Choice Defines QALYs

A US Valuation of the EQ-5D-5L

Benjamin M. Craig, PhD and Kim Rand, PhD†‡*

Background: The 5-level version of the EQ-5D (EQ-5D-5L) was introduced as an improvement on the original 3-level version (EQ-5D-3L). To date, 6 country-specific value sets have been published for EQ-5D-5L and 9 US value sets have been published for other instruments. Our aims were to (1) produce EQ-5D-5L values on a quality-adjusted life year (QALY) scale from the perspective of US adults and (2) compare them with US EQ-5D-3L values and the other country-specific EQ-5D-5L values.

Methods: In 2016, 8222 US respondents from all 50 states and Washington, DC completed an online survey including a discrete choice experiment with 20 paired comparisons. Each comparison asked respondents, “Which do you prefer?” regarding a pair of alternatives described using EQ-5D-5L and lifespan attributes. On the basis of more than 50 choices on each of the 3160 pairs, we estimated EQ-5D-5L values on a QALY scale and compared them with the US EQ-5D-3L values and the other country-specific EQ-5D-5L values.

Results: Ranging from −0.287 (55555) to 0.992 (11121) on a QALY scale, the estimated EQ-5D-5L values were similar to the US EQ-5D-3L values. Compared with the US EQ-5D-3L values, the values exhibited greater sensitivity and specificity and higher correlation with the EQ-5D-5L values of other countries, particularly England.

Conclusions: Like previous US valuation studies, this study produced nationally representative EQ-5D-5L values on a QALY scale. The results further demonstrate the advantages of the EQ-5D-5L over its 3-level predecessor as a preference-based summary measure of health-related quality of life from the perspective of US adults.

Key Words: quality-adjusted life years, patient preference, choice behavior, health valuation, cost-utility analyses, value of life, quality of life, patient reported outcome measures

(*Med Care* 2018;00: 000–000)

Clinical trials often conduct surveys with patient-reported outcome (PRO) instruments to measure health-related quality of life (HRQOL) as a primary or secondary endpoint. However, treatment effects on HRQOL can be difficult to summarize using distributions of PRO responses or scores alone. To summarize PRO evidence better, preference researchers sometimes present PRO response patterns to stakeholders and ask about their preferences between these outcomes.[1] That is, each item response on each PRO domain is translated into an adjectival statement describing a health attribute, so that stakeholders can better express the value that they place on each outcome. The stakeholders are also told who will experience each outcome and for how long (ie, lifespan attribute). Using evidence from health valuation studies, we can summarize treatment effects on HRQOL as values from the perspective of stakeholders.

The 5-level version of the EQ-5D (EQ-5D-5L) is one of the most commonly used PRO instruments for the measurement of HRQOL in clinical trials and health technology assessments.[2–6] Using health preference evidence from a nationally representative sample of US respondents (Supplementary Fig. 1, Supplemental Digital Content 1, http://links.lww.com/MLR/B557), this study produced EQ-5D-5L values on a quality-adjusted life-year (QALY) scale, with 0 representing the value “immediate death” and 1 representing the value “1 year with no health problems then die.”[1] Like other country-specific value set of the EQ-5D-5L, these results (Appendix, Supplemental Digital Content 4, http://links.lww.com/MLR/B560) will aid in the summary of HRQOL evidence in clinical trials and health technology assessments from the perspective of US adults.[7,8]

In health valuation, preference-elicitation tasks may collect choices, rankings, or ratings.[9,10] In this study, respondents were asked, “Which do you prefer?” of 2 alternatives (ie, paired

From the *Department of Economics, University of South Florida, Tampa, FL; †Department of Health Management and Health Economics, University of Oslo, Oslo; and ‡Health Services Research Centre, Akershus University Hospital, Lørenskog, Norway.

This US valuation study of the EQ-5D-5L was conducted in collaboration with the International Academy of Health Preference Research (IAHPR). The collaboration ensured the authors’ independence in designing the study, interpreting the data, writing, and publishing the report. The views expressed by the authors in the publication do not necessarily reflect the views of IAHPR.

Supported by the EuroQol Research Foundation and a grant from the National Institutes of Health, Department of Health and Human Services, through the National Cancer Institute (1R01CA160104). This study was funded by the Foundation under a call for methodological research on discrete choice experiments, rather than for the creation of an EQ-5D-5L value set using an interview-based time trade-off (TTO) task. The EuroQol funding agreements ensured the authors’ independence in designing the study, interpreting the data, writing, and publishing the report.

The views expressed by the authors in the publication do not necessarily reflect the views of the EuroQol Group.

The authors declare no conflict of interest.

Reprints: Benjamin M. Craig, PhD, Department of Economics, University of South Florida, 4202 E. Fowler Avenue, CMC206A, Tampa, FL 33620. E-mail: bcraig@usf.edu.

Supplemental Digital Content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal’s website, www.lww-medicalcare.com.

ISSN: 0025-7079/18/000-000

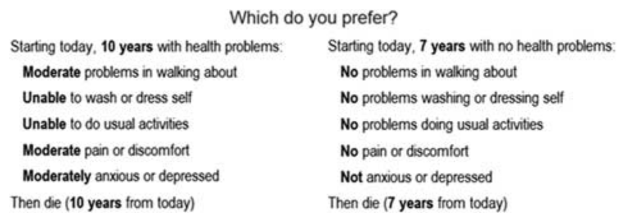

FIGURE 1. Paired comparison.

comparison; Fig. 1). By contrast, a time trade-off (TTO) is a matching algorithm in which respondents complete an iterative procedure composed by series of consecutive and adapted choices (ie, ping-pong) adjusting the lifespan attribute of one alternative until indifference is achieved (ie, match). To date, 7 US valuation studies have been published covering a wide variety of PRO instruments [TTO: 3-level version of the EQ-5D (EQ-5D-3L; 3 value sets); Paired Comparison (6 value sets): Short-Form Six-Dimension Health Index (SF-6D), the Patient-Reported Outcomes Measurement Information System 29-item Profile (PROMIS-29), the child-friendly version of the EQ-5D (EQ-5D-Y), the Behavioral Problems Index (BPI), the National Survey of Children with Special Health Care Needs (NS-CSHCN), and the Patient-Reported Outcomes version of the Common Terminology Criteria for Adverse Events (PRO-CTCAE)].[4,9,11–19]

The first EQ-5D valuation study conducted in the United States used an interviewer-based TTO task, replicating earlier protocols.[19,20] In the original paper, the authors arbitrarily transformed a quarter of the TTO responses (ie, all responses lower than immediate death) to improve the face validity of the estimated EQ-5D values. Subsequently, the original authors and others proposed alternative estimation techniques that avoid data manipulation,[12,13,21] but these did not address the limitations inherent to the TTO task.[15] Since then, paired-comparison and ranking tasks were introduced as an alternative to the TTO.[1,4,9,11,15–18,22–25] As an attempt to compromise, hybrid models were introduced that merged discrete-choice and TTO responses.[12]

Using paired comparison evidence from a large nationally representative survey (N = 8222; Supplementary Fig. 1 and Table 1, Supplemental Digital Content 1, http://links.lww.com/MLR/B557), our aims were to (1) produce EQ-5D-5L values on a QALY scale from the perspective of US adults and (2) compare these value with US EQ-5D-3L values and the other country-specific EQ-5D-5L values. The comparative analysis included the EQ-5D-5L values from the first aim as well as a published outcomes dataset[3] with EQ-5D-3L and EQ-5D-5L responses and a series of published EQ-5D-3L and EQ-5D-5L value sets.[26–31] The valuation evidence also served as the basis for an econometrics competition on the predictive validity of alternative modeling approaches, which was recently published.[32,33]

METHODS

Aim 1: US Valuation of the EQ-5D-5L

As the largest health valuation study ever conducted, with 8222 respondents, this study was intentionally over-powered so that it could demonstrate differences in modelling approaches in health valuation. In total, this survey collected 164,440 paired comparison responses (Fig. 1; ie, 20 pairs×8222 respondents). Further details of the survey instrument, choice sets, recruitment, and quota sampling can be found on the study web site and related publications.[32,33]

Like previous US valuation studies, respondents were recruited from each US state and Washington, DC (Supplementary Fig. 1 and Table 1, Supplemental Digital Content 1, http://links.lww.com/MLR/B557) using a nationally representative panel.[34] Each of the 3160 pairs had 50 or more respondents following 18 demographic quotas: [Men, Women]×[Age 18–34, 35–54, 55+]×[Hispanic, non-Hispanic black or African American, non-Hispanic, white, other]. To better understand the demographic and socioeconomic representativeness of the sample, the respondent characteristics are provided in Supplementary Table 1 (Supplemental Digital Content 1, http://links.lww.com/MLR/B557).

The Effect of EQ-5D-5L and Lifespan Attributes on Value

The EQ-5D-5L instrument was introduced as an improvement on the original 3-level version for HRQOL measurement and valuation.[2,3,5,6] Specifically, it was designed to increase sensitivity and reduce the ceiling effect and gaps in the value sets. Its descriptive system includes 5 problems: Mobility, Self-Care, Usual Activities, Pain/Discomfort, and Anxiety/Depression. Each problem is characterized as being at one of 5 levels: none (level 1), slight, moderate, severe, unable/extreme (level 5). In shorthand notation, these attributes are shown as vectors of 5 numbers (eg, Fig. 1 includes 35533 and 11111). Aside from health attributes, the lifespan attribute is shown in days, weeks, months and years and ranging from "immediate death" to 20 years (eg, Fig. 1 includes 10 and 7 years).

Value is a function of EQ-5D-5L and lifespan attributes and expressed on a QALY scale. A previous analysis demonstrated that the relationship between value and these attributes may be best represented by a power function (ie, value = V×Lifespan$^{\alpha}$, where $\alpha \leq 1$). For example, if V(33333) is 0.75 (see below for details) and the power, α, is 0.5 (square root), then the values of 33333 for 3 months, 1 year, and 4 years are 0.375, 0.75, and 1.5 QALYs, respectively. Like the previous study, the value function includes different power parameters depending on the temporal unit of lifespan (days, months, weeks, or years).

Econometric Analysis

Each respondent completed 20 paired comparisons (Fig. 1); therefore, the responses are not fully independent. Instead of fully parameterizing their correlation (ie, random-effects), we estimated the parameters by maximizing likelihood with clusters [ie, maximize, cluster (respondent identifier)]. This clustered maximization process does not assume any particular specification for within-cluster correlation, yet it relaxes the assumption of independence within clusters.

For each paired comparison, the likelihood of a choice is described with a Zermelo-Bradley-Terry (ZBT; Technical Appendix 1, Supplemental Digital Content 2, http://links.

lww.com/MLR/B558) cumulative density function (CDF), P (A > B) = $D_B/(D_A+D_B)$, where D_B represents the opportunity cost of choosing B over A and D_A represents the opportunity cost of choosing A over B (eg, longer lifespan vs. relief of health problems). For comparison, a logit CDF assumes that the log odds [ie, ln(P(A > B)/P(B > A)] is proportional to the difference in opportunity costs (ie, $D_B–D_A$). Unlike logit CDF, the opportunity costs, D_A and D_B, in a ZBT CDF are always non-negative (ie, there is no such thing as a free lunch); otherwise, the choice would be deterministic [ie, P (A > B) is bounded by 0 and 1]. Under both CDFs, the likelihood of choosing A over B decreases with the cost of A (D_A) and increases with the cost of B (D_B).

In Figure 1, $D_B = V(11111)\times(10^{\alpha}–7^{\alpha})$ and $D_A = (V(11111)–V(35533))\times7^{\alpha}$. If V(35533) is 0.75 and the power, α, is 1, $D_B = 1\times(10–7) = 3$ and $D_A = (1–0.75)\times7 = 0.25\times7 = 1.75$; therefore, the likelihood is 3/(3+1.75), or 0.63, favoring the longer lifespan (A; left). However, if the power, α, were 0.5 (square root), $D_B = 1\times(10^{0.5}–7^{0.5}) = 0.52$ and $D_A = (1–0.75)\times7^{0.5} = 0.25\times7^{0.5} = 0.66$; therefore, the likelihood would be 0.52/(0.52+0.66), or 0.44, favoring relief of health problems (B; right). If power, α, was erroneously constrained to be 1 (ie, constant proportionality), the model would poorly predict choices and produce biased estimates, compressing the range (ie, mild outcomes would become more severe and severe outcomes would become more mild).

Two regression models were estimated to produce EQ-5D-5L values (eg, V(33333)). The first model included only the 20 effect-coded dummy variables, which is the standard approach to modeling the EQ-5D-5L. Each coefficient represents the loss in EQ-5D-5L value incurred by an increase of 1 level-increment in a domain [eg, going from level 2 (slight problems) to level 3 (moderate problems) on mobility]. For example, V(33333) equals 1 minus the sum of the first 2 coefficients of each domains. The second model included the same 20 effect-coded dummy variables along with counts of the number of domains with levels > 1, 2, 3, and 4. A positive (negative) coefficient of a level count implies that the sum of its problems is worse (better) than its parts.

We hypothesized that each of the 20 effect-coded coefficients is positive (ie, worse health reduces value) with and without level counts, and that each of the 4 powers (days, weeks, months, years) is <1, rejecting the constant proportionality assumption.

Aim 2: Comparisons of EQ-5D-3L and EQ-5D-5L Values

Do the EQ-5D-5L Values Discriminate Better than the EQ-5D-3L Values?

To assess discrimination and convergent validity, we used a published dataset of US respondents who completed the EQ-5D-5L and EQ-5D-3L instruments as well as self-reported health status on a 5-point scale (excellent, very good, good, fair and poor) and a 0–100 numerical visual analog scale (VAS).[3] This online survey used a similar recruitment strategy of the US general population (N = 2619), and showed that health problems are more common, but less severe when measured using EQ-5D-5L. In the current study, we compare the values of 2 "known groups" (297 with fair and 75 with poor self-reported health) and estimated the association between values and VAS by instrument.

Sensitivity is defined as the likelihood that the equivalence in means between respondents in fair and poor health is rejected at a significance of 0.01. For each of 10,000 iterations, we randomly drew samples (of 25–75 respondents) from each strata with replacement and tested for equivalence in means.

Specificity is defined as the likelihood of failing to reject equivalence in means of 2 randomly drawn samples with equal number of respondents in fair and poor health at a significance of 0.01. After combining these 2 samples, we split the respondents into 2 new samples that were perfectly balanced in terms of fair and poor health. Then we tested for equivalence in means. We hypothesized that the EQ-5D-5L values have higher sensitivity and specificity, and thus discriminate better than the EQ-5D-3L values.

Do the Values Change When You Switch From the EQ-5D-3L to EQ-5D-5L?

The comparison between the EQ-5D-3L and EQ-5D-5L value sets is challenging because some of the differences are due to measurement and others are due to values. However, among the 3125 possible combinations (5^5) of EQ-5D-5L attributes, 72 are identical to those of the EQ-5D-3L descriptive system. For example, the description for EQ-5D-5L 35533 (Fig. 1) is the same as that for EQ-5D-3L 23322. These 72 common descriptions illustrate differences among the value sets. Furthermore, the EQ-5D-5L values were compared to the crosswalk values provided by the EuroQol Foundation as an interim value set and based on the US EQ-5D-3L values.[35] For the 72 EQ-5D-3L and 3124 crosswalk values, we estimated their unadjusted association with the EQ-5D-5L values by Pearson correlation, Spearman correlation, and Lin concordance coefficient.

Do the EQ-5D-5L Values Change When You Switch Between Countries?

We compared the EQ-5D-5L values from this study to the values of 6 countries (Table 2; Appendix, Supplemental Digital Content 4, http://links.lww.com/MLR/B560).[26–31] In addition to comparing their ranges, we applied each value set to the published dataset of US respondents (297 fair and 75 poor self-reported health) and showed the difference in means. Furthermore, we estimated their unadjusted association between country-specific values (Pearson correlation, Spearman correlation, and Lin concordance coefficient).

EQ-5D-5L value sets will generally display high correlation because they are based on the same descriptive system and not because of the values themselves (ie, descriptive system bias; Supplementary Fig. 2; Supplemental Digital Content 1, http://links.lww.com/MLR/B557, Technical Appendix 2, Supplemental Digital Content 3, http://links.lww.com/MLR/B559). Using the combined value sets from the 7 countries, we regressed the EQ-5D-5L value on the descriptive system (ie, misery index) and 8 country-specific indicators (8 parameters) and examined the association between the country-specific residuals and the US residuals. These adjusted measures of association may be more informative that the unadjusted ones, because they describe the correlation and concordance between country-specific values after controlling for their common descriptive system.

RESULTS

Aim 1: US Valuation of the EQ-5D-5L

Table 1 shows the parameter estimates for the models without and with the level counts. In both models, the effect-coded coefficients are positive and each of the 4 powers is > 1, letting us reject the constant-proportionality assumption. Power increases with temporal unit (ie, respondents had shorter time horizons when lifespan was expressed in days rather than years). The confidence intervals are smaller than those of the EQ-5D-3L valuation study, which failed to show significant differences between levels in each domain.

The effect of shifting from none to unable/extreme on value varies by domain and is similar with and without the level counts. Relief of pain/discomfort is the most valuable domain (0.369 and 0.324, respectively), and the largest parameter is the shift from moderate to severe pain/discomfort (0.218 and 0.204, respectively). The level-count coefficients are nonzero individually (*t* tests) and together (Wald test) at a significance level of <0.01, favoring their inclusion. Therefore, the EQ-5D-5L values shown in the Appendix, Supplemental Digital Content 4 (http://links.lww.com/MLR/B560), are based on the 28-parameter model.

Aim 2: Comparisons of EQ-5D-3L and EQ-5D-5L Values

Figure 2 shows the sensitivity and specificity of the EQ-5D-5L and EQ-5D-3L values by sample size. A sample size of 50 would be sufficient to reject the equivalence in means between respondents in fair and poor health 87% and 84% of the time, respectively (*P*-value <0.01). If the fair-health patients were evenly mixed into both groups, a sample size of 50 patients would be sufficient to fail to reject equivalence in means 99.4% and 99.3% of the time, respectively. Compared with EQ-5D-3L values, EQ-5D-5L values are more correlated with self-reported health on a VAS (Pearson ρ, 0.500 vs. 0.503; Spearman ρ, 0.478 vs. 0.504). Although the differences in sensitivity, specificity, and convergent validity are modest (< 0.05), this evidence shows that the EQ-5D-5L discriminates between fair and poor health better and concords with VAS better than the EQ-5D-3L.

Figure 3A illustrates the relationship between the EQ-5D-5L and EQ-5D-3L values on the 72 common descriptions. Three noteworthy patterns emerged. First, the EQ-5D-3L values display a large gap at the top and a shorter range: the EQ-5D-3L values range from –0.102 (33333) to 0.86 (11211), and the EQ-5D-5L values range from –0.287 (55555) to 0.992 (11121). Second, there appears to be a series of shifts corresponding to the level counts. The EQ-5D-3L value set is partly based on a count of level 3 (ie, N3 term), and the EQ-5D-5L values are based on 4 level counts. Nevertheless, the EQ-5D-5L and EQ-5D-3L values are highly correlated (unadjusted Pearson ρ, 0.96; Spearman ρ, 0.98; Lin ρ, 0.81).

TABLE 1. Models of 5-Level Version of the EQ-5D Values*

8222 US Respondents	Without Level Counts			With Level Counts†		
	Coefficient	95% CI		Coefficient	95% CI	
Mobility						
Level 1–2	0.014	0.012	0.016	0.021	0.017	0.024
Level 2–3	0.016	0.013	0.018	0.012	0.009	0.015
Level 3–4	0.131	0.120	0.142	0.036	0.030	0.043
Level 4–5	0.084	0.070	0.097	0.042	0.031	0.053
Self-Care						
Level 1–2	0.018	0.015	0.020	0.023	0.018	0.027
Level 2–3	0.032	0.027	0.037	0.021	0.017	0.026
Level 3–4	0.123	0.111	0.135	0.023	0.015	0.030
Level 4–5	0.104	0.090	0.119	0.056	0.044	0.069
Usual Activities						
Level 1–2	0.011	0.009	0.013	0.016	0.013	0.020
Level 2–3	0.016	0.013	0.018	0.009	0.006	0.013
Level 3–4	0.110	0.100	0.119	0.020	0.015	0.026
Level 4–5	0.061	0.050	0.073	0.027	0.017	0.037
Pain/Discomfort						
Level 1–2	0.009	0.007	0.011	0.014	0.011	0.018
Level 2–3	0.016	0.014	0.019	0.011	0.008	0.014
Level 3–4	0.218	0.203	0.233	0.090	0.080	0.099
Level 4–5	0.126	0.109	0.144	0.070	0.056	0.084
Anxiety/Depression						
Level 1–2	0.022	0.019	0.024	0.025	0.021	0.029
Level 2–3	0.020	0.017	0.022	0.013	0.009	0.016
Level 3–4	0.164	0.152	0.176	0.052	0.045	0.060
Level 4–5	0.041	0.033	0.049	0.011	0.004	0.018
Level counts						
Count of levels > 1	—	—	—	–0.006	–0.010	–0.003
Count of levels > 2	—	—	—	0.008	0.005	0.011
Count of levels > 3	—	—	—	0.114	0.104	0.124
Count of levels > 4	—	—	—	0.023	0.014	0.031
Power (α)						
Lifespan in days	0.158	0.142	0.173	0.138	0.124	0.153
Lifespan in weeks	0.185	0.167	0.202	0.164	0.147	0.181
Lifespan in months	0.228	0.206	0.249	0.209	0.189	0.229
Lifespan in years	0.335	0.304	0.365	0.304	0.277	0.332

*Without the level counts, experiencing the worst possible EQ5L description for 1 year (55555) equals the loss of 1.336 quality-adjusted life years (sum of all 20 effects-coded coefficients). With level counts, this loss decreases to 1.287 quality-adjusted life years (sum of all 20 effects-coded coefficients plus 5 times the level-count coefficients). Regardless, experiencing the worst possible EQ-5L description for 1 year (55555) is worse than "immediate death" (ie, –0.336 and –0.287, respectively).

†For example, the value of 1 year in 22211 equals 1–(0.021+0.023+0.016–0.006×3) based on the model with level counts.

CI indicates confidence interval.

The comparison between the EQ-5D-5L and crosswalk values (Fig. 3B)[35] show 2 other noteworthy patterns. First, the crosswalk values also have a shorter range than the EQ-5D-5L values, but they have a wider range than the EQ-5D-3L values: –0.109 (55555) to 0.888 (11211). Second, the EQ-5D-5L and crosswalk values are also correlated (unadjusted Pearson ρ, 0.88; Spearman ρ, 0.88; Lin ρ, 0.80).

This valuation of EQ-5D-5L from the perspective of US adults was compared with the value vets from 6 other

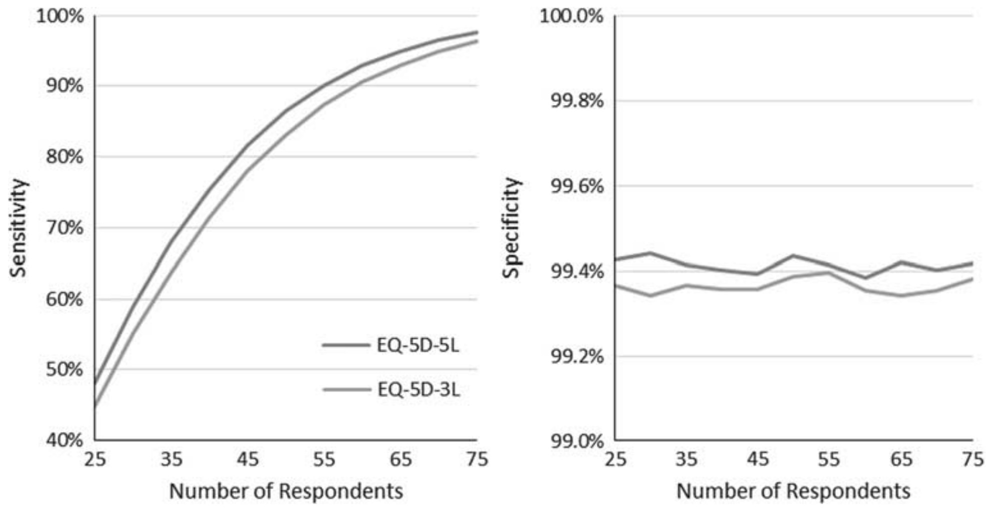

FIGURE 2. Discrimination of the EQ-5D-5L and EQ-5D-3L values on a quality-adjusted life year scale*. *Sensitivity is the likelihood that the equivalence in means between respondents in fair and poor health is rejected at a significance of 0.01. Specificity is the likelihood of failing to reject equivalence in means of 2 randomly drawn samples with equal number of respondents in fair and poor health. EQ-5D-3L indicates 3-level version of the EQ-5D; EQ-5D-5L, 5-level version of the EQ-5D.

countries (Table 2; Supplementary Fig. 2, Supplemental Digital Content 1, http://links.lww.com/MLR/B557; Appendix, Supplemental Digital Content 4, http://links.lww.com/MLR/B560).[26–31] Japan and Korea have the narrowest ranges in values (0.920 and 0.949). The Netherlands and the United States have the widest ranges (1.364 and 1.279). Among the United States and the 6 other countries, there are large differences between the lowest values (from −0.399 the Netherlands to −0.066 Korea) and the highest (0.883 Korea to 0.992 the United States); however, the differences in maximum, minimum, and range do not explain why the United States mean value for fair and poor health is greater than those of the other countries.

At first glance, EQ-5D-5L value sets appears highly correlated. After the country-specific ceiling effects and descriptive system bias are controlled for, however, the values of US adults appear more similar to the values of England. In fact, the concordance between the United States and England (Lin ρ, 0.814) is greater than the concordance between Canada and England (0.780), Canada and the Netherlands (0.640), and the Netherlands and England (0.782), even though these 3 countries used the same interview-based TTO approach.

DISCUSSION

This study produced EQ-5D-5L values (Appendix, Supplemental Digital Content 4, http://links.lww.com/MLR/B560) that are representative of US adults, on a QALY scale, and may be directly implemented by researchers interested in the summary EQ-5D-5L evidence in clinical trials and health technology assessments to inform resource allocation decisions. On the basis of the Aim 2 comparative results, the

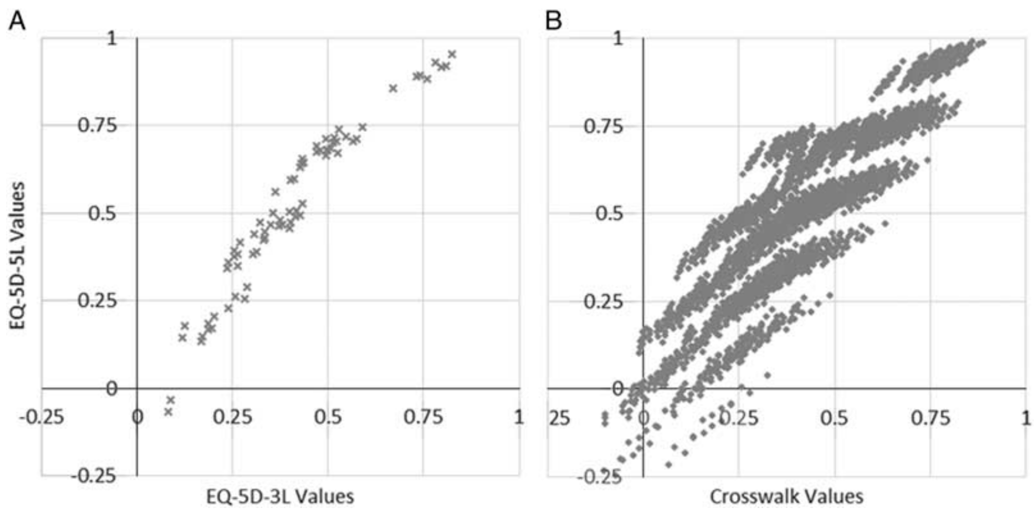

FIGURE 3. Comparison of EQ-5D-5L, EQ-5D-3L*, and crosswalk values. *The EQ-5D-3L values are shown only for the 72 descriptions shared with the EQ-5D-5L. EQ-5D-3L indicates 3-level version of the EQ-5D; EQ-5D-5L, 5-level version of the EQ-5D.

TABLE 2. Relationship Between the 5-Level Version of the EQ-5D Values of the United States and 6 Other Countries

	Japan	Korea	Uruguay	The Netherlands	Canada	England	USA
QALYs							
Lowest value	−0.025	−0.066	−0.264	−0.399	−0.148	−0.281	−0.287
Highest value	0.895	0.883	0.978	0.965	0.929	0.951	0.992
Fair health*	0.700	0.729	0.841	0.713	0.729	0.740	0.860
Poor health	0.545	0.574	0.690	0.477	0.544	0.537	0.672
Measures of association, unadjusted (N = 3124)†							
Pearson ρ	0.924	0.943	0.947	0.935	0.972	0.974	—
Spearman ρ	0.920	0.940	0.949	0.929	0.973	0.972	—
Lin ρ	0.817	0.847	0.925	0.762	0.868	0.876	—
Measures of association, adjusted for ceiling effects and descriptive system bias (N = 3124)							
Pearson ρ	−0.120	0.282	0.625	0.665	0.752	0.823	—
Spearman ρ	−0.126	0.255	0.606	0.663	0.742	0.816	—
Lin ρ	−0.099	0.274	0.620	0.636	0.697	0.814	—

*The mean values of US respondents in fair and poor self-reported health (N = 297 and 75, respectively) are shown by country-specific value set. Regardless of country, the differences in mean value between fair and poor health is positive ($P < 0.01$).

†The EQ-5D-5L descriptive systems contains 5 domains with 5 levels ($5^5 = 3125$). We excluded the value for 11111 from the measures of association, because its value is constant (ie, V(11111) = 1 QALY) across all descriptive systems, except for Canada (0.949 QALYs).

QALYs indicates quality-adjusted life years.

EQ-5D-5L values were similar to the EQ-5D-3L values, except with smaller SE, no large gap near full health, and a wider range. The results demonstrated the advantages of the EQ-5D-5L values over its 3-level predecessor in terms of discrimination and convergent validity.

The Aim 1 analysis showed that all 20 effect-coded parameters were significantly positive, which was a concern because of evidence on preference inversions between level 4 and 5 on anxiety/depression.[2,4] Its results also confirmed that the count of level influences the EQ-5D-5L values (interaction effects) and that the power parameters on lifespan were <1 (rejecting the constant proportionality assumption; $P < 0.01$) and varied by temporal unit.

The Aim 2 analysis showed that, after adjustments for ceiling effects and descriptive system biases, US values were highly correlated with the values of England and Canada. Although the valuation studies of the Netherlands, England, and Canada applied similar interview-based TTO protocol, the concordances among these 3 countries are lower than the concordance between the English and US values. This suggests that similarity in language (English) may be more important than protocols.

Recent Innovations in Health Valuation

In this study, we applied 3 recent innovations in health valuation. First, the initial paired comparison studies (A vs. B; Fig. 1) used sigmoidal CDFs to estimate US value sets.[9,12,14,17] These functions are common in biostatistics and preference research (eg, logits) and are based on additive differences in value (eg, A–B). In terms of predictive validity, sigmoidal CDFs have been shown to be inferior to proportional CDFs, ones based on value ratios (A/B), for health valuation.[32] Except for the first 2 (EQ-5D-3L and SF-6D), all US health valuation studies have used proportional CDF, namely ZBT (Technical Appendix 1, Supplemental Digital Content 2, http://links.lww.com/MLR/B558). Unlike other health preference studies, health valuation studies examine the value of a wide range of alternatives (eg, mild to severe problems; immediate death to 20-y lifespan). Proportional CDF are particularly useful in health valuation, because scale cancels within the value ratio. For example, doubling time with pain and depression has a large effect on their additive difference (A–B), but no effect on their ratio (A/B) and little effect on choice. Sigmoidal CDF may work well if the range of alternatives is narrow (eg, comparing similar treatments), but proportional CDFs are more suitable for wider ranges of values.[4,11,15,16,18,32] Replacing sigmoidal CDFs with proportional CDFs improves predictive validity and controls for scaling biases.

Second, all previous US value sets were estimated under the assumption of a constant proportional relationship between value and lifespan (ie, value = V×Lifespan). Yet, respondents typically exhibit decreasing marginal value of lifespan (ie, discounting, time preferences).[36–44] A recent study demonstrated that a power function can relax the constant-proportionality assumption (ie, value = V×Lifespan$^{\alpha}$, where $\alpha \leq 1$) by adding only one parameter (Supplementary Fig. 3, Supplemental Digital Content 1, http://links.lww.com/MLR/B557).[45] Not only was the constant proportionality rejected, this assumption caused a bias that compressed the range of values on the QALY scale, namely increasing the gap between full health and mild outcomes and causing severe problems to appear less burdensome. Relaxing this assumption using the power function improves predictive validity and reduces the compression in values. This is the first health valuation study to relax the constant proportionality assumption to produce a country-specific value set for any PRO instrument. In future research, the US TTO-based value set of the EQ-5D-3L may be reestimated to incorporate the power function.

Third, multiple studies have compared alternative preference-elicitation methods and concluded that online surveys produce preference evidence equivalent to that of interviews,[46] avoid interviewer effects,[8] and have fewer sequence effects (ie, decreases in modal response with each additional choice task) than interviewer-based tasks.[7,15] With the lower cost and increasing generalizability of online surveys, discrete choice experiments (DCEs), such as paired comparisons, have become the most common approach to producing US value sets (6/7) replacing the more expensive, interview-based tasks (eg, TTO). A study of this size and with coverage of all 50 US states (Supplementary Fig. 1, Supplemental Digital Content 1, http://links.lww.com/MLR/B557) would have been cost-prohibitive without the recent innovations in online surveys.

Limitations

The sample was representative of the United States in terms of self-reported age, gender, race, ethnicity and geographic region (Supplementary Fig. 1, Supplemental Digital Content 1, http://links.lww.com/MLR/B557 and Table 1) due to its use of an online panel.[34] Such quota sampling techniques do not address unobservable factors and online panels typically underrepresent people of low income and low educational attainment. Compared with the 2010 US Census, the sample has similar proportions of Hispanic respondents (12.16% vs. 14.22%) and black or African American respondents (12.24% vs. 11.97%), but has a lower proportion other nonwhite respondents (6.40% vs. 13.37%). This difference may be attributed to a lack of quota sampling for other racial categories (eg, Asian), the lack of a “2 or more races” category in the survey instrument, or the use of an online panel. This limitation is commonplace in country-specific valuation studies. Door-to-door interview survey would have made the study cost-prohibitive and may not have improved the generalizability, particularly to low socioeconomic status respondents.

Furthermore, this study did not use the interview-based TTO approach and constant proportionality assumption applied in the 6 other country-specific valuation studies of the EQ-5D-5L, which complicates the comparison EQ-5D-5L value sets. Additional challenges in EQ-5D translation to non-English languages may explain further differences.[2] Although these differences have scientific and practical motives, they detract from efforts to promote a uniform approach in health valuation and some may seek to improve consistency further by enhancing methodologic standardization and translation processes, which is beyond the scope of this study.

CONCLUSIONS

“Choice defines value” is a mantra within the health preference community.[32] This paper shows how choices produce EQ-5D-5L values on a QALY scale (ie, choice defines QALYs). Except for EQ-5D-3L,[13,19–21] all US valuation studies used online paired comparisons.[4,11,15–18] The resulting values (Appendix, Supplemental Digital Content 4, http://links.lww.com/MLR/B560) are similar to the US EQ-5D-3L values and the EQ-5D-5L values of other countries. Although the gains in discrimination are modest, these EQ-5D-5L values should be suitable for the summary of HRQOL evidence in clinical trials and cost-utility analyses from the perspective of US adults.

Each of the 6 other country-specific EQ-5D-5L studies conducted ~1000 interviews using a structured interview protocol and quality control tools developed over the last decade (EQ-VT).[7,8] For the US, 1000 similar interviews would cost more than 10 times as much as 8000 online surveys.[47] Considering the demonstrated potential of paired comparisons, our approach appears to be a cost-effective method to conduct a national valuation study. Future US valuation studies of the EQ-5D-5L may be directly compared with this one in terms of cost, values, and uncertainty, which may be informative for those planning studies in other countries.[47]

REFERENCES

1. Stolk EA, Craig BM, Mulhern B, et al. Health valuation: demonstrating the value of health and lifespan. *Patient*. 2017;10:515–517.
2. Craig BM, Monteiro AL, Herdman M, et al. Further evidence on EQ-5D-5L preference inversion: a Brazil/US collaboration. *Qual Life Res*. 2017;26: 2489–2496.
3. Craig BM, Pickard AS, Lubetkin EI. Health problems are more common, but less severe when measured using newer EQ-5D versions. *J Clin Epidemiol*. 2014;67:93–99.
4. Craig BM, Brown DS, Reeve BB. Valuation of child behavioral problems from the perspective of US adults. *Med Decis Making*. 2016;36:199–209.
5. Feng Y, Devlin N, Herdman M. Assessing the health of the general population in England: how do the three- and five-level versions of EQ-5D compare? *Health Qual Life Outcomes*. 2015;13:171.
6. Herdman M, Gudex C, Lloyd A, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Qual Life Res*. 2011;20:1727–1736.
7. Oppe M, Rand-Hendriksen K, Shah K, et al. EuroQol protocols for time trade-off valuation of health outcomes. *Pharmacoeconomics*. 2016;34: 993–1004.
8. Ramos-Goni JM, Oppe M, Slaap B, et al. Quality Control Process for EQ-5D-5L Valuation Studies. *Value Health*. 2017;20:466–473.
9. Craig BM, Busschbach JJ, Salomon JA. Modeling ranking, time trade-off, and visual analog scale values for EQ-5D health states: a review and comparison of methods. *Med Care*. 2009;47:634–641.
10. Craig BM, Lancsar E, Muhlbacher AC, et al. Health preference research: an overview. *Patient*. 2017;10:507–510.
11. Craig BM, Brown DS, Reeve BB. The value adults place on child health and functional status. *Value Health*. 2015;18:449–456.
12. Craig BM, Busschbach JJ. The episodic random utility model unifies time trade-off and discrete choice approaches in health state valuation. *Popul Health Metr*. 2009;7:3.
13. Craig BM, Busschbach JJ. Revisiting United States valuation of EQ-5D states. *J Health Econ*. 2011;30:1057–1063.
14. Craig BM, Busschbach JJ, Salomon JA. Keep it simple: ranking health states yields values similar to cardinal measurement approaches. *J Clin Epidemiol*. 2009;62:296–305.
15. Craig BM, Greiner W, Brown DS, et al. Valuation of child health-related quality of life in the United States. *Health Econ*. 2015;25:768–777.
16. Craig BM, Mitchell SA. Examining the value of menopausal symptom relief among US women. *Value Health*. 2016;19:158–166.
17. Craig BM, Pickard AS, Stolk E, et al. US valuation of the SF-6D. *Med Decis Making*. 2013;33:793–803.
18. Craig BM, Reeve BB, Brown PM, et al. US valuation of health outcomes measured using the PROMIS-29. *Value Health*. 2014;17:846–853.
19. Shaw JW, Johnson JA, Coons SJ. US valuation of the EQ-5D health states: development and testing of the D1 valuation model. *Med Care*. 2005;43:203–220.
20. Gudex C. *Time Trade-Off User Manual: Props and Self-Completion Methods Report of the Centre for Health Economics*. York, UK: University of York; 1994.
21. Shaw JW, Pickard AS, Yu S, et al. A median model for predicting United States population-based EQ-5D health state preferences. *Value Health*. 2010;13:278–288.
22. Craig BM, Reeve BB, Cella D, et al. Demographic differences in health preferences in the United States. *Med Care*. 2014;52:307–313.
23. Maydeu-Olivares A, Bockenholt U. Modeling subjective health outcomes: top 10 reasons to use Thurstone’s method. *Med Care*. 2008; 46:346–348.
24. Krabbe PF. Thurstone scaling as a measurement method to quantify subjective health outcomes. *Med Care*. 2008;46:357–365.

25. Veit CT, Rose BJ, Ware JE Jr. Effects of physical and mental health on health-state preferences. *Med Care*. 1982;20:386–401.
26. Augustovski F, Rey-Ares L, Irazola V, et al. An EQ-5D-5L value set based on Uruguayan population preferences. *Qual Life Res*. 2016;25:323–333.
27. Devlin N, Shah K, Mulhern B, et al. Valuing health-related quality of life: an EQ-5D-5L value set for England. Office of Health Economics Research Paper Series. 2016. Available at: https://www.ohe.org/publications/valuing-health-related-quality-life-eq-5d-5l-value-set-england. Accessed April 2, 2018.
28. Kim SH, Ahn J, Ock M, et al. The EQ-5D-5L valuation study in Korea. *Qual Life Res*. 2016;25:1845–1852.
29. MV M, MV K, MAAE S, et al. Dutch Tariff for the Five-Level Version of EQ-5D. *Value Health*. 2016;19:343–352.
30. Shiroiwa T, Ikeda S, Noto S, et al. Comparison of value set based on DCE and/or TTO data: scoring for EQ-5D-5L health states in Japan. *Value Health*. 2016;19:648–654.
31. Xie F, Pullenayegum E, Gaebel K, et al. A time trade-off-derived value set of the EQ-5D-5L for Canada. *Med Care*. 2016;54:98–105.
32. Jakubczyk M, Craig BM, Barra M, et al. Choice defines value: a predictive modeling competition in health preference research. *Value Health*. 2018;21:229–238.
33. Craig BM. EQ DCE predictive modeling competition. 2016. Available at: http://iahpr.org/eq-dce-competition/. Accessed February 10, 2018.
34. Craig BM, Hays RD, Pickard AS, et al. Comparison of US panel vendors for online surveys. *J Med Internet Res*. 2013;15:e260.
35. van Hout B, Janssen MF, Feng YS, et al. Interim scoring for the EQ-5D-5L: mapping the EQ-5D-5L to EQ-5D-3L value sets. *Value Health*. 2012;15:708–715.
36. Attema AE, Brouwer WB. On the (not so) constant proportional trade-off in TTO. *Qual Life Res*. 2010;19:489–497.
37. Attema AE, Brouwer WB. Constantly proving the opposite? A test of CPTO using a broad time horizon and correcting for discounting. *Qual Life Res*. 2012;21:25–34.
38. Craig BM. The duration effect: a link between TTO and VAS values. *Health Econ*. 2009;18:217–225.
39. Dolan P, Stalmeier P. The validity of time trade-off values in calculating QALYs: constant proportional time trade-off versus the proportional heuristic. *J Health Econ*. 2003;22:445–458.
40. Stalmeier PF, Busschbach JJ, Lamers LM, et al. The gap effect: discontinuities of preferences around dead. *Health Econ*. 2005;14:679–685.
41. Stalmeier PF, Lamers LM, Busschbach JJ, et al. On the assessment of preferences for health and duration: maximal endurable time and better than dead preferences. *Med Care*. 2007;45:835–841.
42. Sutherland HJ, Llewellyn-Thomas H, Boyd NF, et al. Attitudes toward quality of survival. The concept of "maximal endurable time". *Med Decis Making*. 1982;2:299–309.
43. Buckingham KJ, Devlin NJ. A note on the nature of utility in time and health and implications for cost utility analysis. *Soc Sci Med (1982)*. 2009;68:362–367.
44. Macran S, Kind P. "Death" and the valuation of health-related quality of life. *Med Care*. 2001;39:217–227.
45. Stevens SS. *Psychophysics: Introduction to its Perceptual, Neural, and Social Prospects*. New York, NY: Wiley; 1975.
46. Mulhern B, Longworth L, Brazier J, et al. Binary choice health state valuation and mode of administration: head-to-head comparison of online and CAPI. *Value Health*. 2013;16:104–113.
47. Pickard AS. *US Valuation Study of the EQ-5D-5L*. USA: EuroQol Foundation; 2016.