dimension.

Based on these data, we can identify differences in mean health state values in the WePP model vs the four value sets from the CENS countries across the whole severity range. We take the summary score (from 11111 = 5 to 55555 = 25) and transform it to an unweighted [0 – 1]-scale, in which 11111 is assigned 1.00, and 55555 is assigned 0.00. Figure 2a provides detailed distribution of the mean values for the 21 unique levels of the transformed summary score, for the total sample. Interestingly, the WePP values lie within the confidence intervals for the value sets in Canada, England, Spain across the whole severity distribution.

Figure 2 a-b: Mean index values by transformed summary score, CENS countries and WePP

Generally, the index values in WePP appear to lie close to the English and Canadian values. An analysis of the agreement between WePP and each of the four CENS value sets revealed strong degrees of concordance: [ICC=0.990; 95% CI: (0.977, 0.995)] for the English value set, followed by Canada [ICC=0.989; 95% CI: (0.987, 0.990)], Spain [ICC= 0.985; 95% CI: (0.968, 0.991)] and the Netherlands [ICC=0.945; (95% CI: 0.787, 0.976)]. Similarly, the mean difference is minimal at the group level when compared with the value sets in Canada (-0.006), England (0.014), and Spain (-0.016). The mean difference when compared with the Dutch value set is higher (-0.045), which is expected since the Dutch index values were generally lower than the other three CENS countries. The WePP model has a narrow variation in the 95% limits of agreement (LOA) when compared to the English value set [95% LOA: (-0.031, 0.058)], indicating small variation between the two measures at the individual level as well. While this 95% LOA is relatively moderate in Canada [95% LOA: (-0.058, 0.046)] and Spain [95% LOA: (-0.072, 0.041)], it is considerably higher for the Dutch value set (particularly among individuals with poorer health states); i.e. the 95% LOA: (-0.153, 0.064).

Furthermore, to see if the WePP model performs differently for physical *vs* mental/psychological symptoms, it was assessed in two disease groups, arthritis (N=929) and depression (N=917). In the arthritis group, the results are consistent with the analyses based on the total sample (see Appendix Figure A2). The findings reveal a very strong agreement between the WePP model and the Canadian tariff [ICC=0.995; 95% CI: (0994, 0.995)] followed by the English tariff [ICC=0.988; 95% CI: (0.977, 0.992)]. A relatively high agreement has been observed with Spain [ICC=0.977; 95% CI: (0.924, 0.989)] and the Netherlands [ICC=0.933; 95% CI: (0.927, 0.973)]. Similar performance was observed in the depression group: the level of agreement as measured by ICC between the WePP and the Canadian, the English, the Spanish, and the Dutch tariffs were 0.995, 0.991, 0.983, and 0.891,

respectively. The variation between the proposed WePP model and the CENS value sets (as measured by the 95% LOA) in the two disease groups were also small. In the arthritis group, the 95% LOA with Canada was (-0.038, 0.039), with England (-0.042, 0.070), with Spain (-0.090, 0.042), and with the Netherlands (-0.173, 0.063). The respective values for the depression group were (-0.025, 0.045), (0.038, 0.062), (-0.082, 0.047), and (-0.225, 0.039). Our results indicate the highest discrepancy from the Netherlands tariff (at the individual level) with the width of the 95% limits of agreement equal to 0.236 for the arthritis group and 0.264 for the depression group. Hence, the WePP model appears to perform equally well for physical and mental symptoms.

To test if the WePP model performs better than the CENS median model, we compared the degree of agreement that the WePP model represents, with those between the CENS median value set for each of the four countries. The WePP model performed better in three of the four countries. Only for Canada did the CENS median model give a slightly stronger degree of concordance (0.995 as compared to 0.990 for WePP).

Despite its breadth in terms of diagnoses, the MIC data only includes 566 (18%) of the 3,125 unique health state combinations in the 5L instrument. In order to locate WePP in relation to each of the four CENS countries' value sets, Figure 2b considers the same dimensions as Figure 2a, by taking the mean preference based value (and confidence interval) at each level of the transformed summary score. The mean values of the WePP lie within the confidence intervals of the value sets from Canada (except at the bottom end), England and Spain.

A quite different test of WePP's performance is to compare its index values with the observed TTO values of the health state combinations on which the EQ-5D-5L modelling is based. The Dutch study reports these cTTO values for each of the 86 combinations included. However, most of them appear to be extremely rare or non-existent in practice. The 7,933 subjects in the MIC study report only 19 of the 86 combinations. However, 10 of which are rare with only 1 to 4 subjects in each. The remaining 9 more prevalent health state combinations (i.e. with 5 or more subjects) represent mild conditions in that they include level 2 in one or two dimensions. When comparing the cTTO based values with those of the Dutch model, and the WePP value set, the WePP comes closer to the observed values than does the Dutch model (Appendix Table A3) for 8 of these 9 health states. However, it remains to test the degree of misspecifications in the moderate or severe states (levels 3, 4 and 5), and for the other countries' value sets.

5: Discussion and conclusion

The seven recently published value sets for the EQ-5D-5L instrument are expected to differ because of differences in cultures, norms and wealth impact upon people's health state preferences. In addition to variations explained by preference diversity across countries, differences in the published value sets will also depend on: i) which elicitation method was used (TTO or DCE); ii) which modelling was chosen, and; iii) the quality of data collection.

Our inquiry into (dis)similarities across the seven value sets revealed some characteristic preference patterns in the four Western countries, Canada, England, the Netherlands and Spain. As to the relative importance of the dimensions, the two symptom dimensions (PD and AD) have similar weights. Their total weight is roughly similar to the total weight of the three functioning dimensions (MO, SC, UA), which also appear to have quite similar weights. As to the relative decrement from one level to the next one down, these are non-linear with three distinct 'kinks'; small decrements from level 2 to 3; large decrements from level 3 to 4, and; small decrements from level 4 to 5. As for the total scale lengths, they all stretch below zero to include states worse than dead. However, the value for the 55555 state differs across the countries.

Most researchers with some experience in preference elicitation would admit that respondents do not hold precise, stable preferences over hypothetical health states (see e.g. [14]). Rather, a respondent's valuation of a described health state would depend on the elicitation method used (the question asked), and a wide range of framing effects. However, while people may have difficulties in expressing exact cardinal values for alternative health states, they may have a clear idea of their ordinal ranking. Furthermore, they may have an idea of the relative values, e.g. that a symptom dimension is one and a half times as important as a functioning dimension. The rationale behind our search for a common currency was to identify any such preference patterns observed when comparing value sets from different countries.

An amalgam model, WePP, was presented to accommodate the key characteristics of preference patterns observed in the CENS countries (Canada, England, the Netherlands and Spain). The agreement of this model with the value set from each of these countries proved to be very strong, and better than the agreement produced by the CENS median model. Our comparisons lend support to the finding that 'East does not meet West' in health state utilities, and that there is less variation in the Western countries [15].

We have demonstrated a successful implementation of an approach based on the distilled common characteristics of several value sets. Still, the WePP value set will be vulnerable to problems characterising the source of the data, namely hypothetical bias. Modelled values will always differ from observed values. The modelled values may be a better approximation to the population values than those observed. However, in order to obtain a model that reflects the complete severity range, some health state combinations are included that are extremely rare in practice. The problem is that the more unreal – or constructed – a described health state combination appears to respondents, the more hypothetical bias is likely to be introduced in the preference elicitation exercise. This problem may arise for those health state combinations which few respondents have experienced, i.e. those which include levels 4 or 5 ('severe' and 'extreme'/'unable'), and particularly when there is no corresponding severities for the symptom items (PD or AD). For example, in the MIC data, only 94 out of the 6,403 respondents with a health state different from 11111 reported a decrement in any of the functioning items without also reporting a decrement in either of the two symptom items. Ten health state combinations covered 50% of respondents, while 50 combinations covered more than 75%. When the vast majority of the 3,125 possible combinations in the 5L system involve potential hypothetical bias, and there are strong arguments put forward for eliciting experienced preferences, there is a case for taking prevalence into account when designing data collection.

The EQ-5D has a dominant position in the estimation of QALY-gains. In order to maintain its key role in applied analyses of high policy relevance, it is important to apply valid and reliable value sets. Some unresolved issues remain: to what extent do the observed differences in value sets reflect genuine preference heterogeneities across these countries, and to what extent do they result from differences in modelling, or the quality of the data collection? This paper has identified some consistent preference patterns. The use of a 'common currency' like the WePP might be a useful option in other Western countries that have yet to develop their own value sets.

References

1. Richardson J, M. J., Bariola E. (2011). Review and critique of health related multi attribute utility instruments. http://www.buseco.monash.edu.au/centres/che/pubs/researchpaper64.pdf. Accessed 18 April 2014.
2. Wisløff, T., Hagen, G., Hamidi, V., Movik, E., Klemp, M., & Olsen, J. A. (2014). Estimating QALY Gains in Applied Studies: A Review of Cost-Utility Analyses Published in 2010. *Pharmacoeconomics*, *32*(4), 367-375, doi:http://10.1007/s40273-014-0136-z.
3. Augustovski, F., Rey-Ares, L., Irazola, V., Garay, O. U., Gianneo, O., Fernandez, G., Morales, M., Gibbons, L., & Ramos-Goni, J. M. (2015). An EQ-5D-5L value set based on Uruguayan population preferences. *Qual Life Res*, doi:10.1007/s11136-015-1086-4.
4. Devlin, N., Shah, K., Feng, Y., Mulhern, B., & Van Hout, B. (2016). Valuing Health-Related Quality of Life: An EQ-5D-5L Value Set for England. https://www.ohe.org/publications/valuing-health-related-quality-life-eq-5d-5l-value-set-england. Accessed March 15 2016.
5. Ikeda, S., Shiroiwa, T., Igarashi, A., Noto, S., Fukuda, T., Saito, S., & Shimozuma, K. (2015). Developing a Japanese version of the EQ-5D-5L value set. *J. Natl. Inst. Public Health*, *64*(1), 47-55.
6. Kim, S.-H., Ahn, J., Ock, M., Shin, S., Park, J., Luo, N., & Jo, M.-W. (2016). The EQ-5D-5L valuation study in Korea. [journal article]. *Quality of Life Research*, *25*(7), 1845-1852, doi:10.1007/s11136-015-1205-2.
7. Ramos-Goni, J. M., Pinto-Prades, J. L., Oppe, M., Cabases, J. M., Serrano-Aguilar, P., & Rivero-Arias, O. (2014). Valuation and Modeling of EQ-5D-5L Health States Using a Hybrid Approach. *Med Care*, doi:10.1097/mlr.0000000000000283.
8. Versteegh, M. M., Vermeulen, K. M., Evers, S. M. A. A., de Wit, G. A., Prenger, R., & Stolk, E. A. (2016). Dutch Tariff for the Five-Level Version of EQ-5D. *Value in Health*, doi:http://dx.doi.org/10.1016/j.jval.2016.01.003.
9. Xie, F., Pullenayegum, E., Gaebel, K., Bansback, N., Bryan, S., Ohinmaa, A., Poissant, L., & Johnson, J. A. (2015). A Time Trade-off-derived Value Set of the EQ-5D-5L for Canada. *Med Care*, doi:10.1097/mlr.0000000000000447.
10. Oppe, M., Devlin, N. J., van Hout, B., Krabbe, P. F. M., & de Charro, F. (2014). A Program of Methodological Research to Arrive at the New International EQ-5D-5L Valuation Protocol. *Value in Health*, *17*(4), 445-453, doi:http://dx.doi.org/10.1016/j.jval.2014.04.002.
11. Janssen, B. M. F., Oppe, M., Versteegh, M. M., & Stolk, E. A. (2013). Introducing the composite time trade-off: a test of feasibility and face validity. *The European Journal of Health Economics*, *14*(Suppl 1), 5-13, doi:10.1007/s10198-013-0503-2.
12. Dolan, P. (1997). Modeling valuations for EuroQol health states. *Med Care*, *35*(11), 1095-1108.
13. Richardson, J., Iezzi, A., & Maxwell, A. (2012). Cross-national comparison of twelve quality of life instruments: MIC Paper 1 Background, questions, instruments. Research Paper 76. http://www.buseco.monash.edu.au/centres/che/pubs/researchpaper76.pdf. Accessed April 10 2014.
14. Lloyd, A. J. (2003). Threats to the estimation of benefit: are preference elicitation methods accurate? *Health Econ*, *12*(5), 393-402, doi:10.1002/hec.772.
15. Xie, F., Pullenayegum, E., Pickard, A. S., Ramos Goni, J. M., Jo, M. W., & Igarashi, A. (2016). Transforming Latent Utilities to Health Utilities: East Does Not Meet West. *Health Econ*, doi:10.1002/hec.3444.